Chapter 6

Categorical Data Models

6.1 Models for Binary Variable

6.1.1 Logit Models for Binary Variable

– Consider the situation where the realization of the random variable Y_i can have only two different outcomes. Every binary outcome situations outcomes can be coded as values 0 and 1.

– The binary random variable Y_i is said to follow Bernoulli distribution $Y_i \sim Ber(\mu_i)$, where the probabilities $P(Y_i = 1)$ and $P(Y_i = 0)$ are denoted as

$$P(Y_i = 1) = \mu_i, \quad P(Y_i = 0) = 1 - \mu_i. \tag{6.1}$$

– The probability mass function for $Y_i \sim Ber(\mu_i)$ is

$$P(Y_i = y_i) = p(y_i) = \mu_i^{y_i}(1 - \mu_i)^{1-y_i}, \qquad y_i \in \{0, 1\}. \tag{6.2}$$

– When $Y_i \sim Ber(\mu_i)$, then the expected value and the variance of the random variable Y_i are

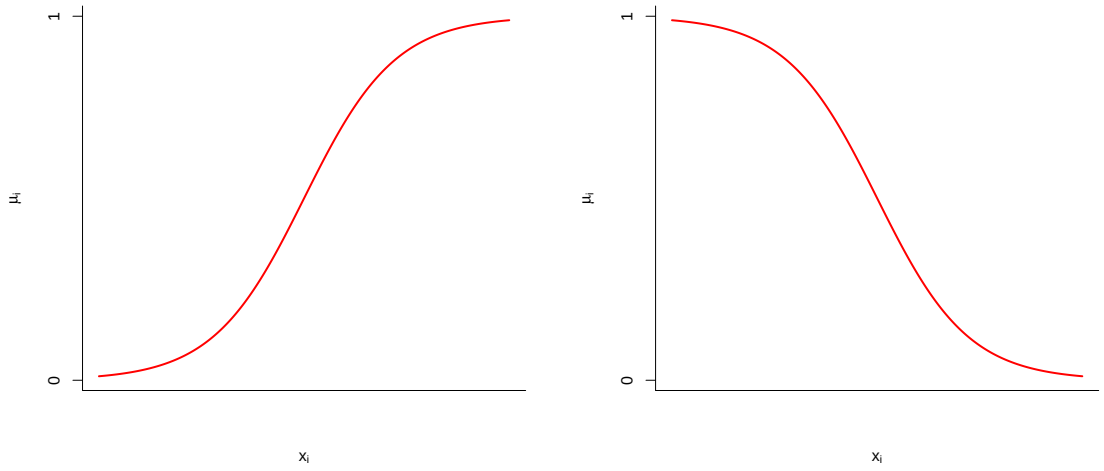
$$\mathrm{E}(Y_i) = \mu_i, \quad \mathrm{Var}(Y_i) = \mu_i(1 - \mu_i). \tag{6.3}$$

– Under the Bernoulli's distribution $Y_i \sim Ber(\mu_i)$, the most used link function is the logit link function

$$\text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}. \tag{6.4}$$

– logit link is nonlinear link function by inducing the expected value to have a form

$$\mu_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}}. \tag{6.5}$$



– $\log$it link is logarithm transformation for the odds:

$$\gamma_i = \frac{\mu_i}{1-\mu_i}. \tag{6.6}$$

– Logarithm of the odds ratio

$$\psi_{x_1+1|x_1} = \frac{\frac{\mu(x_1+1)}{1-\mu(x_1+1)}}{\frac{\mu(x_1)}{1-\mu(x_1)}} \tag{6.7}$$

has the form

$$\begin{aligned}\log(\psi_{x_1+1|x_1}) &= \log\left(\frac{\mu(x_1+1)}{1-\mu(x_1+1)}\right) - \log\left(\frac{\mu(x_1)}{1-\mu(x_1)}\right) \\ &= \beta_0 + \beta_1(x_1+1) + \cdots + \beta_p x_p - (\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) = \beta_1. \end{aligned} \tag{6.8}$$

– Thus the odds ratio has the form

$$\psi_{x_1+1|x_1} = e^{\beta_1}. \tag{6.9}$$

– If the random variable Z follows the standard logistic distribution, then the cumulative distribution function for Z has the form

$$F_Z(z) = \frac{e^z}{1 + e^z}. \tag{6.10}$$

– The expected value μ_i can be viewed as an cumulative distribution function of the standard logistic distribution

$$\mu_i = F_Z(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}}, \tag{6.11}$$

and then logit link is the inverse function of the cumulative distribution function F_Z:

$$\mathrm{logit}(\mu_i) = F_Z^{-1}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}. \tag{6.12}$$

– By using other cumulative distribution functions, alternative link functions can be formed for the expected value μ_i:

$$\Phi^{-1}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \textit{probit (normal) link}, \tag{6.13a}$$

$$F^{-1}_{\mathsf{cauchy}}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \textit{cauchy link}, \tag{6.13b}$$

$$\log\left(-\log(1 - \mu_i)\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \textit{Gumbel link}. \tag{6.13c}$$

6.1.2 Quasi-Bernoulli Models

– If $Y_i \sim Ber(\mu_i)$, then

$$\mathrm{E}(Y_i) = \mu_i, \quad \mathrm{Var}(Y_i) = \mu_i(1-\mu_i). \tag{6.14}$$

– In practice, $\mathrm{Var}(Y_i) = \mu_i(1-\mu_i)$ may not hold.

– In Quasi-Bernoulli situation, the expected value and the variance have the structure

$$\mathrm{E}(Y_i) = \mu_i, \tag{6.15a}$$
$$\mathrm{Var}(Y_i) = \phi\mu_i(1-\mu_i), \tag{6.15b}$$

where $\phi > 0$ unknown dispersion parameter.

– If $\phi > 1$, then the binary model is said to have overdispersion, and if $\phi < 1$, then model has underdispersion.

– Unbiased estimate for ϕ has the form

$$\tilde{\phi} = \frac{\sum_{i=1}^{n} \frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i(1-\hat{\mu}_i)}}{n-(p+1)} = \frac{X^2}{n-p-1}. \tag{6.16}$$

Example 6.1.

Consider the dataset malaria.txt:

```
    subject age   ab mal
1         1  15  546   0
2         2  14  268   0
3         3  12  284   0
4         4  15   38   0
.
100     100   5  138   0
```

A random sample of 100 children aged 3-15 years from a village in Ghana. The children were followed for a period of 8 months. At the beginning of the study, values of a particular antibody were assessed. Based on observations during the study period, the children were categorized into two groups: individuals with and without symptoms of malaria.

Denote variables as following

$$Y = \text{mal},\ X_1 = \text{age},\ X_2 = \text{ab}.$$

(a) Consider the model

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

Calculate the estimate for the expected value μ_{i_*} when $x_{i_*1} = 10, x_{i_*2} = 100$.

(b) Investigate if there is a need to include interaction effect of variables X_1 and X_2 to model. Consider also the model, where one of the explanatory variables is the logarithm of the variable X_2. Explore which link function fits best to the data.

Example 6.2.

Consider the dataset babyfood.txt:

```
  disease nondisease  sex   food
1      77        381  Boy Bottle
2      19        128  Boy  Suppl
3      47        447  Boy Breast
4      48        336 Girl Bottle
5      16        111 Girl  Suppl
6      31        433 Girl Breast

Study on infant respiratory disease, namely the proportions of children
developing bronchitis or pneumonia in their first year of life by type of feeding and sex.
```

Research problem is to model how the explanatory variables X_1 =sex, X_2 =food are effecting to the probability of the outcome disease to appear.

(a) Consider the model

$$\text{logit}(\mu_{jh}) = \beta_0 + \beta_j + \alpha_h.$$

Test at 5% significance level, is the explanatory variable X_2 =food statistically significant variable in the above model.

(b) Assume $\text{Var}(Y_{ijh}) = \phi\mu_{jh}(1-\mu_{jh})$. Test at 5% significance level, is the explanatory variable X_2 =food statistically significant variable in the above model.

(c) Calculate the estimate of the odds ratio $\psi_{\text{Boy, Breast|Girl, Suppl}}$.

6.2 Models for Multinomial Variable

6.2.1 Multinomial Logit Model

– Let the random variable Y_i have m distinctive possible outcomes. Then Y_i said to follow categorical distribution $Y_i \sim Cat(\theta_{i1}, \theta_{i2}, \ldots, \theta_{im})$, where

$$P(Y_i = "1") = \theta_{i1}, \;\; P(Y_i = "2") = \theta_{i2}, \;\; \ldots, P(Y_i = "m") = \theta_{im}. \tag{6.17}$$

– Let the probability $P(Y_i = "1") = \theta_{i1}$ be chosen as an baseline probability in analysis.

– Multinomial logit model has the form

$$\log\left(\frac{\theta_{ik}}{\theta_{i1}}\right) = \beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2} + \cdots + \beta_{pk}x_{ip}, \qquad k = 2, 3, \ldots, m. \tag{6.18}$$

– Under the multinomial logit model, the probabilities $P(Y_i = "k") = \theta_{ik}$ have the forms

$$\theta_{i1} = \frac{1}{1 + \sum_{j=2}^{m} e^{\beta_{0j} + \beta_{1j}x_{i1} + \beta_{2j}x_{i2} + \cdots + \beta_{pj}x_{ip}}}, \tag{6.19a}$$

$$\theta_{ik} = \frac{e^{\beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2} + \cdots + \beta_{pk}x_{ip}}}{1 + \sum_{k=2}^{m} e^{\beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2} + \cdots + \beta_{pk}x_{ip}}}, \qquad k = 2, 3, \ldots, m. \tag{6.19b}$$

Example 6.3.

Based on airborne laser scanning, a total of $N = 1795$ trees could be identified from the forest area, and measures on the height of the tree X_1 =height and the width of the tree X_2 = width was recorded. From the identified trees, 50 trees were randomly selected to the sample and the species of the selected trees were also recorded. The dataset can be found in the file laserscanning.txt

Let us assume that for each tree i the response variable Y = species follows the categorical distribution $Y_i \sim Cat(\theta_{i1}, \theta_{i2}, \ldots, \theta_{i4})$, where

$$k = \begin{cases} 1, & \text{pine}, \\ 2, & \text{spruce}, \\ 3, & \text{birch}, \\ 4, & \text{other species}, \end{cases}$$

Consider the multinomial logit model

$$\mathcal{M}_{1|2}: \quad \text{logit}(\theta_{ik}) = \log\left(\frac{\theta_{ik}}{\theta_{i1}}\right) = \beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2}, \quad k = 2, 3, 4.$$

(a) Calculate the maximum likelihood estimate for the probability $P(Y_{i_*} = 2 = \text{"spruce"} = \theta_{i_*2}$ when

$$x_{i_*1} = 18, \qquad x_{i_*2} = 1.4.$$

(b) Calculate the maximum likelihood estimate for the probability that the tree i_* is either birch or some other species, when

$$x_{i_*1} = 18, \qquad x_{i_*2} = 1.4.$$

(c) Test at 5% significance level, is the explanatory variable $X_2 = \mathsf{width}$ statistically significant variable in the model $\mathcal{M}_{1|2}$.

6.2.2 Ordered Logit Model

– Let the random variable Y_i be defined on ordinal scale with m distinctive possible outcomes. Let the possible outcomes have natural order $"1" < "2" < \cdots < "m"$.

– In ordered logit models, the aim is to model cumulative probabilities

$$P(Y_i \leq k) = \theta_{i1} + \theta_{i2} + \cdots + \theta_{ik}, \qquad k = 1, \ldots, m. \tag{6.20}$$

– So called Cumulative proportional odds logit model has the form

$$\begin{aligned} \log\left(\frac{P(Y_i \leq k)}{1 - P(Y_i \leq k)}\right) &= \text{logit}\left(P(Y_i \leq k)\right) \\ &= \beta_{0k} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \qquad k = 1, \ldots, m-1. \end{aligned} \tag{6.21}$$

– In R, functions `polr` and `clm` are estimating the model

$$\text{logit}\left(P(Y_i \leq k)\right) = \beta_{0k} - (\beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}), \qquad k = 1, \ldots, m-1. \tag{6.22}$$

Example 6.4.

Consider the dataset Housing.txt:

```
     Sat   Infl      Type Cont Freq
1    Low    Low     Tower  Low   21
2 Medium    Low     Tower  Low   21
3   High    Low     Tower  Low   28
.
```

Denote variables as following $Y = \mathsf{Sat},\ X_1 = \mathsf{Infl},\ X_2 = \mathsf{Type},\ X_3 = \mathsf{Cont}$.

(a) Use cumulative proportional odds logit model to model cumulative probabilities of the response variable Y=Sat by including X_1=Infl, X_2 =Type, and X_3=Cont in the model. In case of main effect model, where indexes j, h, l are related to categories of X_1, X_2, X_3,

$$\log\left(\frac{P(Y_i \le k)}{1 - P(Y_i \le k)}\right) = \text{logit}\left(P(Y_i \le k)\right) = \beta_{0k} - (\beta_j + \alpha_h + \delta_l)$$

what sort of estimates the parameters β_{0k} have?

(b) Calculate in main effect model the maximum likelihood estimates for the probabilities θ_{i_*k} when $x_{i_*1} = \mathsf{High}, x_{i_*2} = \mathsf{Apartment}, x_{i_*3} = \mathsf{Low}$.

(c) Test at 5% significance level, is the explanatory variable X_2 =Type statistically significant variable in the above model.

Example 6.5.

The company has large stores in Helsinki and Tampere. The company wanted to find the customer satisfaction of the stores. In the outlets of the stores, customers were able to express their satisfaction by the following type of survey:

The file customersurvey.txt contains the data of the survey. The following contingency table (crosstab) between variables X=location and Y=satisfaction can be created.

```
> data<-read.table(file="customersurvey.txt", sep="\t", dec=".", header=TRUE)
> attach(data)
> T<-table(satisfaction,location)
> T
            location
satisfaction Helsinki Tampere
           1     1054      342
           2     1689      489
           3      865     1192
           4      409      662
           5      188      660
```

The company wanted to find out which proportion of the customers at Tampere store were not happy at all about the customer service they were given while shopping in the store. Thus estimate the probability

$$P(Y_i \geq 4 | x_i = \text{Tampere})$$

by the multinomial logit model

$$\log\left(\frac{\theta_{ik}}{\theta_{i1}}\right) = \beta_{0k} - \beta_{kj}, \qquad k = 2, 3, 4, 5,$$

and by the ordered logit model

$$\text{logit}\left(P(Y_i \leq k)\right) = \beta_{0k} - \beta_j, \qquad k = 1, 2, 3, 4,$$

where β_{kj} and β_j are parameters related to categories of the variable $X = \text{location}$. Calculate the mean square error value $\text{MSE} = \frac{\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\boldsymbol{\theta}}_i)'(\mathbf{y}_i - \hat{\boldsymbol{\theta}}_i)}{n}$ for these models. Which model fits best to the data?

6.3 Statistical Inference in Logit Models

6.3.1 Estimation in Logit Models

– Let $n_1Y_1, n_2Y_2, \ldots, n_nY_n$ be random sample from the binomial distribution with known values of n_i, i.e., $n_iY_i \sim Bin(n_i, \mu_i)$.

– The random variable Y_i belongs to the Exponential Family of Distributions and hence the expected value and the variance are

$$\mathrm{E}(Y_i) = \mu_i, \qquad \mathrm{Var}(Y_i) = \frac{\mu_i(1-\mu_i)}{n_i}. \tag{6.23}$$

– Consider the maximum likelihood estimation of the parameters $\boldsymbol{\beta}$ under logit link function

$$g(\mu_i) = \mathrm{logit}(\mu_i) = \mathrm{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}, \tag{6.24}$$

– Denote

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{n} = \begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ n_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix} = \left(\mathbf{x}_{(1)} : \mathbf{x}_{(2)} : \cdots : \mathbf{x}_{(m)}\right), \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix},$$

and

$$g(\boldsymbol{\mu}) = \text{logit}(\boldsymbol{\mu}) = \text{logit}(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \qquad \boldsymbol{\mu} = \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{1+e^{\mathbf{X}\boldsymbol{\beta}}} = \begin{pmatrix} \frac{e^{\mathbf{x}_1'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_1'\boldsymbol{\beta}}} \\ \frac{e^{\mathbf{x}_2'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_2'\boldsymbol{\beta}}} \\ \vdots \\ \frac{e^{\mathbf{x}_n'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_n'\boldsymbol{\beta}}} \end{pmatrix},$$

$$\mathbf{n} * \mathbf{y} = \begin{pmatrix} n_1 y_1 \\ n_2 y_2 \\ \vdots \\ n_n y_n \end{pmatrix}, \quad \mathbf{n} * \boldsymbol{\mu} = \begin{pmatrix} n_1 \mu_1 \\ n_2 \mu_2 \\ \vdots \\ n_n \mu_n \end{pmatrix}.$$

– The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ must satisfy the likelihood equations

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\text{Var}(Y_i)} \frac{1}{g'(\mu_i)} x_{ij} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\text{Var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \tag{6.25}$$

$$= \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\frac{\mu_i(1-\mu_i)}{n_i}} x_{ij} \left(\mu_i(1-\mu_i) \right) = \sum_{i=1}^{n} n_i (y_i - \mu_i) x_{ij} = 0, \qquad j = 0, 1, 2 \ldots, p.$$

– Hence it must hold that

$$\frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \beta_j} = \mathbf{x}_{(j)}'(\mathbf{n} * \mathbf{y} - \mathbf{n} * \boldsymbol{\mu}) = 0, \qquad j = 0, 1, 2 \ldots, p, \tag{6.26}$$

and thus

$$\mathbf{u}_{\boldsymbol{\beta}} = \frac{\partial l(\boldsymbol{\beta}, \phi)}{\partial \boldsymbol{\beta}} = \mathbf{X}'(\mathbf{n} * \mathbf{y} - \mathbf{n} * \boldsymbol{\mu}) = \mathbf{X}' \left(\mathbf{n} * \mathbf{y} - \mathbf{n} * \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{1 + e^{\mathbf{X}\boldsymbol{\beta}}} \right) = \mathbf{0}. \tag{6.27}$$

– In logit model, the weighted least squares estimation method

$$\mathbf{X}'\mathbf{W}_t\mathbf{X}\boldsymbol{\beta}_{t+1} = \mathbf{X}'\mathbf{W}_t(\mathbf{X}\boldsymbol{\beta}_t + \mathbf{D}_t^{-1}(\mathbf{n} * \mathbf{y} - \boldsymbol{\mu}_t)) \tag{6.28}$$

has the form

$$\begin{aligned} &\mathbf{X}' \left(\mathbf{diag} \left(\mathbf{n} * \tfrac{e^{\mathbf{X}\boldsymbol{\beta}_t}}{(1+e^{\mathbf{X}\boldsymbol{\beta}_t})^2} \right) \right) \mathbf{X}\boldsymbol{\beta}_{t+1} \qquad (6.29) \\ &\quad = \mathbf{X}' \left(\mathbf{diag} \left(\mathbf{n} * \tfrac{e^{\mathbf{X}\boldsymbol{\beta}_t}}{(1+e^{\mathbf{X}\boldsymbol{\beta}_t})^2} \right) \right) \left[\mathbf{X}\boldsymbol{\beta}_t + \left(\mathbf{diag} \left(\tfrac{e^{\mathbf{X}\boldsymbol{\beta}_t}}{(1+e^{\mathbf{X}\boldsymbol{\beta}_t})^2} \right) \right)^{-1} \left(\mathbf{n} * \mathbf{y} - \mathbf{n} * \frac{e^{\mathbf{X}\boldsymbol{\beta}_t}}{1 + e^{\mathbf{X}\boldsymbol{\beta}_t}} \right) \right]. \end{aligned}$$

– Iterative process is continued until the difference $l(\boldsymbol{\beta}_{t+1}) - l(\boldsymbol{\beta}_t)$ is sufficiently small. Starting values are set as $\mathbf{n} * \boldsymbol{\mu}_0 = \mathbf{n} * \mathbf{y}$.

– Estimated covariance matrix $\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}})$ is asymptotically

$$\widehat{\mathrm{cov}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1} = (\mathbf{X}'(\mathbf{diag}(n_i\hat{\mu}_i(1-\hat{\mu}_i)))\mathbf{X})^{-1} = \left(\mathbf{X}' \left(\mathbf{diag} \left(\mathbf{n} * \tfrac{e^{\mathbf{X}\hat{\boldsymbol{\beta}}}}{(1+e^{\mathbf{X}\hat{\boldsymbol{\beta}}})^2} \right) \right) \mathbf{X} \right)^{-1} \tag{6.30}$$

where

$$W_{ii} = \frac{1}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 = \frac{n_i}{\mu_i(1-\mu_i)} (\mu_i(1-\mu_i))^2 = n_i\mu_i(1-\mu_i).$$

6.3.2 Hypotheses Testing in Logit Models

– Consider the following models

$$\begin{aligned}\mathcal{M}_1: &\quad \text{logit}(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1,\\ \mathcal{M}_2: &\quad \text{logit}(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,\end{aligned}$$

where $\mathbf{X} = (\mathbf{X}_1 : \mathbf{X}_2)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$.

– Let us test the hypotheses

$$\begin{aligned}H_0 &: \text{Model } \mathcal{M}_1 \text{ is the true model},\\ H_a &: \text{Model } \mathcal{M}_2 \text{ is the true model}.\end{aligned}$$

– The difference of the deviancies is

$$\Delta D = 2\sum_{i=1}^{n} n_i y_i \log\left(\frac{n_i\hat{\mu}_{i;\mathcal{M}_2}}{n_i\hat{\mu}_{i;\mathcal{M}_1}}\right) + 2\sum_{i=1}^{n}(n_i - n_i y_i)\log\left(\frac{n_i - n_i\hat{\mu}_{i;\mathcal{M}_2}}{n_i - n_i\hat{\mu}_{i;\mathcal{M}_1}}\right). \tag{6.31}$$

– Under H_0 hypothesis, $\Delta D \sim \chi^2_{(q)}$, where $q = \text{rank}(\mathbf{X}_2)$.

– In Quasi-Logit situation $\text{Var}(Y_i) = \phi(\mu_i(1-\mu_i))$, the test statistic is

$$F = \frac{\frac{1}{\phi}\cdot\left(D(\mathcal{M}_1) - D(\mathcal{M}_2)\right)/\text{rank}(\mathbf{X}_2)}{\frac{1}{\phi}\cdot X^2/n - (p+1)} = \frac{\left(D(\mathcal{M}_1) - D(\mathcal{M}_2)\right)/\text{rank}(\mathbf{X}_2)}{\tilde{\phi}} \sim F_{(q,n-(p+1))}. \tag{6.32}$$

6.3.3 Confidence and Prediction Intervals in Logit Models

– In logit model

$$\mathcal{M}: \ \text{logit}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$$

confidence interval for the probability μ_{i_*} can be created based on Wald statistic.

– The maximum likelihood estimator for the link function is

$$\text{logit}(\hat{\mu}_{i_*}) = \mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}}.$$

– Estimated variance is

$$\widehat{\text{Var}}\left(\text{logit}(\hat{\mu}_{i_*})\right) = \mathbf{x}'_{i_*}\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_{i_*}. \tag{6.33}$$

– The $100(1-\alpha)\%$ confidence interval for the logit link function is

$$\left[\mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}} - z_{\alpha/2}\sqrt{\mathbf{x}'_{i_*}\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_{i_*}}, \mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}} + z_{\alpha/2}\sqrt{\mathbf{x}'_{i_*}\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_{i_*}}\right], \tag{6.34}$$

where $P(Z > z_{\alpha/2}) = \alpha/2$ as $Z \sim N(0,1)$.

– The $100(1-\alpha)\%$ confidence interval for the probability $\mu(\mathbf{x})$ is

$$\left[\frac{e^{\mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}} - z_{\alpha/2}\sqrt{\mathbf{x}'_{i_*}\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_{i_*}}}}{1+e^{\mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}} - z_{\alpha/2}\sqrt{\mathbf{x}'_{i_*}\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_{i_*}}}}, \frac{e^{\mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}} + z_{\alpha/2}\sqrt{\mathbf{x}'_{i_*}\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_{i_*}}}}{1+e^{\mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}} + z_{\alpha/2}\sqrt{\mathbf{x}'_{i_*}\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_{i_*}}}}\right]. \tag{6.35}$$

– In binary situation, it is meaningful to predict the sum $Y_S = \sum_{i=n+1}^{N} y_{if}$.

– The maximum likelihood predictor for the sum $Y_S = \sum_{i=n+1}^{N} y_{if}$ is

$$\widehat{Y}_S = \sum_{i=n+1}^{N} \hat{y}_{if} = \sum_{i=n+1}^{N} \hat{\mu}_{if} = \sum_{i=n+1}^{N} \frac{e^{\mathbf{x}'_{if}\hat{\boldsymbol{\beta}}}}{1 + e^{\mathbf{x}'_{if}\hat{\boldsymbol{\beta}}}}. \tag{6.36}$$

– The prediction interval for the sum Y_S can be based on the interval

$$\left[\widehat{Y}_S - z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}_b(e_{Y_S})}, \widehat{Y}_S + z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}_b(e_{Y_S})}\right], \tag{6.37}$$

where $\widehat{Y}_S$ is the point prediction of the Y_S, $\widehat{\mathrm{Var}}_b(e_{Y_S})$ is the bootstrap estimate of the variance of the prediction error $e_{Y_S} = Y_S - \widehat{Y}_S$, and $z_{\alpha/2}$ is $1 - \alpha/2$ quantile of the standard normal distribution.

Parametric bootstrap based method

1. Find the estimates $\hat{\mu}_i$ for all $i = 1, 2, \ldots, n, n+1, \ldots, N$.
2. Simulate y_{ib} for all $i = 1, 2, \ldots, n, n+1, \ldots, N$ from the distribution $Y_i \sim Ber(\hat{\mu}_i)$.
3. Based on simulated values $y_{1b}, y_{2b}, \ldots, y_{nb}$ create bootstrap point prediction $\widehat{Y}_{S_b}$ for the sum Y_S.
4. Find the bootstrap prediction error $e_{Y_{S_b}} = Y_{S_b} - \widehat{Y}_{S_b}$.

5. Repeat M times the steps 2-4, and then calculate the sample variance $\widehat{\mathrm{Var}}_b(e_{Y_S})$ of the simulated values $e_{Y_{S_b}}$.
6. Create the prediction interval

$$\left[\widehat{Y}_S - z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}_b(e_{Y_S})}, \widehat{Y}_S + z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}_b(e_{Y_S})}\right] .$$

Example 6.6.

Consider the dataset christmastree.txt:

```
> head(data)
  age distance disease
1 old      242       0
2 old      107       0
3 old       30       1
4 old      103       1
5 old      239       0
6 old       28       0
> tail(data)
        age distance disease
17545 young      295      NA
17546 young      283      NA
17547 young      155      NA
17548 young       99      NA
17549 young      168      NA
17550 young      257      NA


age - classified age of the tree,
distance - distance to the nearby leafy forest,
disease - is a tree affected by fungal disease, 0=no, 1=yes.
```

Research problem is to model how the explanatory variables X_1 =distance, X_2 =age are effecting to the probability of the fungal disease to occur. Model the disease variable with logit model

$$\mathrm{E}(Y_i) = \mu_i = \frac{e^{\beta_0+\beta_1 x_i+\alpha_j}}{1+e^{\beta_0+\beta_1 x_i+\alpha_j}},$$
$$\mathrm{Var}(Y_i) = \mu_i(1-\mu_i).$$

(a) Construct the 95% confidence interval for μ_{i_*} when

```
age distance
old      242
```

(b) Create the maximum likelihood prediction for how many Christmas trees have been affected by fungal disease in the breeding area, i.e., predict the value $\sum_{i=1}^{n} Y_i + \sum_{i=n+1}^{N} Y_i$.

(c) Create bootstrap estimate for variance $\mathrm{Var}\left(e_{Y_{S_\Omega}}\right)$ of the prediction error $e_{Y_S} = Y_S - \widehat{Y}_S$.

(d) Create a 80% prediction interval for how many Christmas trees have affected fungal disease in the breeding area by using the above logistic model.