Chapter 5

Count Data Models

5.1 Poisson Models and Beyond

5.1.1 Poisson Models

– In count data models, the random response variable Y_i can have a realization as an nonnegative integer $y_i = \{0, 1, 2, 3, 4, \dots\}$.

– In count data models, the random response variable Y_i is assumed to follow either *Poisson* distribution or *negative binomial* distribution.

– If the random variable Y_i follows the Poisson distribution $Y_i \sim Poi(\mu_i)$, then the realization will have nonnegative integer value $y_i = \{0, 1, 2, 3, 4, \dots\}$, and $\mathrm{E}(Y_i) = \mu_i$ and $\mathrm{Var}(Y_i) = \mu_i$.

– Under Poisson distribution, possible link functions $g(\mu_i)$ are

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \textit{identity link}, \tag{5.1a}$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \log \textit{link}, \tag{5.1b}$$

$$\sqrt{\mu_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \textit{square root link}. \tag{5.1c}$$

5.1.2 Quasi-Poisson Models

- Poisson models include assumption of $\mathrm{E}(Y_i) = \mu_i = \mathrm{Var}(Y_i)$. In practice, this often does not hold.
- In quasi-Poisson situation, the expected value and the variance have the structure

$$\mathrm{E}(Y_i) = \mu_i, \tag{5.2a}$$

$$\mathrm{Var}(Y_i) = \phi\mu_i, \tag{5.2b}$$

  where $\phi > 0$ unknown dispersion parameter.
- If $\phi > 1$, then the count model is said to have overdispersion, and if $\phi < 1$, then model has underdispersion.
- Unbiased estimate for ϕ has the form

$$\tilde{\phi} = \frac{\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}}{n - (p+1)} = \frac{X^2}{n - (p+1)}. \tag{5.3}$$

5.1.3 Negative Binomial Models

– If there is a clear evidence of overdispersion under Poisson model, then negative binomial model is useful an alternative model in case of overdispersion.

– The random variable Y_i follows negative binomial distribution $Y_i \sim NegBin(\mu_i, \theta)$, if the probability mass function has the form

$$f(y_i|\mu_i, \theta) = \frac{\Gamma(y_i + \theta)}{\Gamma(\theta) y_i!} \cdot \frac{\mu_i^{y_i} \theta^{\theta}}{(\mu_i + \theta)^{(y_i + \theta)}}, \quad y_i = 0, 1, 2, 3, \ldots$$

– If $Y_i \sim NegBin(\mu_i, \theta)$, then

$$\mathrm{E}(Y_i) = \mu_i, \tag{5.4a}$$

$$\mathrm{Var}(Y_i) = \mu_i + \frac{\mu_i^2}{\theta}. \tag{5.4b}$$

– The parameter θ is usually estimated by the maximum likelihood method.

– Under negative binomial distribution $Y_i \sim NegBin(\mu_i, \theta)$, the most common link function is $\log$ link $g(\mu_i) = \log(\mu_i)$.

Example 5.1.

Consider modeling the number of plant species in the data set species.txt:

```
     pH    Biomass Species
1  high 0.46929722      30
2  high 1.73087043      39
3  high 2.08977848      44
4  high 3.92578714      35
.
.
90  low 4.87050789       3
```

In the dataset, the response is a count of the number of plant species on plots that have different biomass (a continous explanatory variable) and different soil pH (a categorical variable with three levels: high=1, mid=2, and low=3).

Denote variables as following

$$Y = \text{Species}, \quad X_1 = \text{Biomass} \quad X_2 = \text{pH}.$$

Assume first that $Y_i \sim Poi(\mu_i)$. Consider the models

$$\begin{aligned}
\mathcal{M}_{2-\log} : &\quad \log(\mu_i) = \beta_0 + \alpha_j, \\
\mathcal{M}_{1|2-\log} : &\quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \alpha_j, \\
\mathcal{M}_{12-\log} : &\quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i2}, \\
\mathcal{M}_{12-\sqrt{\cdot}} : &\quad \sqrt{\mu_i} = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i2}, \\
\mathcal{M}_{12-I} : &\quad \mu_i = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_j x_{i2}.
\end{aligned}$$

(a) Under the model $\mathcal{M}_{2-\log}$, consider the hypotheses

$$H_0 : \alpha_j = 0,$$
$$H_1 : \alpha_j \neq 0.$$

Select the appropriate test statistic to test the above hypotheses. Calculate the value of the test statistic.

(b) Under the model $\mathcal{M}_{1|2-\log}$, calculate the maximum likelihood estimate for the parameter β_1.

(c) Consider the hypotheses

$$H_0 : \text{Model } \mathcal{M}_{1|2-\log} \text{ is the true model,}$$
$$H_1 : \text{Model } \mathcal{M}_{12-\log} \text{ is the true model.}$$

Calculate the value of the test statistic and p-value. What if we assume $\operatorname{Var}(Y_i) = \phi\mu_i$. Calculate the estimate for the parameter ϕ. Calculate the value of the test statistic and p-value in this case too.

(d) Which one of the models $\mathcal{M}_{12-\log}$, $\mathcal{M}_{12-\sqrt{\cdot}}$ or $\mathcal{M}_{12-I}$ fits best in the data?

(e) Assume $Y_{ij} \sim NegBin(\mu_{ij}, \theta)$. Calculate the maximum likelihood estimate for the parameter β_1 in model $\mathcal{M}_{1|2-\log}$.

5.1.4 Ratio Models

- Often in count data situations, there is index variable t_i, which sets the conditions in which time, quantity, or space the random variable Y_i can have its realization.
- In ratio models, the main interest is to model the expected value of the ratio $Z_i = \frac{Y_i}{t_i}$, where Y_i is the nonnegative integer random variable, and where t_i in known nonrandom index variable.
- The expected value of the ratio $Z_i = \frac{Y_i}{t_i}$ is

$$\mathrm{E}(Z_i) = \mathrm{E}\left(\frac{Y_i}{t_i}\right) = \frac{\mu_i}{t_i}. \tag{5.5}$$

- In ratio models, the $\log$ link function is the most suitable way of modeling the expected value of ratio:

$$\log\left(\frac{\mu_i}{t_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}. \tag{5.6}$$

- Log-linear ratio model can be written as

$$\log(\mu_i) = \log(t_i) + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \tag{5.7}$$

where $\log(t_i)$ is a *offset* variable meaning that related parameter is set on value 1.

Example 5.2.

Consider the data in the file ratescancer.txt, where lung cancer cases occur in certain cities at certain ages. In dataset, the response variable is the $Y = \text{cases}$ and the index variable is $t = \text{pop}$. The explanatory variables are $X_1 = \text{age}$ and $X_2 = \text{city}$.

```
> ratescancer
         city   age  pop cases
1  Fredericia 40-54 3059    11
2     Horsens 40-54 2879    13
3     Kolding 40-54 3142     4
4       Vejle 40-54 2520     5
5  Fredericia 55-59  800    11
6     Horsens 55-59 1083     6
7     Kolding 55-59 1050     8
8       Vejle 55-59  878     7
9  Fredericia 60-64  710    11
10    Horsens 60-64  923    15
11    Kolding 60-64  895     7
12      Vejle 60-64  839    10
13 Fredericia 65-69  581    10
14    Horsens 65-69  834    10
15    Kolding 65-69  702    11
16      Vejle 65-69  631    14
17 Fredericia 70-74  509    11
18    Horsens 70-74  634    12
19    Kolding 70-74  535     9
20      Vejle 70-74  539     8
21 Fredericia   75+  605    10
22    Horsens   75+  782     2
23    Kolding   75+  659    12
24      Vejle   75+  619     7
```

(a) Consider the ratio model

$$\log\left(\frac{\mu_{jh}}{t_i}\right) = \beta_0 + \beta_j + \alpha_h,$$

when it is assumed $Y_i \sim Poi(\mu_{jh})$. Test at 5% significance level, is the explanatory variable X_2=city statistically significant variable in the model. Repeat the testing under assumption of $\mathrm{Var}(Y_i) = \phi\mu_{jh}$.

(b) Consider the ratio model

$$\log\left(\frac{\mu_{jh}}{t_i}\right) = \beta_0 + \beta_j + \alpha_h.$$

Calculate the maximum likelihood estimate for the ratio $\frac{\mu_{jh}}{t_i}$ when $x_1 = 70 - 74$ and $x_2 = \text{Kolding}$. Construct also 95% confidence interval for the ratio $\frac{\mu_{jh}}{t_i}$.

5.1.5 Zero-Inflated Models

– In count data models, there may be situations where the number of zero occurs significantly more frequently than it should occur under Poisson or negative binomial distribution.

– Zero-inflated models are one way to model values of the nonnegative integer response variable Y_i when realizations $y_1, y_2, \ldots, y_n$ in the data are having the value zero too frequently.

– The zero-inflated Poisson model is a mixture model with two zero generating processes. The first process just generates zeros, and the second process is a Poisson distribution which generates nonnegative integers, some of which may be zero.

– The zero-inflated Poisson model has the probability structure

$$P(Y_i = 0) = \theta_i + (1 - \theta_i)e^{-\mu_i}, \tag{5.8}$$

$$P(Y_i = y_i) = (1 - \theta_i)\frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, \quad y_i = 1, 2, 3, \ldots \tag{5.9}$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}. \tag{5.10}$$

– In zero inflated model, the expected value μ_i with $\log$ link has the form

$$\mu_i = \theta_i \cdot 0 + (1 - \theta_i) \cdot e^{(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})}. \tag{5.11}$$

– The parameter θ_i can also depend on the explanatory variables by the logit link structure

$$\mathrm{logit}(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_p x_{ip}, \tag{5.12}$$

where $\alpha_0, \alpha_1, \ldots, \alpha_p$ are unknown parameters.

Example 5.3.

Consider the data set related to file ships.txt:

```
The file has 34 rows corresponding to the observed combinations of type of ship,
year of construction and period of operation.
Each row has information on five variables as follows:
 - ship type, coded 1-5 for A, B, C, D and E,
 - year of construction (1=1960-64, 2=1965-70, 3=1970-74, 4=1975-79),
 - period of operation (1=1960-74, 2=1975-79)
 - months of service, ranging from 63 to 20,370, and
 - damage incidents, ranging from 0 to 53.
  type construction operation months damage
1    A      1960-64   1960-74    127      0
2    A      1960-64   1975-79     63      0
3    A      1965-69   1960-74   1095      3
4    A      1965-69   1975-79   1095      4
.
```

Interest in study is to model how the explanatory variables $X_1 =$type, $X_2 =$construction, $X_3 =$operation are effecting to the expected values of Y $=$damage variable, and to the expected value of the ratio $Z = \frac{Y}{t}$, where the index variable is $t =$months.

(a) Consider the following zero-inflated model

$$\mu_i = \theta_i \cdot 0 + (1 - \theta_i) \cdot e^{\log(t_i)+\beta_0+\beta_j+\alpha_h+\delta_k},$$

where is assumed that $Y_i \sim \theta_i \cdot I_{\{0\}} + (1-\theta_i) \cdot Poi(\mu_i),\ \operatorname{logit}(\theta_i) = \alpha_0 + \alpha_1 \log(t_i)$.Calculate the maximum likelihood estimates for the expected value μ_{i_*} and the ratio $\frac{\mu_{i*}}{t_{i*}}$ when $x_{i_*1} = \mathsf{C}$, $x_{i_*2} = \mathsf{1975\text{-}79}$, $x_{i_*3} = \mathsf{1975\text{-}79}$, and $t_{i_*} = 1000$.

(b) Test at 5% significance level, is the explanatory variable type statistically significant variable in the above model.

(c) Which one of the following models fits best in the data when the comparison is based on the mean squared error, MSE, value?

$$\mathcal{M}_1 : Y_i \sim Poi(\mu_i), \qquad \mu_i = e^{\log(t_i)+\beta_0+\beta_{1j}+\alpha_h+\delta_k},$$

$$\mathcal{M}_2 : Y_i \sim \theta_i \cdot I_{\{0\}} + (1-\theta_i) \cdot Poi(\mu_i),$$

$$\mu_i = \theta_i \cdot 0 + (1-\theta_i) \cdot e^{\log(t_i)+\beta_0+\beta_j+\alpha_h+\delta_k}, \quad \operatorname{logit}(\theta_i) = \alpha_0 + \alpha_1 \log(t_i),$$

$$\mathcal{M}_3 : Y_i \sim \theta_i \cdot I_{\{0\}} + (1-\theta_i) \cdot NegBin(\mu_i, \theta),$$

$$\mu_i = \theta_i \cdot 0 + (1-\theta_i) \cdot e^{\log(\log(t_i)+\beta_0+\beta_j+\alpha_h+\delta_k}, \quad \operatorname{logit}(\theta_i) = \alpha_0 + \alpha_1 \log(t_i).$$

5.2 Statistical Inference in Poisson Models

5.2.1 Estimation in Poisson Models

- The random variable is said to follow Poisson distribution if its the probability mass function has the form

$$f(y_i|\mu_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \exp\left(\frac{y_i\log(\mu_i) - e^{\log(\mu_i)}}{1} - \log(y_i!)\right), \qquad y_i = 0, 1, 2, 3, \ldots \tag{5.13}$$

- The Poisson distribution belongs to the Exponential Family of Distributions and the expected value and variance are $\mathrm{E}(Y_i) = \mu_i$, $\mathrm{Var}(Y_i) = \mu_i$.

- Consider the maximum likelihood estimation in Poisson log-linear model

$$Y_i \sim Poi(\mu_i), \qquad \log(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \qquad \boldsymbol{\mu} = e^{\mathbf{X}\boldsymbol{\beta}}. \tag{5.14}$$

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

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial\beta_j} &= \sum_{i=1}^{n}\frac{(y_i-\mu_i)}{\mathrm{Var}(Y_i)}x_{ij}\left(\frac{\partial\mu_i}{\partial\eta_i}\right) = \sum_{i=1}^{n}\frac{(y_i-\mu_i)}{\mu_i}\cdot x_{ij}\cdot\mu_i \\ &= \sum_{i=1}^{n}(y_i-\mu_i)x_{ij} = 0, \qquad j = 0, 1, 2, \ldots, p.\end{aligned} \tag{5.15}$$

– Thus the likelihood equations have the form

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} &= u_j = \mathbf{x}'_{(j)}(\mathbf{y} - e^{\mathbf{X}\boldsymbol{\beta}}) = 0, \qquad j = 0, 1, 2, \dots p, \\ \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \mathbf{u}_{\boldsymbol{\beta}} = \mathbf{X}'(\mathbf{y} - e^{\mathbf{X}\boldsymbol{\beta}}) = \mathbf{0} \qquad (5.16) \end{aligned}$$

– Since in Poisson log-linear model

$$\begin{aligned} \mathrm{E}\left[\frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_j} \cdot \frac{\partial l_i(\boldsymbol{\beta})}{\partial \beta_h}\right] &= \mathrm{E}\left[\frac{(y_i - \mu_i)}{\mathrm{Var}(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right) \cdot \frac{(y_i - \mu_i)}{\mathrm{Var}(Y_i)} x_{ih} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right] \\ &= \frac{x_{ij} x_{ih}}{\mathrm{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 = \frac{x_{ij} x_{ih}}{\mu_i} (\mu_i)^2 = x_{ij} \mu_i x_{ih}, \qquad (5.17) \end{aligned}$$

the Fisher information matrix is

$$\begin{aligned} \mathbf{F} &= \mathrm{E}\left[\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right] = -\,\mathrm{E}\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right] \\ &= \mathbf{X}'\mathbf{W}\mathbf{X} = \mathbf{X}' \,\mathbf{diag}(\boldsymbol{\mu})\mathbf{X} = \mathbf{X}' \,\mathbf{diag}[e^{\mathbf{X}\boldsymbol{\beta}}]\mathbf{X}. \qquad (5.18) \end{aligned}$$

– In Poisson log-linear model, the weighted least squares estimation method has the form

$$\begin{aligned}\mathbf{X}'\mathbf{W}_t\mathbf{X}\boldsymbol{\beta}_{t+1} &= \mathbf{X}'\mathbf{W}_t(\mathbf{X}\boldsymbol{\beta}_t + \mathbf{D}_t^{-1}(\mathbf{y}-\boldsymbol{\mu}_t)) \\ \mathbf{X}'\,\mathbf{diag}[e^{\mathbf{X}\boldsymbol{\beta}_t}]\mathbf{X}\boldsymbol{\beta}_{t+1} &= \mathbf{X}'\,\mathbf{diag}[e^{\mathbf{X}\boldsymbol{\beta}}](\mathbf{X}\boldsymbol{\beta}_t + \left(\mathbf{diag}[e^{\mathbf{X}\boldsymbol{\beta}}]\right)^{-1}(\mathbf{y}-e^{\mathbf{X}\boldsymbol{\beta}_t})). \end{aligned} \tag{5.19}$$

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

– The maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is asymptotically efficient estimator, and hence

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \mathbf{F}^{-1} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} = (\mathbf{X}'\,\mathbf{diag}(\boldsymbol{\mu})\mathbf{X})^{-1} = (\mathbf{X}'\,\mathbf{diag}[e^{\mathbf{X}\boldsymbol{\beta}}]\mathbf{X})^{-1}, \tag{5.20}$$

and furthermore,

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \widehat{\mathbf{F}}^{-1} = (\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1} = (\mathbf{X}'\,\mathbf{diag}(\hat{\boldsymbol{\mu}})\mathbf{X})^{-1} = (\mathbf{X}'\,\mathbf{diag}[e^{\mathbf{X}\hat{\boldsymbol{\beta}}}]\mathbf{X})^{-1}, \tag{5.21}$$

– Approximately

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})). \tag{5.22}$$

5.2.2 Hypotheses Testing in Poisson Models

– Consider the hypotheses related to competing Poisson log-linear models $\mathcal{M}_1$ and $\mathcal{M}_2$:

$$\begin{aligned} H_0 &: \text{Model } \mathcal{M}_1 : \ \log(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 \text{ is the true model,} \\ H_1 &: \text{Model } \mathcal{M}_2 : \ \log(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 \text{ is the true model.} \end{aligned}$$

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

– Let us assume that the intercept β_0 is in both of the models $\mathcal{M}_1$ and $\mathcal{M}_2$.

– Above hypotheses correspond the hypotheses

$$\begin{aligned} H_0 &: \boldsymbol{\beta}_2 = \mathbf{0}, \\ H_a &: \boldsymbol{\beta}_2 \neq \mathbf{0}. \end{aligned}$$

– Testing can be based on the likelihood ratio statistic

$$LR = -2\left[l(\hat{\boldsymbol{\beta}}_1|\mathcal{M}_1) - l(\hat{\boldsymbol{\beta}}|\mathcal{M}_2)\right]. \tag{5.23}$$

– The likelihood ratio statistic LR is equal to the difference of the deviancies

$$LR = \Delta D = D(\mathcal{M}_1) - D(\mathcal{M}_2). \tag{5.24}$$

- For the log-linear Poisson Model, the difference of the deviancies is

$$\Delta D = D(\mathcal{M}_1) - D(\mathcal{M}_2) = 2\left(\sum_{i=1}^{n} y_i \log\left(\frac{\hat{\mu}_i(\mathcal{M}_2)}{\hat{\mu}_i(\mathcal{M}_1)}\right)\right) = 2\left(\sum_{i=1}^{n} y_i \log\left(\frac{e^{\mathbf{x}_{i1}'\hat{\boldsymbol{\beta}}_1 + \mathbf{x}_{i2}'\hat{\boldsymbol{\beta}}_2}}{e^{\mathbf{x}_{i1}'\hat{\boldsymbol{\beta}}_1}}\right)\right).$$

- Under H_0 hypothesis, $\Delta D \sim \chi^2_{(q)}$, where $q = \mathrm{rank}(\mathbf{X}_2)$.
- In Quasi-Poisson situation $\mathrm{Var}(Y_i) = \phi\mu_i$, the ratio $\frac{\Delta D}{\phi}$ follows under H_0 hypothesis $\frac{\Delta D}{\phi} \sim \chi^2_{(q)}$ and $\frac{1}{\phi} \cdot \frac{\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}}{n-(p+1)} \sim \chi^2_{n-(p+1)}$.
- Thus in Quasi-Poisson situation, the test statistic is

$$F = \frac{\frac{1}{\phi} \cdot (D(\mathcal{M}_1) - D(\mathcal{M}_2)) \,/\, \mathrm{rank}(\mathbf{X}_2)}{\frac{1}{\phi} \cdot X^2/n - (p+1)} = \frac{(D(\mathcal{M}_1) - D(\mathcal{M}_2)) \,/\, \mathrm{rank}(\mathbf{X}_2)}{\tilde{\phi}},$$

  which follows $F_{(q,n-(p+1))}$ distribution under H_0 hypothesis, where $q = \mathrm{rank}(\mathbf{X}_2)$.

5.2.3 Confidence and Prediction Intervals

– In Poisson log-linear model, the maximum likelihood estimator for the link function $\log(\mu_{i_*}) = \eta_{i_*} = \mathbf{x}'_{i_*}\boldsymbol{\beta}$ is

$$\widehat{\log(\mu_{i_*})} = \hat{\eta}_{i_*} = \mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}}. \tag{5.25}$$

– Estimated variance for $\hat{\eta}_{i_*} = \mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}}$ is

$$\widehat{\mathrm{Var}}\left(\mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}}\right) = \mathbf{x}'_{i_*}\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_{i_*} = \mathbf{x}'_{i_*}(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{x}_{i_*}. \tag{5.26}$$

– Since approximately

$$\mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}} \sim N(\mathbf{x}'_{i_*}\boldsymbol{\beta}, \mathbf{x}'_{i_*}(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{x}_{i_*}), \tag{5.27}$$

the $100(1-\alpha)\%$ confidence interval for the link function $\eta_{i_*} = \mathbf{x}'_{i_*}\boldsymbol{\beta}$ is

$$\left[\mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}} - z_{\alpha/2}\sqrt{\mathbf{x}'_{i_*}\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_{i_*}}, \mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}} + z_{\alpha/2}\sqrt{\mathbf{x}'_{i_*}\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_{i_*}}\right] = \left[L_{\alpha/2}, U_{\alpha/2}\right], \tag{5.28}$$

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

– The $100(1-\alpha)\%$ confidence interval for the expected value μ_{i_*} is

$$\left[\exp(L_{\alpha/2}), \exp(U_{\alpha/2})\right]. \tag{5.29}$$

- The maximum likelihood predictor for the new observation Y_f (observable in future) with given values of the explanatory variables $\mathbf{x}_f$ is

$$\hat{Y}_f = \exp(\mathbf{x}'_f\hat{\boldsymbol{\beta}}). \tag{5.30}$$

- By so called delta method, it can be shown that approximately

$$\mathrm{Var}(\hat{Y}_f) = \left(\frac{\partial \mu_f}{\partial \eta_f}\right)^2 \cdot \mathbf{x}'_f \,\mathrm{Cov}(\hat{\boldsymbol{\beta}})\mathbf{x}_f = \mu_f^2 \cdot \mathbf{x}'_f(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_f. \tag{5.31}$$

- Since for the prediction error $e_f = Y_f - \hat{Y}_f$ it holds

$$\mathrm{Var}(e_f) = \mathrm{Var}(Y_f) + \mathrm{Var}(\hat{Y}_f) = \mu_f + \mu_f^2 \cdot \mathbf{x}'_f \,\mathrm{Cov}(\hat{\boldsymbol{\beta}})\mathbf{x}_f, \tag{5.32}$$

  the estimated variance of the prediction error is

$$\widehat{\mathrm{Var}}(e_f) = \hat{\mu}_f \left[1 + \hat{\mu}_f \cdot \mathbf{x}'_f\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_f\right], \tag{5.33}$$

- Approximately $Y_f - \hat{Y}_f \sim N(0, \widehat{\mathrm{Var}}(e_f))$, and hence the $100(1-\alpha)\%$ prediction interval for the new observation Y_f is

$$\left[\exp(\mathbf{x}'_f\hat{\boldsymbol{\beta}}) - z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(e_f)}, \exp(\mathbf{x}'_f\hat{\boldsymbol{\beta}}) + z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(e_f)}\right]. \tag{5.34}$$

Example 5.4.

Consider the data in the file ratescancer.txt, where lung cancer cases occur in certain cities at certain ages. In dataset, the response variable is the $Y = \textsf{cases}$ and the index variable is $t = \textsf{pop}$. The explanatory variables are $X_1 = \textsf{age}$ and $X_2 = \textsf{city}$.

```
> ratescancer
        city   age  pop cases
1  Fredericia 40-54 3059    11
2     Horsens 40-54 2879    13
3     Kolding 40-54 3142     4
.
```

```
19     Kolding 70-74  535     9
20       Vejle 70-74  539     8
21 Fredericia   75+  605    10
22     Horsens   75+  782     2
23     Kolding   75+  659    12
24       Vejle   75+  619     7
```

Consider the ratio model

$$\log\left(\frac{\mu_{jh}}{t_i}\right) = \beta_0 + \beta_j + \alpha_h.$$

Create 80% prediction interval for the ratio $\frac{Y_f}{t_f}$ when $x_1 = 70 - 74$ and $x_2 = \textsf{Kolding}$.