

Chapter 3

Estimation in Generalized Linear Models

3.1 Introduction to Generalized Linear Models

3.1.1 Generalized Linear Normal Model

- Let us assume that random variable Y_1, Y_2, \dots, Y_n are following the normal distribution $Y_i \sim N(\mu_i, \sigma^2)$.
- In generalized linear models, the explanatory variables X_1, X_2, \dots, X_p are effecting to the expected value μ_i through the link function g :

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}. \quad (3.1)$$

- In practice, possible link functions $g(\mu_i)$ are

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad \textit{identity link}, \quad (3.2a)$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad \textit{log-link}, \quad (3.2b)$$

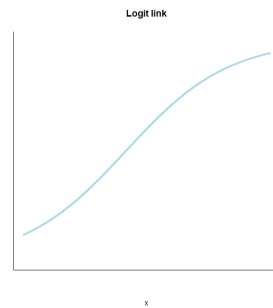
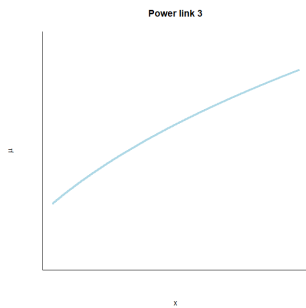
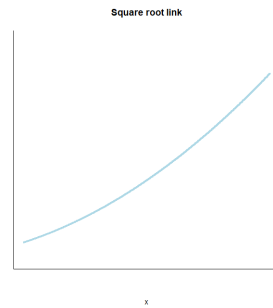
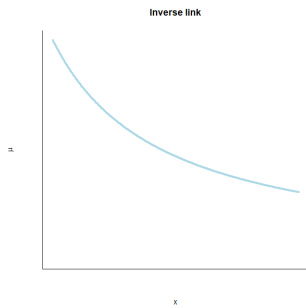
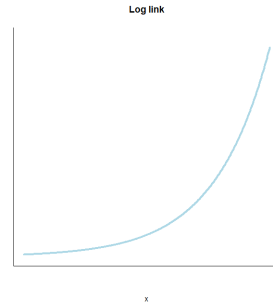
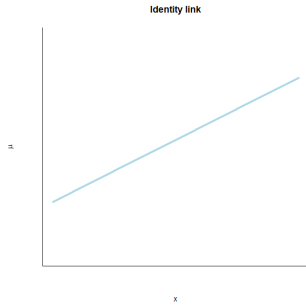
$$\frac{1}{\mu_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad \textit{inverse link}, \quad (3.2c)$$

$$\sqrt{\mu_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad \textit{square root link}, \quad (3.2d)$$

$$\mu_i^k = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad \textit{Power link } k, \quad (3.2e)$$

$$\log\left(\frac{\mu_i}{1 - \mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad \textit{logit link}. \quad (3.2f)$$

-
- In case on nonlinear link functions, the models are nonlinear with respect to explanatory variables X_1, X_2, \dots, X_p with inverse function $\mu_i = g^{-1}(\mathbf{x}_i, \boldsymbol{\beta}) = h(\mathbf{x}_i, \boldsymbol{\beta})$ being:



$$\mu_i = \mathbf{x}_i' \boldsymbol{\beta}, \quad \text{identity link}, \quad (3.3a)$$

$$\mu_i = e^{\mathbf{x}_i' \boldsymbol{\beta}}, \quad \text{log-link}, \quad (3.3b)$$

$$\mu_i = \frac{1}{\mathbf{x}_i' \boldsymbol{\beta}}, \quad \text{inverse link}, \quad (3.3c)$$

$$\mu_i = (\mathbf{x}_i' \boldsymbol{\beta})^2, \quad \text{square root link}, \quad (3.3d)$$

$$\mu_i = (\mathbf{x}_i' \boldsymbol{\beta})^{\frac{1}{k}}, \quad \text{Power link } k, \quad (3.3e)$$

$$\mu_i = \frac{e^{\mathbf{x}_i' \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i' \boldsymbol{\beta}}}, \quad \text{logit link}. \quad (3.3f)$$

-
- Polynomial models are often used models which are nonlinear with respect to explanatory variables. For example (in case of X_1, X_2), second degree polynomial model with interaction term (and with identity link) has the form

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1} x_{i2}. \quad (3.4)$$

- So called *exponential model* has the forms

$$\mu_i = e^{\beta_0} x_{i1}^{\beta_1} x_{i2}^{\beta_2} * \cdots * x_{ip}^{\beta_p}, \quad (3.5a)$$

$$\log(\mu_i) = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \log(x_{i2}) + \cdots + \beta_p \log(x_{ip}). \quad (3.5b)$$

Example 3.1.

Consider the research problem where there was interest to determine whether glucose had an effect on single-cell Tetrahymena cell size when cells were grown in cell culture on the growth medium. The cell size was measured by the (average) diameter of cells $Y = \text{diameter}$. In research, explanatory variables were the concentration of Tetrahymena cells $X_1 = \text{conc}$ and $X_2 = \text{glucose}$ which was coded in the following way:

$$x_{i2} = \begin{cases} \text{yes} = 1, & \text{when glucose was added to growth medium in experiment } i, \\ \text{no} = 2, & \text{when glucose was not added to growth medium in experiment } i. \end{cases}$$

The dataset can be found on the file `tetrahymena.txt`.

The data contains diameter and concentration of Tetrahymena cells with and without glucose added to growth medium.

glucose: a numeric vector code, 1: yes, 2: no.

conc: a numeric vector, cell concentration (counts/ml).

diameter: a numeric vector, cell diameter (micrometre).

	glucose	conc	diameter
1	1	631000	21.2
2	1	592000	21.5
.			
50	2	13000	24.3
51	2	11000	24.2

Let us assume $Y_i \sim N(\mu_i, \sigma^2)$. Let us consider modeling the expected value μ_i by the following models

$$\mathcal{M}_{1|2_{\log}} : \log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \alpha_j, \quad \mathcal{M}_{1|2_{\text{exponential}}} : \mu_i = e^{\beta_0} x_{i1}^{\beta_1} e^{\alpha_j},$$

```
> model.log<-glm(diameter~conc+factor(glucose), family=gaussian(link="log"), data=data)
> summary(model.log)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.149e+00	9.424e-03	334.156	< 2e-16 ***
conc	-3.607e-07	2.964e-08	-12.170	2.80e-16 ***
factor(glucose)yes	6.836e-02	1.068e-02	6.398	6.18e-08 ***

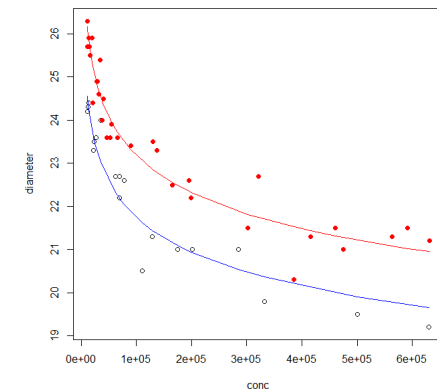
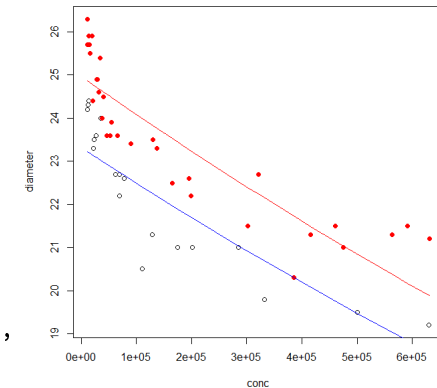
(Dispersion parameter for gaussian family taken to be 0.6971037)

```
> model.exponential<-glm(diameter~log(conc)+factor(glucose), family=gaussian(link="log"),
> summary(model.exponential)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.711625	0.025437	145.91	< 2e-16 ***
log(conc)	-0.054908	0.002259	-24.31	< 2e-16 ***
factor(glucose)yes	0.063899	0.006013	10.63	3.33e-14 ***

(Dispersion parameter for gaussian family taken to be 0.2215944)



3.1.2 Gamma and Inverse Gaussian Generalized Linear Models

- If the random variable Y_i follow the Gamma distribution $Y_i \sim \text{Gamma}(\mu_i, \phi)$, then for realization y_i it holds that $y_i > 0$. Furthermore

$$f(y_i|\mu_i, \nu) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu_i} \right)^\nu y_i^{\nu-1} e^{-\left(\frac{y_i \nu}{\mu_i}\right)}, \quad y_i > 0$$

$$E(Y_i) = \mu_i,$$

$$\text{Var}(Y_i) = \phi \mu_i^2,$$

where $\phi = \nu^{-1}$.

- Gamma distribution is suitable in modeling situations when the considered response variable can have only positive values and when the variance increases (proportionally to rate μ_i^2) same time as the expected value μ_i increases.
- Under Gamma distribution, usually used link functions $g(\mu_i)$ are

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \textit{identity link}, \quad (3.6a)$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \textit{log link}, \quad (3.6b)$$

$$\frac{1}{\mu_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \textit{inverse link}. \quad (3.6c)$$

-
- If the random variable Y_i follow the **Inverse Gaussian** distribution $Y_i \sim IG(\mu_i, \phi)$, then for realization y_i it holds that $y_i > 0$. Furthermore

$$f(y_i|\mu_i, \gamma) = \sqrt{\frac{\gamma}{2\pi y_i^3}} \exp\left(\frac{-\gamma(y_i - \mu_i)^2}{2\mu_i^2 y_i}\right), \quad y_i > 0$$

$$E(Y_i) = \mu_i,$$

$$\text{Var}(Y_i) = \phi\mu_i^3,$$

where $\phi = \gamma^{-1}$.

- Inverse Gaussian distribution is suitable in modeling situations when the considered response variable can have only positive values and when the variance increases (proportionally to rate μ_i^3) same time as the expected value μ_i increases. Inverse Gaussian distribution is very much competing distribution for the Gamma distribution.
- Under Inverse Gaussian distribution, usually used link functions $g(\mu_i)$ are

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \textit{identity link}, \quad (3.7a)$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \textit{log link}, \quad (3.7b)$$

$$\frac{1}{\mu_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \textit{inverse link}, \quad (3.7c)$$

$$\frac{1}{\mu_i^2} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \textit{canonical link}. \quad (3.7d)$$

Example 3.2.

In biodiesel study, methyl ester was produced from waste canola oil. In experiments, it was measured what kind of effect the factors $X_1 = \text{Time (15,30,45min)}$, $X_2 = \text{Temperature (240,255,270C)}$, and level of Methanol/Oil weight ratio (1,1.5,2), $X_3 = \text{Methanol}$, have on yield of methyl ester, $Y = \text{Yield}$. Data obtained from experiments is available in a file canoladiesel.txt.

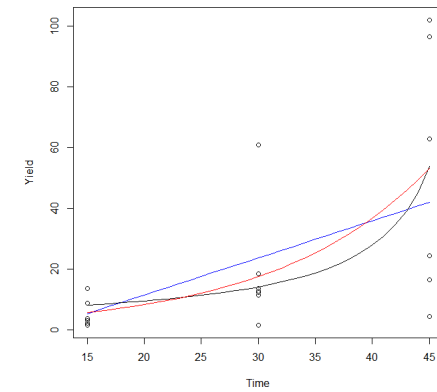
	Time	Temp	Methanol	Yield
1	15	240	1.0	1.5
2	15	240	1.5	3.2
3	15	240	2.0	3.8
.				
18	45	270	1.0	96.4
19	45	270	2.0	102.0

Source: S. Lee, D. Posarac, N. Ellis (2012). "An Experimental Investigation of Biodiesel Synthesis from Waste Canola Oil Using Supercritical Methanol," Fuel, Vol. 91, pp. 229-237.

Let us assume $Y_i \sim \text{Gamma}(\mu_i, \phi)$, and let us consider the models

$$\mathcal{M}_{\text{identity}} : \quad \mu_i = \beta_0 + \beta_1 x_{i1}, \quad \mathcal{M}_{\text{inverse}} : \quad \frac{1}{\mu_i} = \beta_0 + \beta_1 x_{i1}, \quad \mathcal{M}_{\text{log}} : \quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1}.$$

```
> model.identity<-glm(Yield~Time, family=Gamma(link="identity"), data=data)
> coef(model.identity)
(Intercept)      Time
-12.912377    1.219704
> model.inverse<-glm(Yield~Time, family=Gamma(link="inverse"), data=data)
> coef(model.inverse)
(Intercept)      Time
0.174318305 -0.003460192
> model.log<-glm(Yield~Time, family=Gamma(link="log"), data=data)
> coef(model.log)
(Intercept)      Time
0.63975713  0.07409982
```



3.1.3 Beta Distribution Models

- If the random variable Y_i follow the beta distribution $Y_i \sim \text{Beta}(\mu_i, \phi)$, then for realization y_i it holds that $0 < y_i < 1$ and density function has the form

$$f(y|\mu_i, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_i\phi)\Gamma((1-\mu_i)\phi)} y_i^{\mu_i\phi-1} (1-y_i)^{(1-\mu_i)\phi-1}, \quad (3.8)$$

where $0 < \mu_i < 1$ and $\phi > 0$.

- Furthermore

$$E(Y_i) = \mu_i, \quad (3.9)$$

$$\text{Var}(Y_i) = \frac{\mu_i(1-\mu_i)}{1+\phi}. \quad (3.10)$$

- Beta distribution is suitable in modeling situations when the considered response variable can naturally have values between some open interval (a, b) .
- Since beta distribution is defined on interval $(0, 1)$, it is common that transformation $\frac{y-a}{b-a}$ is done on the original response variable.
- Most often used link function with beta distribution is the logit link function

$$\text{logit}(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \quad \text{logit link.} \quad (3.11)$$

Example 3.3.

The coffee company had developed a new dark coffee that the company wanted to explore its tastefulness with the help of sensory analysis. In the study, consumers were given a sample of coffee, after tasting, the consumers were asked to evaluate how tasteful the coffee was on the scale of 0 to 100. The consumers who were selected for the survey were also asked about what level of roasting they usually like their coffee to have. The dataset can be on the file `coffeerating.txt`.

	age	gender	flavor	rating
1	33	female	light	40
2	32	female	dark	87
3	49	female	medium	86
4	39	male	light	36
.				
324	46	female	medium	88
325	42	male	medium	71

By applying Beta model, $Y_i \sim \text{Beta}(\mu_i, \phi)$,

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_h$$

we obtain the following parameter estimates.

```
> model.main<-betareg(Y~age+factor(gender)+factor(flavor), data=data, link=c("logit"))
> summary(model.main)
Call:
betareg(formula = Y ~ age + factor(gender) + factor(flavor), data = data,
        link = c("logit"))

Standardized weighted residuals 2:
      Min       1Q   Median       3Q      Max 
-2.6086 -0.6613 -0.0303  0.5190  4.8256 

Coefficients (mean model with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.258107   0.066829   18.83  <2e-16 ***
age             0.035037   0.001408   24.89  <2e-16 ***
factor(gender)male -0.516011  0.021166  -24.38  <2e-16 ***
factor(flavor)light -2.796063  0.038925  -71.83  <2e-16 ***
factor(flavor)medium -1.397279  0.038970  -35.85  <2e-16 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)   163.71     12.82   12.77  <2e-16 ***

Type of estimator: ML (maximum likelihood)
Log-likelihood: 663.2 on 6 Df
Pseudo R-squared: 0.9475
Number of iterations: 27 (BFGS) + 3 (Fisher scoring)
```

3.1.4 Count and Categorical Data Models

- In count data models, the random response variable Y_i can have a realization as a 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 non-negative integer value $y_i = \{0, 1, 2, 3, 4, \dots\}$, and $E(Y_i) = \mu_i$ and $Var(Y_i) = \mu_i$.
- Under Poisson distribution, possible link functions $g(\mu_i)$ are

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad \textit{identity link}, \quad (3.12a)$$

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad \textit{log link}, \quad (3.12b)$$

$$\sqrt{\mu_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}, \quad \textit{square root link}. \quad (3.12c)$$

-
- In categorical data models, the most simple situation is 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

$$P(Y_i = 1) = \mu_i, \quad P(Y_i = 0) = 1 - \mu_i. \quad (3.13)$$

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

$$E(Y_i) = \mu_i, \quad \text{Var}(Y_i) = \mu_i(1 - \mu_i). \quad (3.14)$$

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

$$\text{logit}(\mu_i) = \log \left(\frac{\mu_i}{1 - \mu_i} \right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}. \quad (3.15)$$

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

$$\mu_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}}. \quad (3.16)$$

3.2 Maximum Likelihood Estimation

3.2.1 Exponential Family of Distributions

- The distribution of a random variable Y_i belongs to the exponential family if its probability density function can be written in the form

$$f(y_i|\Theta_i, \phi) = \exp \left(\frac{y_i \Theta_i - b(\Theta_i)}{a(\phi)} + c(y_i, \phi) \right), \quad (3.17)$$

where Θ_i is the natural or canonical parameter, ϕ is the dispersion parameter, and a , b and c are specific functions.

Example 3.4.

For the normal distribution $Y_i \sim N(\mu_i, \sigma^2)$ it holds that

$$\begin{aligned} f(y_i|\mu_i, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(-\frac{1}{2} \cdot \frac{(y_i - \mu_i)^2}{\sigma^2} \right) = \exp \left(-\frac{1}{2} \cdot \frac{(y_i^2 - 2y_i\mu_i + \mu_i^2)}{\sigma^2} + \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \right) \\ &= \exp \left(\frac{y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} + \log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \right) = \exp \left(\frac{y_i\Theta_i - b(\Theta_i)}{a(\phi)} + c(y_i, \phi) \right), \end{aligned}$$

$$\Theta_i = \mu_i, \quad \phi = \sigma^2, \quad a(\phi) = \phi, \quad b(\Theta_i) = \frac{1}{2}\Theta_i^2, \quad c(y_i, \phi) = -\frac{y_i^2}{2\phi^2} + \log \left(\frac{1}{\sqrt{2\pi\phi^2}} \right).$$

-
- Let us consider now what kind properties the log-likelihood of the single random variable Y_i has when Y_i is belonging to the exponential family of distributions.
 - Log-likelihood $l(\Theta_i, \phi) = \log(L(\Theta_i, \phi))$ for the random variable Y_i has when Y_i is belonging to the exponential family of distributions has the form

$$l(\Theta_i, \phi) = \log(f(y_i|\Theta_i, \phi)) = \frac{y_i\Theta_i - b(\Theta_i)}{a(\phi)} + c(y_i, \phi), \quad (3.18)$$

and the the partial derivative with respect to canonical parameter Θ_i is

$$\frac{\partial l(\Theta_i, \phi)}{\partial \Theta_i} = \frac{y_i - b'(\Theta_i)}{a(\phi)}. \quad (3.19)$$

- Generally, for any log-likelihood, the partial derivative with respect to unknown parameter Θ_i has the property

$$\frac{\partial l}{\partial \Theta_i} = \frac{\partial \log(L)}{\partial \Theta_i} = \frac{1}{L} \frac{\partial L}{\partial \Theta_i}, \quad (3.20)$$

and hence

$$\frac{\partial l}{\partial \Theta_i} \cdot L = \frac{\partial L}{\partial \Theta_i}. \quad (3.21)$$

-
- Thus the expected value of the partial derivative $\frac{\partial l}{\partial \Theta_i}$ is (under the certain regularity conditions holding in the case of the exponential family)

$$\begin{aligned} E\left(\frac{\partial l}{\partial \Theta_i}\right) &= \int_{-\infty}^{\infty} \frac{\partial l}{\partial \Theta_i} \cdot f(y_i|\Theta_i, \phi) dy_i = \int_{-\infty}^{\infty} \frac{\partial l}{\partial \Theta_i} \cdot L dy_i = \int_{-\infty}^{\infty} \frac{\partial L}{\partial \Theta_i} dy_i \\ &= \frac{\partial}{\partial \Theta_i} \int_{-\infty}^{\infty} L dy_i = \frac{\partial}{\partial \Theta_i} \int_{-\infty}^{\infty} f(y_i|\Theta_i, \phi) dy_i = \frac{\partial}{\partial \Theta_i} \cdot 1 = 0. \end{aligned} \quad (3.22)$$

- By applying (3.22) to the exponential family of distributions, we get

$$E\left(\frac{\partial l(\Theta_i, \phi)}{\partial \Theta_i}\right) = \frac{E(Y_i) - b'(\Theta_i)}{a(\phi)} = 0, \quad (3.23)$$

and hence it must hold that

$$E(Y_i) = \mu_i = b'(\Theta_i). \quad (3.24)$$

- Furthermore, the second partial derivative with respect to unknown parameter Θ_i has the property

$$\frac{\partial^2 l}{\partial \Theta_i^2} = \frac{\partial}{\partial \Theta_i} \left(\frac{1}{L} \frac{\partial L}{\partial \Theta_i} \right) = -\frac{1}{L^2} \left(\frac{\partial L}{\partial \Theta_i} \right)^2 + \frac{1}{L} \frac{\partial^2 L}{\partial \Theta_i^2} = -\left(\frac{\partial l}{\partial \Theta_i} \right)^2 + \frac{1}{L} \frac{\partial^2 L}{\partial \Theta_i^2}. \quad (3.25)$$

-
- Thus the variance of the partial derivative $\frac{\partial l}{\partial \Theta_i}$ is

$$\begin{aligned}
\text{Var} \left(\frac{\partial l}{\partial \Theta_i} \right) &= \text{E} \left[\left(\frac{\partial l}{\partial \Theta_i} \right)^2 \right] = -\text{E} \left(\frac{\partial^2 l}{\partial \Theta_i^2} \right) + \text{E} \left(\frac{1}{L} \frac{\partial^2 L}{\partial \Theta_i^2} \right) = -\text{E} \left(\frac{\partial^2 l}{\partial \Theta_i^2} \right) + \int_{-\infty}^{\infty} \frac{1}{L} \frac{\partial^2 L}{\partial \Theta_i^2} \cdot L \, dy_i \\
&= -\text{E} \left(\frac{\partial^2 l}{\partial \Theta_i^2} \right) + \int_{-\infty}^{\infty} \frac{\partial^2 L}{\partial \Theta_i^2} \, dy_i = -\text{E} \left(\frac{\partial^2 l}{\partial \Theta_i^2} \right) + \frac{\partial^2}{\partial \Theta_i^2} \int_{-\infty}^{\infty} L \, dy_i \\
&= -\text{E} \left(\frac{\partial^2 l}{\partial \Theta_i^2} \right). \tag{3.26}
\end{aligned}$$

- For the exponential family of distributions, we have

$$\frac{\partial^2 l(\Theta_i, \phi)}{\partial \Theta_i^2} = \frac{-b''(\Theta_i)}{a(\phi)}, \quad \text{and} \quad \left(\frac{\partial l(\Theta_i, \phi)}{\partial \Theta_i} \right)^2 = \frac{(y_i - b'(\Theta_i))^2}{a(\phi)^2}. \tag{3.27}$$

- By applying the result of (3.26), i.e.,

$$\text{E} \left[\left(\frac{\partial l(\Theta_i, \phi)}{\partial \Theta_i} \right)^2 \right] = -\text{E} \left(\frac{\partial^2 l(\Theta_i, \phi)}{\partial \Theta_i^2} \right), \tag{3.28}$$

we have

$$\frac{\text{E}(Y_i - b'(\Theta_i))^2}{a(\phi)^2} = \frac{b''(\Theta_i)}{a(\phi)}. \tag{3.29}$$

-
- Thus the variance of the random variable Y_i is

$$\text{Var}(Y_i) = E(Y_i - b'(\Theta_i))^2 = b''(\Theta_i)a(\phi). \quad (3.30)$$

- Above we have proved the following theorem.

Theorem 3.1. If the density function of the random variable Y_i has the form

$$f(y_i|\Theta_i, \phi) = \exp \left(\frac{y_i\Theta_i - b(\Theta_i)}{a(\phi)} + c(y_i, \phi) \right),$$

then the expected value of Y_i is

$$E(Y_i) = \mu_i = b'(\Theta_i).$$

and the variance of Y_i is

$$\text{Var}(Y_i) = E(Y_i - b'(\Theta_i))^2 = b''(\Theta_i)a(\phi).$$

- In generalized linear models, the distribution of the random variable Y_i is assumed to be belonging to the exponential family of distributions and the canonical parameter Θ_i is the function of \mathbf{x}_i and $\boldsymbol{\beta}$ through the expected value μ_i , i.e, $\Theta_i(\mu_i(\mathbf{x}_i, \boldsymbol{\beta}))$, where $g(\mu_i) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}$.

3.2.2 Newton–Raphson and Fisher Scoring Methods

- Let assume the random vector \mathbf{y} is belonging to the exponential family of distributions, and let us consider the generalized linear model

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}, \quad (3.31)$$

where $g(\boldsymbol{\mu}) = (g(\mu_1), g(\mu_2), \dots, g(\mu_n))'$.

- Log-likelihood function of \mathbf{y} then is

$$l(\boldsymbol{\Theta}(\boldsymbol{\beta}), \phi) = l(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^n \log(f(y_i|\Theta_i, \phi)) = \sum_{i=1}^n \frac{y_i\Theta_i - b(\Theta_i)}{a(\phi)} + \sum_{i=1}^n c(y_i, \phi). \quad (3.32)$$

- To differentiate the log-likelihood function, we use the chain rule

$$\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_j} = \frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \Theta_i} \frac{\partial \Theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}. \quad (3.33)$$

- Since

$$\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \Theta_i} = \frac{y_i - b'(\Theta_i)}{a(\phi)}, \quad \mu_i = b'(\Theta_i), \quad \frac{\partial \Theta_i}{\partial \mu_i} = \frac{1}{b''(\Theta_i)} \quad \text{Var}(Y_i) = b''(\Theta_i)a(\phi), \quad \frac{\partial \eta_i}{\partial \beta_j} = x_{ij},$$

the partial derivatives $\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_j}$ simplifies to the form

$$\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_j} = \sum_{i=1}^n \frac{(y_i - \mu_i)}{\text{Var}(Y_i)} \cdot x_{ij} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i} \right), \quad j = 0, 1, 2, \dots, p. \quad (3.34)$$

-
- The partial derivative $\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_j}$ can be written as

$$\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_j} = \mathbf{x}'_{(j)} \mathbf{D} \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}), \quad j = 0, 1, 2, \dots, p, \quad (3.35)$$

where

$$\mathbf{D} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \eta_1} & 0 & \dots & 0 \\ 0 & \frac{\partial \mu_2}{\partial \eta_2} & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \frac{\partial \mu_n}{\partial \eta_n} \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} \text{Var}(Y_1) & 0 & \dots & 0 \\ 0 & \text{Var}(Y_2) & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \text{Var}(Y_n) \end{pmatrix}$$

and $\mathbf{x}_{(j)}$ is the j th column of the model matrix \mathbf{X} .

- When the partial derivatives $\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_j}$ are stacked as

$$\mathbf{u}_{\boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} \\ \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1} \\ \vdots \\ \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_p} \end{pmatrix},$$

then the partial derivatives (the score function) can be written as

$$\mathbf{u}_{\boldsymbol{\beta}} = \frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{X}' \mathbf{D} \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}). \quad (3.36)$$

-
- The score function $\mathbf{u}_\beta = \mathbf{X}'\mathbf{D}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ depends on the parameters β implicitly through $\boldsymbol{\mu}$. If $\hat{\beta}$ is a such estimator of β that the *likelihood equations*

$$\mathbf{u}_{\hat{\beta}} = \mathbf{X}'\hat{\mathbf{D}}\hat{\mathbf{V}}^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}, \quad (3.37)$$

are holding, then $\hat{\beta}$ is the maximum likelihood estimator of β .

- Usually no a close form solution for the likelihood equations $\mathbf{X}'\mathbf{D}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}$ exists with respect to the parameter vector β , and hence the maximum likelihood estimate $\hat{\beta}$ needs to be numerically solved.
- Newton–Raphson and Fisher Scoring methods are numerical iterative algorithms to solve the maximum likelihood estimate of β . Both methods are based on the second degree Taylor series expansion of the log-likelihood function:

$$l(\beta|\mathbf{y}) \approx l(\beta_t|\mathbf{y}) + \mathbf{u}'_t(\beta - \beta_t) + \frac{1}{2}(\beta - \beta_t)'\mathbf{H}_t(\beta - \beta_t), \quad (3.38)$$

where the subscript t denotes the approximation of the considered vector or matrix in the iterative process $t = 0, 1, 2, \dots$, and the matrix \mathbf{H} is the Hessian matrix

$$\mathbf{H} = \left(\frac{\partial^2 l(\beta|\mathbf{y})}{\partial \beta \partial \beta'} \right) = \begin{pmatrix} \frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_0} & \frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_0 \partial \beta_p} \\ \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_0} & \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_1} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_1 \partial \beta_p} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_0} & \frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial^2 l(\beta)}{\partial \beta_p \partial \beta_p} \end{pmatrix}.$$

-
- In Newton–Raphson method, the partial derivatives $\frac{\partial l(\beta|\mathbf{y})}{\partial \beta}$ of the Taylor series approximation of the log-likelihood function is obtained and set to zero:

$$\frac{\partial l(\beta|\mathbf{y})}{\partial \beta} \approx \mathbf{u}_t + \mathbf{H}_t(\beta - \beta_t) = \mathbf{0}. \quad (3.39)$$

- After initial values β_0 , the next updated values in Newton–Raphson method are obtained from the equation

$$\beta_{t+1} = \beta_t - \mathbf{H}_t^{-1} \mathbf{u}_t. \quad (3.40)$$

- Iterative process in Fisher Scoring method is the same as in Newton–Raphson method except the Hessian matrix \mathbf{H} is replaced by the Fisher information matrix $\mathbf{F} = -\mathbf{E}(\mathbf{H})$:

$$\beta_{t+1} = \beta_t + \mathbf{F}_t^{-1} \mathbf{u}_t, \quad (3.41)$$

- Both methods continue as long as the difference $\beta_{t+1} - \beta_t$ is sufficiently close to zero, which happens when the score function \mathbf{u}_t is sufficiently close to zero. Then the maximum likelihood estimate of β is $\hat{\beta} = \beta_{t+1}$.

3.2.3 Weighted Least Squares Estimation

- Let $\hat{\beta}$ be the maximum likelihood estimator for β in the generalized linear model $g(\mu) = \eta = \mathbf{X}\beta$.
- Based on the Fisher Scoring method, we may assume that the difference between the true unknown parameter vector β and $\hat{\beta}$ is (at least approximately) equal to

$$\hat{\beta} - \beta = \mathbf{F}_{\hat{\beta}}^{-1} \mathbf{u}_{\hat{\beta}}. \quad (3.42)$$

- Asymptotically as the sample size is increasing, we may assume that the estimated information matrix $\mathbf{F}_{\hat{\beta}}$ becomes a constant \mathbf{F}_{β} . Furthermore, as $n \rightarrow \infty$, the estimated score function becomes a zero vector $\mathbf{u}_{\hat{\beta}} = \mathbf{0}$.
- Hence, asymptotically, the expected value of the difference $\hat{\beta} - \beta$ is

$$\mathbb{E}(\hat{\beta} - \beta) = \mathbb{E}(\mathbf{F}_{\hat{\beta}}^{-1} \mathbf{u}_{\hat{\beta}}) = \mathbf{F}_{\beta}^{-1} \mathbb{E}(\mathbf{u}_{\hat{\beta}}) = \mathbf{F}_{\beta}^{-1} \cdot \mathbf{0} = \mathbf{0}. \quad (3.43)$$

- Asymptotically it also holds

$$\begin{aligned} \text{Cov}(\hat{\beta}) &= \mathbb{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = \mathbb{E}(\mathbf{F}_{\hat{\beta}}^{-1} \mathbf{u}_{\hat{\beta}} \mathbf{u}_{\hat{\beta}}' \mathbf{F}_{\hat{\beta}}^{-1}) = \mathbf{F}_{\beta}^{-1} \mathbb{E}(\mathbf{u}_{\hat{\beta}} \mathbf{u}_{\hat{\beta}}') \mathbf{F}_{\beta}^{-1} \\ &= \mathbf{F}_{\beta}^{-1} \mathbf{F}_{\hat{\beta}} \mathbf{F}_{\beta}^{-1} = \mathbf{F}_{\beta}^{-1} \mathbf{F}_{\beta} \mathbf{F}_{\beta}^{-1} = \mathbf{F}_{\beta}^{-1}. \end{aligned} \quad (3.44)$$

-
- Combining (3.43) and (3.44) together with fact that the maximum likelihood estimators are asymptotically generally following the normal distribution, the estimator $\hat{\beta}$ follows asymptotically the normal distribution

$$\hat{\beta} \sim N(\beta, \mathbf{F}_{\beta}^{-1}). \quad (3.45)$$

- Further, since $\mathbf{F}_{\beta} = E(\mathbf{u}_{\beta}\mathbf{u}_{\beta}')$, we have

$$\mathbf{F}_{\beta} = E(\mathbf{X}'\mathbf{D}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'\mathbf{V}^{-1}\mathbf{D}\mathbf{X}) \quad (3.46)$$

$$= \mathbf{X}'\mathbf{D}\mathbf{V}^{-1} E((\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})') \mathbf{V}^{-1}\mathbf{D}\mathbf{X} \quad (3.47)$$

$$= \mathbf{X}'\mathbf{D}\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{D}\mathbf{X} = \mathbf{X}'\mathbf{D}\mathbf{V}^{-1}\mathbf{D}\mathbf{X} = \mathbf{X}'\mathbf{W}\mathbf{X}, \quad (3.48)$$

where

$$\mathbf{W} = \mathbf{D}\mathbf{V}^{-1}\mathbf{D} = \begin{pmatrix} \frac{\left(\frac{\partial \mu_1}{\partial \eta_1}\right)^2}{\text{Var}(Y_1)} & 0 & \dots & 0 \\ 0 & \frac{\left(\frac{\partial \mu_2}{\partial \eta_2}\right)^2}{\text{Var}(Y_2)} & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \frac{\left(\frac{\partial \mu_n}{\partial \eta_n}\right)^2}{\text{Var}(Y_n)} \end{pmatrix}$$

- Hence, asymptotically

$$\hat{\beta} \sim N(\beta, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}). \quad (3.49)$$

-
- The Fisher Scoring process $\beta_{t+1} = \beta_t + \mathbf{F}_t^{-1} \mathbf{u}_t$ can be written also as

$$\mathbf{F}_t \beta_{t+1} = \mathbf{F}_t \beta_t + \mathbf{u}_t. \quad (3.50)$$

- Since $\mathbf{F}_t = \mathbf{X}' \mathbf{W}_t \mathbf{X}$ and $\mathbf{u}_t = \mathbf{X}' \mathbf{D}_t \mathbf{V}^{-1} (\mathbf{y} - \boldsymbol{\mu}_t) = \mathbf{X}' \mathbf{W}_t \mathbf{D}_t^{-1} (\mathbf{y} - \boldsymbol{\mu}_t)$, we have

$$\begin{aligned} \mathbf{X}' \mathbf{W}_t \mathbf{X} \beta_{t+1} &= \mathbf{X}' \mathbf{W}_t (\mathbf{X} \beta_t + \mathbf{D}_t^{-1} (\mathbf{y} - \boldsymbol{\mu}_t)) \\ \mathbf{X}' \mathbf{W}_t \mathbf{X} \beta_{t+1} &= \mathbf{X}' \mathbf{W}_t \mathbf{z}_t, \end{aligned} \quad (3.51)$$

where $\mathbf{z}_t = \mathbf{X} \beta_t + \mathbf{D}_t^{-1} (\mathbf{y} - \boldsymbol{\mu}_t)$ is called as the *adjusted response variable*.

- Solving (3.51) with respect to β_{t+1} gives us the *weighted least squares estimate*

$$\beta_{t+1} = (\mathbf{X}' \mathbf{W}_t \mathbf{X})^{-1} \mathbf{X}' \mathbf{W}_t \mathbf{z}_t. \quad (3.52)$$

- In weighted least squares estimation, iterative process is continued as long as the difference $\beta_{t+1} - \beta_t$ is sufficiently close to zero. The maximum likelihood estimate of β is $\hat{\beta} = \beta_{t+1}$.
- The maximum likelihood estimator $\hat{\beta}$ follows approximately the normal distribution

$$\hat{\beta} \sim N \left(\beta, (\mathbf{X}' \widehat{\mathbf{W}} \mathbf{X})^{-1} \right), \quad (3.53)$$

where $\widehat{\mathbf{W}}$ is the estimate of \mathbf{W} .

3.2.4 Estimation of $\text{Var}(Y_i)$

- In generalized linear models $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$, the variance of the random variable Y_i has often the form

$$\text{Var}(Y_i) = \phi \cdot v(\mu_i), \quad (3.54)$$

where ϕ is a positive unknown dispersion parameter and $v(\mu_i)$ is the function of unknown expected value μ_i .

- For example,

Normal distribution:	$v(\mu_i) = 1,$
Gamma distribution:	$v(\mu_i) = \mu_i^2,$
Inverse Gaussian distribution:	$v(\mu_i) = \mu_i^3,$
Poisson distribution:	$v(\mu_i) = \mu_i,$
Quasi-Poisson distribution:	$v(\mu_i) = \mu_i,$
Bernoulli distribution:	$v(\mu_i) = \mu_i(1 - \mu_i),$
Quasi-Bernoulli distribution:	$v(\mu_i) = \mu_i(1 - \mu_i).$

- For the Poisson and Bernoulli distribution, the dispersion parameter ϕ is equal to 1, but for all other distributions it needs to be estimated.

-
- The dispersion parameter ϕ is usually estimated by the Chisquared statistic

$$\tilde{\phi} = \frac{\sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}}{n - \text{rank}(\mathbf{X})} = \frac{X^2}{n - \text{rank}(\mathbf{X})}. \quad (3.55)$$

- When $\text{Var}(Y_i) = \phi v(\mu_i)$, the X^2 statistic has asymptotically the property

$$\sum_{i=1}^n \left(\frac{y_i - \hat{\mu}_i}{\sqrt{\phi v(\hat{\mu}_i)}} \right)^2 = \frac{1}{\phi} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} = \frac{1}{\phi} X^2 \sim \chi^2_{(n - \text{rank}(\mathbf{X}))}. \quad (3.56)$$

- Since

$$\text{E} \left(\frac{1}{\phi} X^2 \right) = \frac{1}{\phi} \text{E} (X^2) = n - \text{rank}(\mathbf{X}), \quad (3.57)$$

the estimator $\tilde{\phi}$ is the unbiased estimator of ϕ :

$$\text{E}(\tilde{\phi}) = \frac{\text{E} (X^2)}{n - \text{rank}(\mathbf{X})} = \phi. \quad (3.58)$$

- Note that, the estimate $\tilde{\phi}$ is calculated after the estimates $\hat{\mu}_i$ are obtained.
- The unbiased estimator of $\text{Var}(Y_i)$ is

$$\widehat{\text{Var}}(Y_i) = \tilde{\phi} v(\hat{\mu}_i). \quad (3.59)$$

Example 3.5.

Let us consider the model

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}),$$
$$\log(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{1}\beta_0.$$

The score function has the form

$$\begin{aligned} u_{\beta_0} &= \frac{\partial l(\beta_0|\mathbf{y})}{\partial \beta_0} = \sum_{i=1}^n \frac{(y_i - \mu_i)}{\text{Var}(Y_i)} \cdot x_{i0} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i} \right) = \sum_{i=1}^n \frac{(y_i - e^{\beta_0})}{\phi} \cdot 1 \cdot e^{\beta_0} \\ &= \frac{e^{\beta_0}}{\phi} \left[\sum_{i=1}^n (y_i - e^{\beta_0}) \right] = \frac{e^{\beta_0}}{\phi} \left[\left(\sum_{i=1}^n y_i \right) - n \cdot e^{\beta_0} \right], \end{aligned}$$

since

$$\text{Var}(Y_i) = \sigma^2 = \phi, \quad \frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial e^{\eta_i}}{\partial \eta_i} = e^{\eta_i} = e^{\beta_0}.$$

The likelihood equations

$$u_{\beta_0} = \frac{\partial l(\beta_0|\mathbf{y})}{\partial \beta_0} = \frac{e^{\beta_0}}{\phi} \left[\left(\sum_{i=1}^n y_i \right) - n \cdot e^{\beta_0} \right] = 0$$

can hold only if $(\sum_{i=1}^n y_i) - n \cdot e^{\beta_0} = 0$ holds (for all $\beta_0 > -\infty, \phi < \infty$). Solving the equation $(\sum_{i=1}^n y_i) - n \cdot e^{\beta_0} = 0$ with respect to e^{β_0} gives us the maximum likelihood estimate of μ_i as the sample mean $\hat{\mu}_i = e^{\hat{\beta}_0} = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$, and furthermore, the maximum likelihood estimate of β_0 is $\hat{\beta}_0 = \log(\bar{y})$.

The estimate for ϕ is

$$\tilde{\phi} = \frac{\sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}}{n - \text{rank}(\mathbf{X})} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1} = s_y^2,$$

and hence $\widehat{\text{Var}}(Y_i) = \tilde{\phi} = \tilde{\sigma}^2 = s_y^2$. The estimate of the matrix \mathbf{W} is

$$\widehat{\mathbf{W}} = \begin{pmatrix} \frac{\left(\frac{\partial \hat{\mu}_1}{\partial \hat{\eta}_1}\right)^2}{\widehat{\text{Var}}(Y_1)} & 0 & \dots & 0 \\ 0 & \frac{\left(\frac{\partial \hat{\mu}_2}{\partial \hat{\eta}_2}\right)^2}{\widehat{\text{Var}}(Y_2)} & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \frac{\left(\frac{\partial \hat{\mu}_n}{\partial \hat{\eta}_n}\right)^2}{\widehat{\text{Var}}(Y_n)} \end{pmatrix} = \begin{pmatrix} \frac{\bar{y}^2}{\tilde{\sigma}^2} & 0 & \dots & 0 \\ 0 & \frac{\bar{y}^2}{\tilde{\sigma}^2} & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \frac{\bar{y}^2}{\tilde{\sigma}^2} \end{pmatrix} = \frac{\bar{y}^2}{\tilde{\sigma}^2} \mathbf{I},$$

and thus the estimate of the covariance matrix $\text{Cov}(\hat{\beta}_0)$ is

$$\widehat{\text{Cov}}(\hat{\beta}_0) = (\mathbf{1}' \widehat{\mathbf{W}} \mathbf{1})^{-1} = \left(\frac{\bar{y}^2}{\tilde{\sigma}^2} \mathbf{1}' \mathbf{1} \right)^{-1} = \left(\frac{\bar{y}^2 \cdot n}{\tilde{\sigma}^2} \right)^{-1} = \frac{1}{\bar{y}^2} \cdot \frac{\tilde{\sigma}^2}{n}.$$

3.3 Confidence and Prediction Intervals

3.3.1 Confidence Intervals in Generalized Linear Model

- The maximum likelihood estimator for the link function $g(\mu_{i_*}) = \eta_{i_*} = \mathbf{x}'_{i_*}\boldsymbol{\beta}$ is

$$\widehat{g(\mu_{i_*})} = \hat{\eta}_{i_*} = \mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}}. \quad (3.60)$$

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

$$\widehat{\text{Var}}\left(\mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}}\right) = \mathbf{x}'_{i_*}\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_{i_*} = \mathbf{x}'_{i_*}(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{x}_{i_*}. \quad (3.61)$$

- Since approximately

$$\mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}} \sim N(\mathbf{x}'_{i_*}\boldsymbol{\beta}, \mathbf{x}'_{i_*}(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{x}_{i_*}), \quad (3.62)$$

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

$$\left[\mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}} - z_{\alpha/2}\sqrt{\mathbf{x}'_{i_*}\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_{i_*}}, \mathbf{x}'_{i_*}\hat{\boldsymbol{\beta}} + z_{\alpha/2}\sqrt{\mathbf{x}'_{i_*}\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{x}_{i_*}} \right] = [L_{\alpha/2}, U_{\alpha/2}], \quad (3.63)$$

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

- If $g(\mu_{i_*})$ is monotonically increasing function, then $100(1 - \alpha)\%$ confidence interval for the expected value μ_{i_*} is

$$[g^{-1}(L_{\alpha/2}), g^{-1}(U_{\alpha/2})]. \quad (3.64)$$

3.3.2 Prediction Intervals in Generalized Linear Model

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

$$\hat{Y}_f = g^{-1}(\mathbf{x}'_f \hat{\boldsymbol{\beta}}). \quad (3.65)$$

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

$$\text{Var}(\hat{Y}_f) = \left(\frac{\partial \mu_f}{\partial \eta_f} \right)^2 \cdot \mathbf{x}'_f \text{Cov}(\hat{\boldsymbol{\beta}}) \mathbf{x}_f = \left(\frac{\partial \mu_f}{\partial \eta_f} \right)^2 \cdot \mathbf{x}'_f (\mathbf{X}' \mathbf{W} \mathbf{X})^{-1} \mathbf{x}_f. \quad (3.66)$$

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

$$\text{Var}(e_f) = \text{Var}(Y_f) + \text{Var}(\hat{Y}_f) = \text{Var}(Y_f) + \left(\frac{\partial \mu_f}{\partial \eta_f} \right)^2 \cdot \mathbf{x}'_f \text{Cov}(\hat{\boldsymbol{\beta}}) \mathbf{x}_f, \quad (3.67)$$

the estimated variance of the prediction error is

$$\widehat{\text{Var}}(e_f) = \widehat{\text{Var}}(Y_f) + \left(\frac{\partial \hat{\mu}_f}{\partial \hat{\eta}_f} \right)^2 \cdot \mathbf{x}'_f \widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) \mathbf{x}_f, \quad (3.68)$$

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

$$\left[g^{-1}(\mathbf{x}'_f \hat{\boldsymbol{\beta}}) - z_{\alpha/2} \sqrt{\widehat{\text{Var}}(e_f)}, g^{-1}(\mathbf{x}'_f \hat{\boldsymbol{\beta}}) + z_{\alpha/2} \sqrt{\widehat{\text{Var}}(e_f)} \right]. \quad (3.69)$$

-
- The prediction intervals for the new observation y_f (observable in future) with given values of the explanatory variables \mathbf{x}_f can also be obtained by the bootstrap method.

PARAMETRIC BOOTSTRAP BASED METHOD - NORMAL DISTRIBUTION

1. Find the estimates $\hat{\eta}_f = \mathbf{x}'_f \hat{\boldsymbol{\beta}}$ and $\tilde{\phi} = \tilde{\sigma}^2$.
2. Simulate $\hat{\eta}_{f*}$ from the distribution $\hat{\eta}_{f*} \sim N \left(\hat{\eta}_f, \mathbf{x}'_f (\mathbf{X}' \widehat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{x}_f \right)$.
3. Find the estimates $\hat{\mu}_{f*} = g^{-1}(\hat{\eta}_{f*})$.
4. Simulate y_{f*} from the distribution $y_{f*} \sim N(\hat{\mu}_{f*}, \tilde{\sigma}^2)$.
5. Repeat M times the steps 2-4, and then determine $\alpha/2$ and $1 - \alpha/2$ the quantiles of the simulated values y_{f*} .

PARAMETRIC BOOTSTRAP BASED METHOD - GAMMA DISTRIBUTION

1. Find the estimates $\hat{\eta}_f = \mathbf{x}'_f \hat{\boldsymbol{\beta}}$ and $\tilde{\phi}$.
2. Simulate $\hat{\eta}_{f*}$ from the distribution $\hat{\eta}_{f*} \sim N \left(\hat{\eta}_f, \mathbf{x}'_f (\mathbf{X}' \widehat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{x}_f \right)$.
3. Find the estimates $\hat{\mu}_{f*} = g^{-1}(\hat{\eta}_{f*})$.
4. Simulate y_{f*} from the distribution $y_{f*} \sim \text{Gamma}(\hat{\mu}_{f*}, \tilde{\phi})$. (Simulate y_{f*} from the distribution $y_{f*} \sim \text{Gamma}(a, s_*)$, where $a = \frac{1}{\tilde{\phi}}$ and $s_* = \tilde{\phi} \hat{\mu}_{f*}$.)
5. Repeat M times the steps 2-4, and then determine $\alpha/2$ and $1 - \alpha/2$ the quantiles of the simulated values y_{f*} .

PARAMETRIC BOOTSTRAP BASED METHOD - INVERSE GAUSSIAN DISTRIBUTION

1. Find the estimates $\hat{\eta}_f = \mathbf{x}'_f \hat{\boldsymbol{\beta}}$ and $\tilde{\phi}$.
2. Simulate $\hat{\eta}_{f*}$ from the distribution $\hat{\eta}_{f*} \sim N \left(\hat{\eta}_f, \mathbf{x}'_f (\mathbf{X}' \widehat{\mathbf{W}} \mathbf{X})^{-1} \mathbf{x}_f \right)$.
3. Find the estimates $\hat{\mu}_{f*} = g^{-1}(\hat{\eta}_{f*})$.

-
4. Simulate y_{f*} from the distribution $y_{f*} \sim IG(\hat{\mu}_{f*}, \tilde{\phi})$.
(Simulate y_{f*} from the distribution $y_{f*} \sim IG\left(\hat{\mu}_{f*}, \frac{1}{\tilde{\phi}}\right)$.)
 5. Repeat M times the steps 2-4, and then determine $\alpha/2$ and $1 - \alpha/2$ the quantiles of the simulated values y_{f*} .

Example 3.6.

Consider the following data set related to genetic disorder cystic fibrosis:

```
  age sex height weight bmp fev1  rv frc tlc pemax
1   7   0   109  13.1  68   32 258 183 137   95
2   7   1   112  12.9  65   19 449 245 134   85
3   8   0   124  14.1  64   22 441 268 147  100
.
25  23   0   179  71.5  95   52 225 127 101  195
```

The cystfibr data frame has 25 rows and 10 columns. It contains lung function data for cystic fibrosis patients (7-23 years old).

This data frame contains the following columns:

age - age in years.

sex - 0: male, 1:female.

height - height (cm).

weight - weight (kg).

bmp - body mass (percent of normal).

fev1 - forced expiratory volume.

rv - residual volume.

frc - functional residual capacity.

tlc - total lung capacity.

pemax - maximum expiratory pressure.

O'Neill et al. (1983), The effects of chronic hyperinflation, nutritional status, and posture on respiratory muscle strength in cystic fibrosis, Am. Rev. Respir. Dis., 128:1051-1054.

We denote variables as $Y = \text{pemax}$, $X_1 = \text{height}$, $X_2 = \text{weight}$ $X_3 = \text{sex}$. Further, we assume $Y_i \sim N(\mu_i, \sigma^2)$. We consider the model

$$\mu_i = e^{\beta_0} x_{i1}^{\beta_1} x_{i2}^{\beta_2} e^{\alpha_j},$$

and let us find confidence and prediction intervals for Y_f when

	age	sex	height	weight	bmp	fev1	rv	frc	tlc
1	7	0	109	13.1	68	32	258	183	137

```
> data<-read.table("cysticfibrosis.txt", sep="\t", dec=".", header=TRUE)
> model.exponential<-glm(pemax~log(height)+log(weight)+factor(sex), family=gaussian(link="log"), data=data)
> newdata<-data[1,]
> pred<-predict(model.exponential, newdata=newdata, type="response")
> pred
      1
74.91101
> eta<-predict(model.exponential, newdata=newdata, type="link", se.fit=TRUE)
> link.lowerbound<-eta$fit-qnorm(0.975)*eta$se.fit
> link.upperbound<-eta$fit+qnorm(0.975)*eta$se.fit
> lower.mu<-exp(link.lowerbound)
> upper.mu<-exp(link.upperbound)
```

```
> lower.mu
      1
51.57218
> upper.mu
      1
108.8118

> pred<-predict(model.exponential, newdata=newdata, type="response")
> xf<-cbind(model.matrix(model.exponential)[1,])
> Var.Yf<-summary(model.exponential)$dispersion
> D.f<-pred
> Var.ef<-Var.Yf+(D.f^2)*t(xf)%*%vcov(model.exponential)%*%xf
> lower.yf<-pred-qnorm(0.9)*sqrt(Var.ef)
> upper.yf<-pred+qnorm(0.9)*sqrt(Var.ef)
> lower.yf
      [,1]
[1,] 35.30483
> upper.yf
      [,1]
[1,] 114.5172
```

Assignment 3.1.

Consider the normal distribution with the identity link function

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}),$$
$$\boldsymbol{\mu} = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}.$$

What kind of more simplified form the likelihood equations

$$\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}} = \mathbf{u}_{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{D}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0}$$

have in this case? Note that

$$\mathbf{D} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \eta_1} & 0 & \dots & 0 \\ 0 & \frac{\partial \mu_2}{\partial \eta_2} & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \frac{\partial \mu_n}{\partial \eta_n} \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} \text{Var}(Y_1) & 0 & \dots & 0 \\ 0 & \text{Var}(Y_2) & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \text{Var}(Y_n) \end{pmatrix}.$$

Can you obtain the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ directly by solving likelihood equations with respect to $\boldsymbol{\beta}$?

Assignment 3.2.

Consider the model

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}),$$
$$\log(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{1}\beta_0.$$

Construct the prediction intervals for the new observation Y_f .

Assignment 3.3.

In biodiesel study, methyl ester was produced from waste canola oil. In experiments, it was measured what kind of effect the factors $X_1 = \text{Time (15,30,45min)}$, $X_2 = \text{Temperature (240,255,270C)}$, and level of Methanol/Oil weight ratio (1,1.5,2), $X_3 = \text{Methanol}$, have on yield of methyl ester, $Y = \text{Yield}$. Data obtained from experiments is available in a file canoladiesel.txt.

	Time	Temp	Methanol	Yield
1	15	240	1.0	1.5
2	15	240	1.5	3.2
3	15	240	2.0	3.8
.				
18	45	270	1.0	96.4
19	45	270	2.0	102.0

Source: S. Lee, D. Posarac, N. Ellis (2012). "An Experimental Investigation of Biodiesel Synthesis from Waste Canola Oil Using Supercritical Methanol," Fuel, Vol. 91, pp. 229-237.

Let us consider the models

$$\mathcal{M}_{1_{\text{inverse}}} : \frac{1}{\mu_i} = \beta_0 + \beta_1 x_{i1}, \quad \mathcal{M}_{1_{\log}} : \log(\mu_i) = \beta_0 + \beta_1 x_{i1}.$$

Create prediction intervals under different distributional assumptions.

Assignment 3.4.

The coffee company had developed a new dark coffee that the company wanted to explore its tastefulness with the help of sensory analysis. In the study, consumers were given a sample of coffee, after tasting, the consumers were asked to evaluate how tasteful the coffee was on the scale of 0 to 100. The consumers who were selected for the survey were also asked about what level of roasting they usually like their coffee to have. The dataset can be on the file `coffeerating.txt`.

	age	gender	flavor	rating
1	33	female	light	40
2	32	female	dark	87
3	49	female	medium	86
4	39	male	light	36
.				
324	46	female	medium	88
325	42	male	medium	71

Let us assume $Y_i \sim \text{Beta}(\mu_i, \phi)$. Consider the model

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \alpha_j + \gamma_h + \tau_j x_{i1} + \omega_h x_{i1} + \psi_{jh}$$

Create prediction intervals by parametric bootstrap methods.
