

Generalized Linear Models and their Applications

STAT 431/STAT 831^{*}

Fall 2021 (1219)[†]

Cameron Roopnarine[‡]

Leilei Zeng[§]

26th October 2021

^{*}STAT 431 \equiv STAT 831

[†]Online Course

[‡]~~TeX~~

[§]Instructor

Contents

Topic 1a: Review of Linear Regression	3
Topic 1b: Review of Likelihood Methods	12
Topic 2a: Formulation of Generalized Linear Models	19
Topic 2a: Maximum Likelihood Estimation for Generalized Linear Models	24
Topic 3a: Binary Data and Odds Ratios	30
Topic 3b: Binomial Regression Models for Binary Data	34
Topic 3c: Likelihood Ratio Test for Logistic Regression Models	42
Topic 3d: Residuals for Binomial Data and Neuroblastoma Example	48
Topic 3e: Dose-Response Models	58
Topic 3f: Summary of Binomial Regression Models	67
Topic 4a: Poisson GLMs for Count Data	69

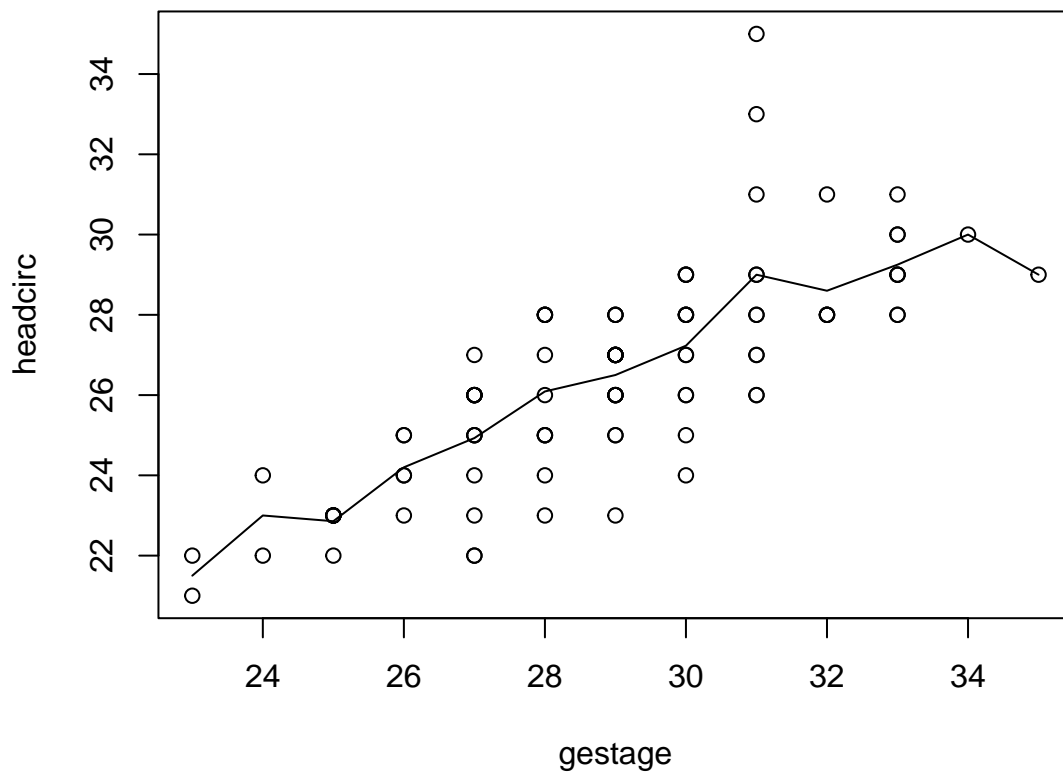
Topic 1a: Review of Linear Regression

Example: low birthweight infants study¹

A study was conducted at two teaching hospitals in Boston, Massachusetts, where the head circumference, gestational age and some other variables are recorded for 100 low birth weight infants.

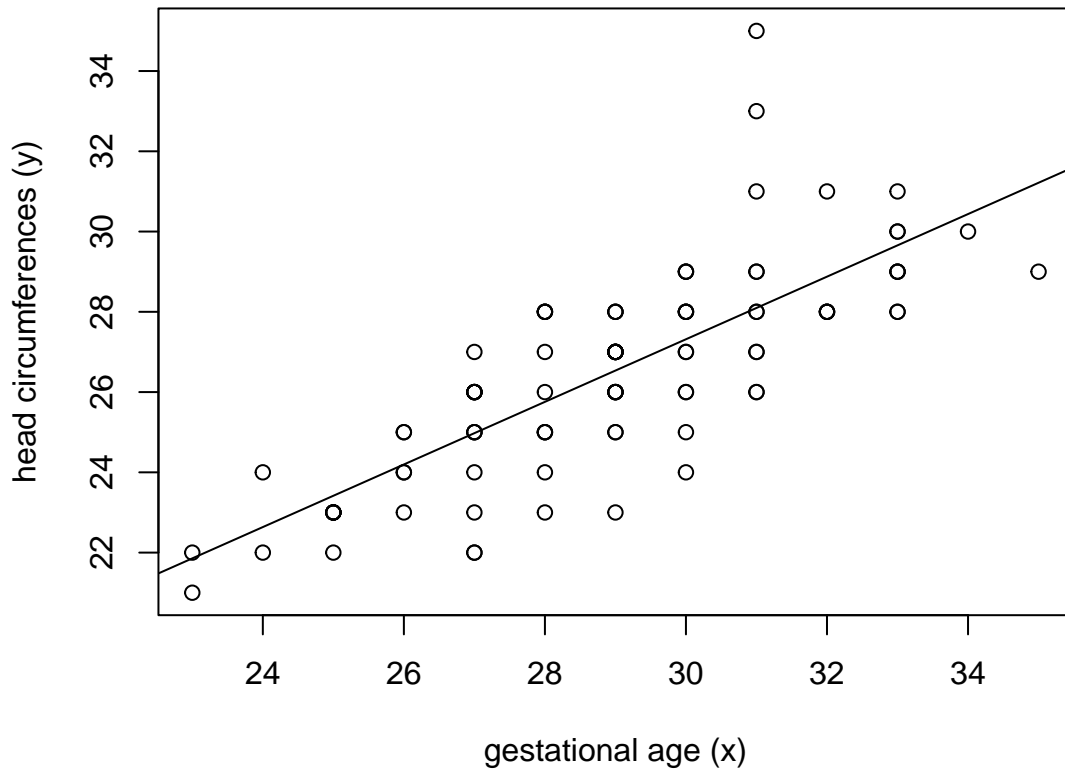
Question: what is the relationship between *gestational age* & *head circumference*?

A Scatterplot of the Data



We wish to model the relationship between *gestational age* and *head circumference* using a straight line!

¹Principles of Biostatistics 2nd Edition by Marcello Pagano, Kimberlee Gauvreau.



The Model Fitting Process

- ① **Model Specification:** select a probability distribution for the response variable and a linear equation linking the response to the explanatory variables.
- ② **Estimation:** finding the equation (the parameters of the model).
- ③ **Model checking:** how well does the model fit the data?
- ④ **Inference:** interpret the fitted model, calculate confidence intervals, conduct hypothesis tests.

① Model Specification

Notation

For each subject $i = 1, \dots, n$ we have:

- Y_i = random variable representing the response, and
- $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^\top$, a vector of explanatory variables.

Specification for Multiple Linear Regression

- Linear regression equation:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \text{ where } \varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

- Equivalently, Y_i 's are independent $\mathcal{N}(\mu_i, \sigma^2)$ random variables or

$$\mu_i = \mathbb{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

- For convenience, we often write linear regression models in matrix form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

and

$$\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

② Estimation

Least Squares Method

We wish to minimize a loss function:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2 \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

The least squares estimators (LSE) are the solutions to the equations:

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0.$$

Maximum Likelihood Method

The probability density function for Y_i is:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2} (y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}))^2\right\}.$$

The log-likelihood function is therefore:

$$\begin{aligned}\ell(\boldsymbol{\beta}, \sigma^2) &= \log\left(\prod_{i=1}^n f(y_i)\right) \\ &= \sum_{i=1}^n \left(-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (y_i - (\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}))^2\right) \\ &= -\frac{n}{2} \log(2\sigma^2) - \frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).\end{aligned}$$

The maximum likelihood estimators (MLE) of $\boldsymbol{\beta}$ are obtained by solving:

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \left[-\frac{1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \right] = 0.$$

- **Parameter Estimates:** For linear regression LSE and MLE of $\boldsymbol{\beta}$ are the same

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

- **Fitted values:** $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$.

- **Residuals:** $\hat{r}_i = (y_i - \hat{y}_i)$.

- **Variance estimates:**

- An unbiased estimate of σ^2 is:

$$\hat{\sigma}^2 = \frac{1}{n - (p + 1)} \sum_{i=1}^n \hat{r}_i^2.$$

- An estimate of the variance of $\hat{\boldsymbol{\beta}}$ is:

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}.$$

Low Birthweight Infant Data Example

- For $n = 100$ infants, we have observed $Y_i =$ head circumference and $x_i =$ gestational age for baby i , $i = 1, \dots, 100$.
- Consider a simple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

- We can fit the model and obtain LSE/MSE using the `lm()` function in R.

```
lowbwt <- read.table("lowbwt.txt", header = T)
fit <- lm(headcirc ~ gestage, data = lowbwt)
summary(fit)
```

```
Call:
lm(formula = headcirc ~ gestage, data = lowbwt)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-3.5358 -0.8760 -0.1458  0.9041  6.9041
```

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.91426    1.82915    2.14   0.0348 *
gestage      0.78005    0.06307   12.37  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.59 on 98 degrees of freedom
Multiple R-squared:  0.6095, Adjusted R-squared:  0.6055
F-statistic: 152.9 on 1 and 98 DF, p-value: < 2.2e-16

```

- What is the interpretation of regression parameters β_0 and β_1 ?
 - β_0 (intercept): expected headcirc for a baby of a gestational age zero ($x = 0$).
 - β_1 (slope): expected change in headcirc associated with a one unit increase in gestational age.

③ Model Checking

Standardized Residuals:

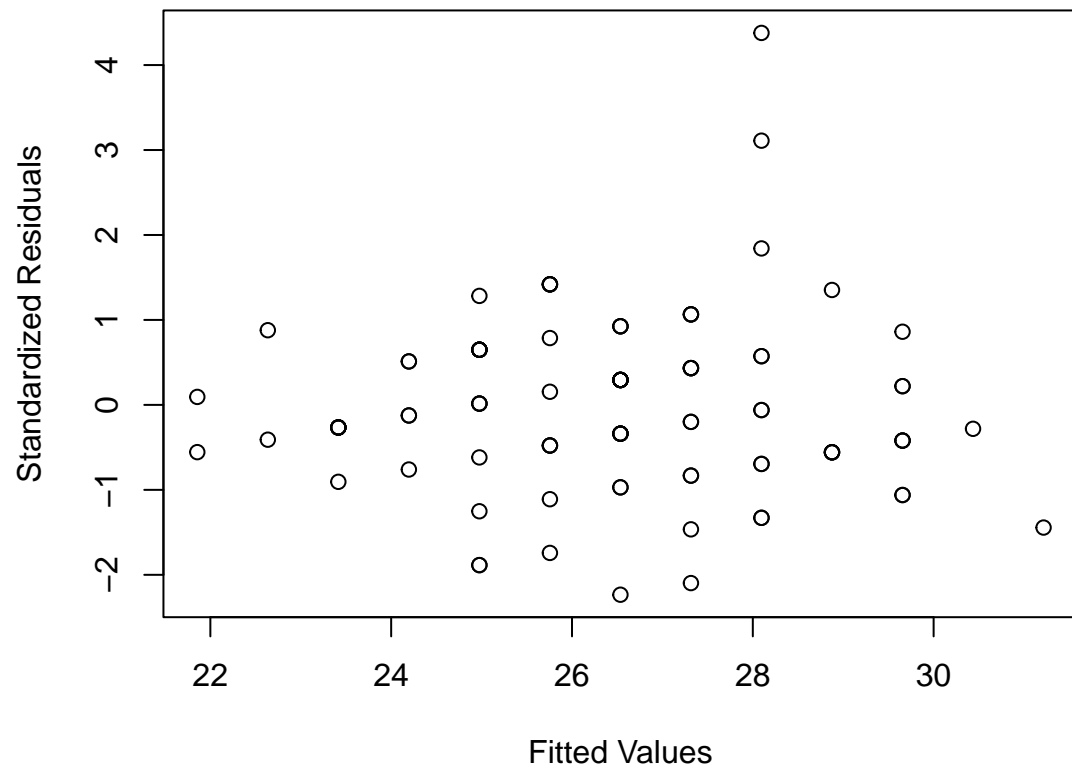
$$d_i = \frac{r_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}},$$

where h_{ii} is the (i, i) element of $\mathbf{H} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$. By asymptotic theory, if the model provides a good fit to the data then we should expect that:

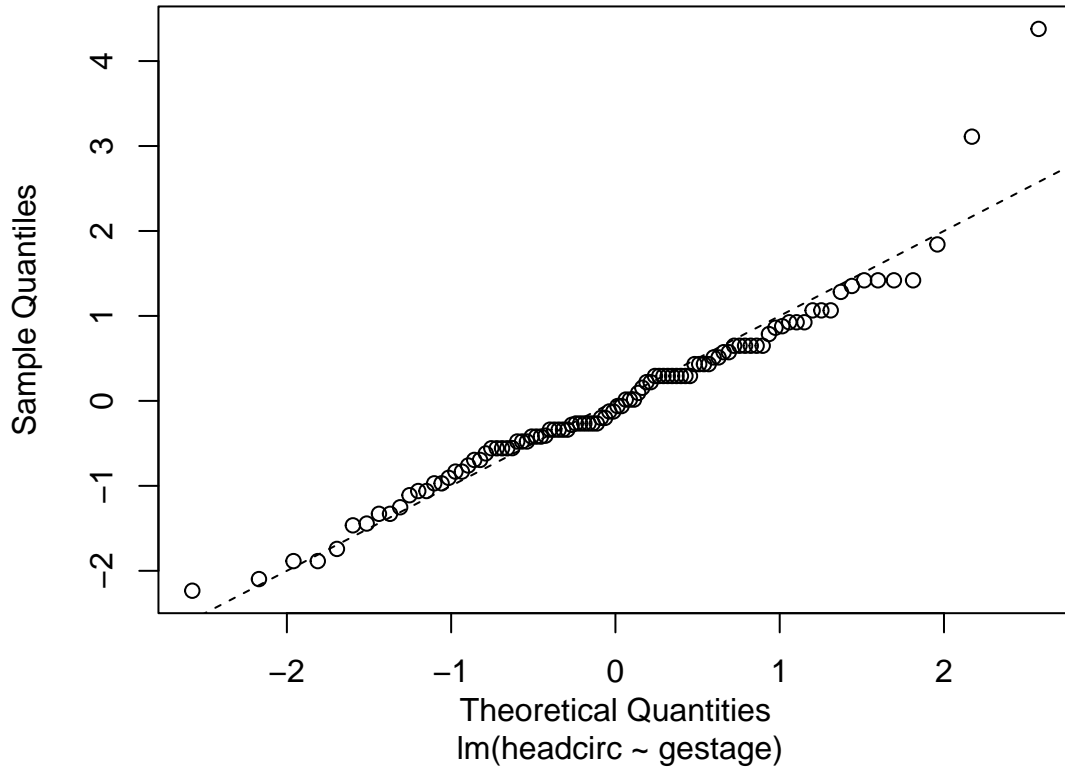
$$d_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1).$$

We visually check this by examining residual plots such as:

- Standardized residuals versus the fitted values.
- Standardized residuals versus the explanatory variable(s).
- Normal probability plot (QQ plot) of the standardized residuals.



Normal Q-Q Plot



④ Inference

- Under suitable assumptions, the fitted regression parameters are asymptotically normally distributed:

$$\begin{aligned}\hat{\beta} &\sim \text{MVN}(\beta, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}), \\ \hat{\beta}_j &\sim \mathcal{N}(\beta_j, \sigma^2 v_{jj}), \quad \text{where } v_{jj} = [(\mathbf{X}^\top \mathbf{X})^{-1}]_{(j,j)}.\end{aligned}$$

- Since σ^2 is generally unknown, we replace it with the unbiased estimate $\hat{\sigma}^2$, and obtain $\text{se}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 v_{jj}}$.
- The inference is then based on the t -distribution result:

$$\frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \sim t_{n-p-1}.$$

Low Birthweight Infant Data Example

- Is there a significant (linear) relationship between head circumference and gestational age?

We wish to test $H_0: \beta_1 = 0$ vs $H_A: \beta_1 \neq 0$.

$$t = \frac{\hat{\beta}_1 - (0)}{\text{se}(\hat{\beta}_1)} \sim t_{98},$$

if H_0 is true, and we reject H_0 if $|t| > t_{98,0.975} = 1.985$. Here we have $t = 0.78/0.063 = 12.37 \gg 1.985$, so we reject H_0 .

- What is the 95 % confidence interval for the expected increase in head circumference when the gestational age of a baby increases by 1 week?

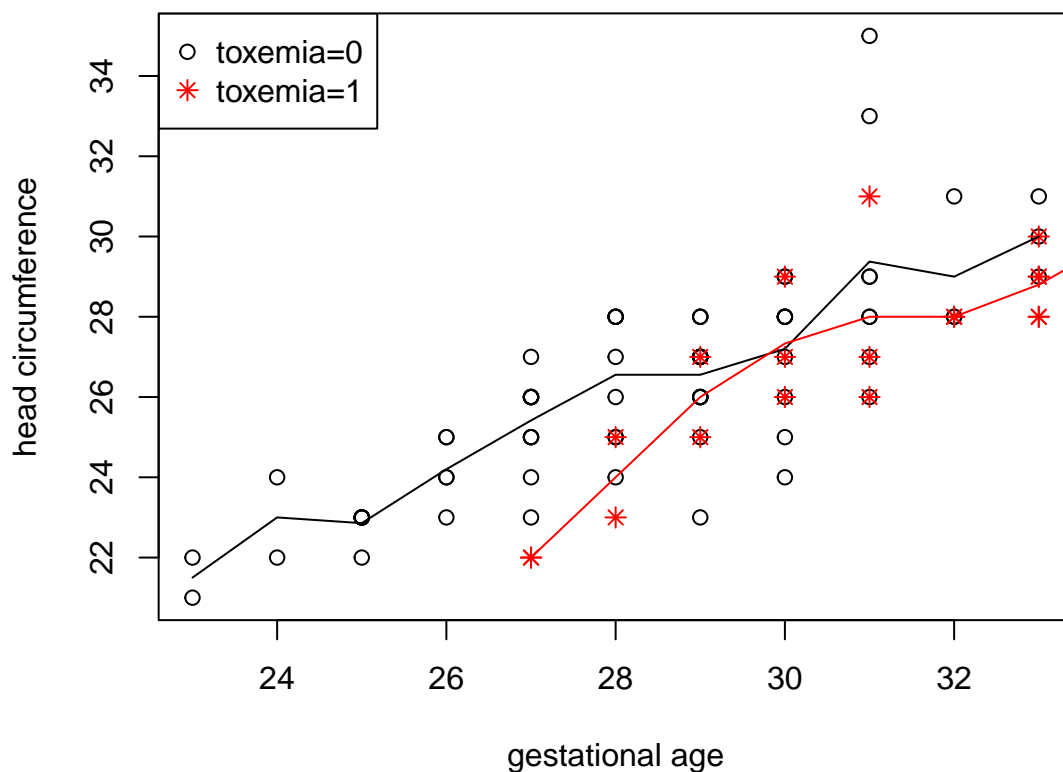
A 95 % CI for β_1 :

$$\hat{\beta}_1 \pm t_{98,0.975} \text{se}(\hat{\beta}_1) = 0.78 \pm 1.985(0.063) = (0.665, 0.905).$$

Linear models with multiple predictors

Low Birthweight Infant Data Example

- Toxemia*, a pregnancy complication characterized by high blood pressure and signs of damage to liver and kidneys, may also have an impact on the development of babies.



- Does *toxemia*, after adjustment for gestational age, also affect the head circumference?

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i.$$

```
fit <- lm(headcirc ~ gestage + factor(toxemia), data = lowbwt)
summary(fit)
```

Call:

```
lm(formula = headcirc ~ gestage + factor(toxemia), data = lowbwt)
```

```

Residuals:
    Min       1Q   Median       3Q      Max
-3.8427 -0.8427 -0.0525  0.8109  6.4092

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.49558    1.86799   0.801  0.42530
gestage        0.87404    0.06561  13.322 < 2e-16 ***
factor(toxemia)1 -1.41233    0.40615  -3.477  0.00076 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.507 on 97 degrees of freedom
Multiple R-squared:  0.6528, Adjusted R-squared:  0.6456
F-statistic: 91.18 on 2 and 97 DF,  p-value: < 2.2e-16

```

What is the interpretation of β_2 ?

$\hat{\beta}_3 = -1.41233$. After adjustment of gestational age, the babies whose mothers had toxemia have smaller (by 1.41 cm) than those whose mothers did not. This difference is significant (test $H_0: \beta_2 = 0$, p -value = $0.0076 < 0.05$).

- Is the rate of increase of head circumference with gestational age the same for infants whose mothers with toxemia as those whose mother without it?

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \varepsilon_i.$$

```

fit <- lm(headcirc ~ gestage * factor(toxemia), data = lowbwt)
summary(fit)

Call:
lm(formula = headcirc ~ gestage * factor(toxemia), data = lowbwt)

Residuals:
    Min       1Q   Median       3Q      Max
-3.8366 -0.8366 -0.0928  0.7910  6.4341

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.76291    2.10225   0.839   0.404
gestage        0.86461    0.07390  11.700 <2e-16 ***
factor(toxemia)1 -2.81503    4.98515  -0.565   0.574
gestage:factor(toxemia)1  0.04617    0.16352   0.282   0.778
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.515 on 96 degrees of freedom
Multiple R-squared:  0.6531, Adjusted R-squared:  0.6422
F-statistic: 60.23 on 3 and 96 DF,  p-value: < 2.2e-16

```

What is the interpretation of β_3 ?

β_3 is the differences in slopes between the two groups (toxemia=1 vs toxemia=0). We want to test $H_0: \beta_3 = 0$, $t = 0.282$, $p\text{-value} = 0.778 > 0.05$. No evidence to reject H_0 .

Limitations of Linear Regression

Linear regression models can be very useful but may not be appropriate to use when response Y is not continuous and can not be assumed to be normally distributed, e.g.,

- Binary data ($Y = 0$ or $Y = 1$),
- Count data ($Y = 0, 1, 2, 3, \dots$).

Generalized Linear Models (GLM) extend the linear regression framework to address the above issue.

- Suitable for continuous and discrete data.
- Normal/Gaussian linear regression is a special case of GLM.
- Inference based on maximum likelihood methods (review next class — 431 Appendix, Stat 330 notes).

WEEK 2
13th to 17th September

Topic 1b: Review of Likelihood Methods

Distributions with a Single Parameter

Setup

- Suppose Y is a random variable with probability density (or mass) function $f(y; \theta)$, where $\theta \in \Omega$ is a continuous parameter.
- The true value of θ is unknown.
- We wish to make inferences about θ (i.e., we may want to estimate θ , calculate a 95 % CI or carry out tests of hypotheses regarding θ).

Likelihood Function

- The **Likelihood function** is any function which is proportional to the probability of observing the data one actually obtained, i.e.,

$$L(\theta; y) = cf(y; \theta) = c\mathbb{P}(Y = y; \theta),$$

where c is a *proportionality constant* that does not depend on θ .

- $L(\theta; y)$ contains all the information regarding θ from the data.
- $L(\theta; y)$ ranks the various parameter values in terms of their consistency with the data.
- Since $L(\theta; y)$ is defined in terms of the random variable y , it is itself a random variable.

Maximum Likelihood Estimator

- For the purposes of estimation we typically want to find θ value that makes the observed data the most likely (hence the term **maximum likelihood**).
- The **maximum likelihood estimator (MLE)** of θ is

$$\hat{\theta} = \arg \max_{\theta} L(\theta; y).$$

- Estimation becomes a simple optimization problem!
- It is often easier to work with the logarithm of the likelihood function, i.e., the **log-likelihood function**

$$\ell(\theta; y) = \log(L(\theta; y)).$$

- Equivalently, since the $\log(\cdot)$ function is monotonic, the value of θ that maximizes $L(\theta; y)$ also maximizes the log-likelihood $\ell(\theta; y)$.
- For simplicity, we drop the y and use $L(\theta) = L(\theta; y)$ and $\ell(\theta) = \ell(\theta; y)$.

A List of Important Functions

- **Log-likelihood function:** $\ell(\theta) = \log(L(\theta))$.
- **Score function:** $S(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = \ell'(\theta)$.
- **Information function:** $I(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta^2} = -\ell''(\theta)$.
- **Fisher information function:** $\mathcal{I}(\theta) = \mathbb{E}[I(\theta)]$.
- **Relative likelihood function:** $R(\theta) = L(\theta)/L(\hat{\theta})$, $0 \leq R(\theta) \leq 1$.
- **Log relative likelihood function:** $r(\theta) = \log(L(\theta)/L(\hat{\theta})) = \ell(\theta) - \ell(\hat{\theta})$, $r(\theta) \leq 0$.

Maximum Likelihood Estimation

- Want θ that maximizes $\ell(\theta)$, or equivalently solves $S(\theta) = 0$.
- Sometimes $S(\theta) = 0$ can be solved explicitly (easy in this case), but often we must solve iteratively.
- Check that the solution corresponds to a maxima of $\ell(\theta)$ by verifying the value of the second derivative at $\hat{\theta}$ is negative, or

$$I(\hat{\theta}) = -\ell''(\hat{\theta}) > 0.$$

- **Invariance property of MLEs:** if $g(\theta)$ is any function of the parameter θ , then the MLE of $g(\theta)$ is $g(\hat{\theta})$.

If $\hat{\theta}$ is the MLE of θ , then $e^{\hat{\theta}}$ is the MLE of e^{θ} .

Example: Binomial Distribution

Example: Binomial Distribution

- A study was conducted to examine the risk for hormone use in healthy postmenopausal women.
- Suppose a group of n women received a combined hormone therapy, and were monitored for the development of breast cancer during 8.5 years follow-up.

- Let

$$Y_i = \begin{cases} 1, & \text{if woman } i \text{ developed breast cancer,} \\ 0, & \text{otherwise,} \end{cases}$$

for $i = 1, \dots, n$.

- Suppose $Y_i \stackrel{\text{iid}}{\sim} \text{BERN}(\pi)$ where $\pi = \mathbb{P}(Y_i = 1)$, then the total number of woman developed breast

cancer is:

$$Y = \sum_{i=1}^n Y_i \sim \text{BIN}(n, \pi).$$

- We wish to find the MLE of unknown parameter π (probability of cancer).

- **Likelihood function:**

$$L(\pi; y) = c \mathbb{P}(Y = y; \pi) = \pi^y (1 - \pi)^{n-y},$$

where we take $c = 1/\binom{n}{y}$ to simplify the likelihood.

- **Log-likelihood function:**

$$\ell(\pi) = y \log(\pi) + (n - y) \log(1 - \pi).$$

- **Score function:**

$$S(\pi) = \frac{y}{\pi} - \frac{n - y}{1 - \pi}.$$

- **Maximum Likelihood Estimator:**

$$S(\pi) = 0 \implies \hat{\pi} = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}.$$

- Second derivative test using **information function:**

$$I(\pi) = -\ell''(\pi) = \frac{y}{\pi^2} + \frac{n - y}{(1 - \pi)^2} > 0 \quad \forall \pi \in (0, 1).$$

Confirms that $\hat{\pi} = \bar{y}$ is the MLE.

Example: Hormone Therapy Data

- A group of $n = 8506$ postmenopausal women aged 50-79 received EPT and $Y = 166$ developed invasive breast cancer during the follow-up.
- Assume $Y \sim \text{BIN}(n, \pi)$ with unknown parameter π .
- The **maximum likelihood estimate** of π is:

$$\hat{\pi} = \bar{y} = \frac{y}{n} = \frac{166}{8506} = 0.0195.$$

Therefore, the probability of breast cancer is estimated to be about 2 %.

Example: Poisson Distribution

Suppose y_1, \dots, y_n is an iid sample from a Poisson distribution with probability mass function:

$$f(y; \lambda) = \mathbb{P}(Y = y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad \lambda > 0, y = 0, 1, 2, \dots$$

- **Likelihood function:**

$$L(\lambda; y_1, \dots, y_n) = \prod_{i=1}^n f(y_i; \lambda) = \frac{\lambda^{\sum y_i} e^{-n\lambda}}{\prod_i y_i!}.$$

- **Log-likelihood function:**

$$\ell(\lambda) = \left(\sum_i y_i \right) \log(\lambda) - n\lambda - \sum_{i=1}^n \log(y_i!).$$

- **Score function:**

$$S(\lambda) = \frac{\sum_i y_i}{\lambda} - n = 0 \implies \hat{\lambda} = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}.$$

Need second derivative test to verify $\hat{\lambda}$ is the MLE.

Newton Raphson Algorithm For Finding MLE

- Sometimes, solving $S(\theta) = 0$ can be challenging and closed form solutions may not be obtained, iterative method need to be used to find the MLE.
- Recall **Taylor Series** expansion of a differentiable function $f(x)$ about a point a :

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \dots$$

- Now suppose we wish to find $\hat{\theta}$, the root of $S(\theta) = 0$ and $\theta^{(0)}$ is a guess that is “close” to $\hat{\theta}$.
- Consider the Taylor series expansion of $S(\theta)$ about $\theta^{(0)}$:

$$S(\theta) = S(\theta^{(0)}) + \frac{S'(\theta^{(0)})}{1!}(\theta - \theta^{(0)}) + \frac{S''(\theta^{(0)})}{2!}(\theta - \theta^{(0)})^2 + \dots$$

- For $|\theta - \theta^{(0)}|$ very small, the second and higher order terms can be dropped to a good approximation:

$$S(\theta) \simeq S(\theta^{(0)}) + S'(\theta^{(0)})(\theta - \theta^{(0)}).$$

$$S(\theta) \simeq S(\theta^{(0)}) - I(\theta^{(0)})(\theta - \theta^{(0)}).$$

- Then at $\theta = \hat{\theta}$,

$$\begin{aligned} S(\hat{\theta}) &\simeq S(\theta^{(0)}) - I(\theta^{(0)})(\hat{\theta} - \theta^{(0)}) \\ I(\theta^{(0)})(\hat{\theta} - \theta^{(0)}) &\simeq S(\theta^{(0)}) \\ (\hat{\theta} - \theta^{(0)}) &\simeq I^{-1}(\theta^{(0)})S(\theta^{(0)}) \\ \hat{\theta} &\simeq \theta^{(0)} + I^{-1}(\theta^{(0)})S(\theta^{(0)}). \end{aligned}$$

- This suggests a revised guess for $\hat{\theta}$ is:

$$\theta^{(1)} = \theta^{(0)} + I^{-1}(\theta^{(0)})S(\theta^{(0)})$$

Newton Raphson Algorithm for finding the MLE

- Begin with an initial estimate $\theta^{(0)}$.
- Iteratively obtain updated estimate by using:

$$\theta^{(i+1)} = \theta^{(i)} + I^{-1}(\theta^{(i)})S(\theta^{(i)}).$$

- Iteration continues until $\theta^{(i+1)} \simeq \theta^{(i)}$ within a specified tolerance.
- Then set $\hat{\theta} = \theta^{(i+1)}$, check that $I(\hat{\theta}) > 0$.

Inference for Scalar Parameters θ

- So far we have discussed estimation of $\hat{\theta}$, next we want to conduct inference about θ , i.e., carry out hypothesis tests and construct confidence intervals of θ .
- Likelihood inference relies on the following **asymptotic distribution results**:

Useful asymptotic distributional results

- (log) Likelihood ratio statistic: $-2 \log(R(\theta)) = -2r(\theta) \sim \chi^2_{(1)}$.
- Score statistic: $(S(\theta))^2 / I(\theta) \sim \chi^2_{(1)}$.
- Wald statistic: $(\hat{\theta} - \theta)^2 I(\hat{\theta}) \sim \chi^2_{(1)}$ or $(\hat{\theta} - \theta) \sqrt{I(\hat{\theta})} \sim \mathcal{N}(0, 1)$ since $Z \sim \mathcal{N}(0, 1) \implies Z^2 \sim \chi^2_1$.

Confidence Interval (CI)

Suppose we want a $100(1 - \alpha)\%$ confidence interval for θ .

- The Likelihood ratio (LR) based pivotal gives a confidence interval:

$$\{\theta : -2r(\theta) < \chi^2_1(1 - \alpha)\},$$

where $\chi^2_1(1 - \alpha)$ is the upper α percentage point of the χ^2_1 distribution.

- The Wald-based pivotal gives an interval:

$$\{\theta : (\hat{\theta} - \theta)^2 I(\hat{\theta}) < \chi^2_1(1 - \alpha)\},$$

or equivalently

$$\hat{\theta} \pm Z_{1-\alpha/2} (I(\hat{\theta}))^{-1/2},$$

where $Z_{1-\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal.

Example: Hormone Therapy Data

Likelihood Ratio based 95% CI: $\{\theta : -2r(\theta) < \chi^2_1(0.95)\}$ where $r(\theta) = \ell(\theta) - \ell(\hat{\theta})$.

- For the Binomial distribution: $\hat{\theta} = y/n$, and

$$r(\theta) = \underbrace{(y \log(\theta) + (n - y) \log(1 - \theta))}_{\ell(\theta)} - \underbrace{\left(y \log\left(\frac{y}{n}\right) + (n - y) \log\left(1 - \frac{y}{n}\right) \right)}_{\ell(\hat{\theta})}.$$

- To find the root of $-2r(\theta) = \chi^2_1(0.95) \iff -2r(\theta) - \chi^2_1(0.95)$:

```
y = 166
n = 8506
LRCI = function(theta, y, n) {
  -2 * (y * log(theta) + (n - y) * log(1 - theta) - y * log(y/n) -
    (n - y) * log(1 - y/n)) - qchisq(0.95, 1)
}
mle = y/n
uniroot(LRCI, c(0, mle), y = y, n = n)$root

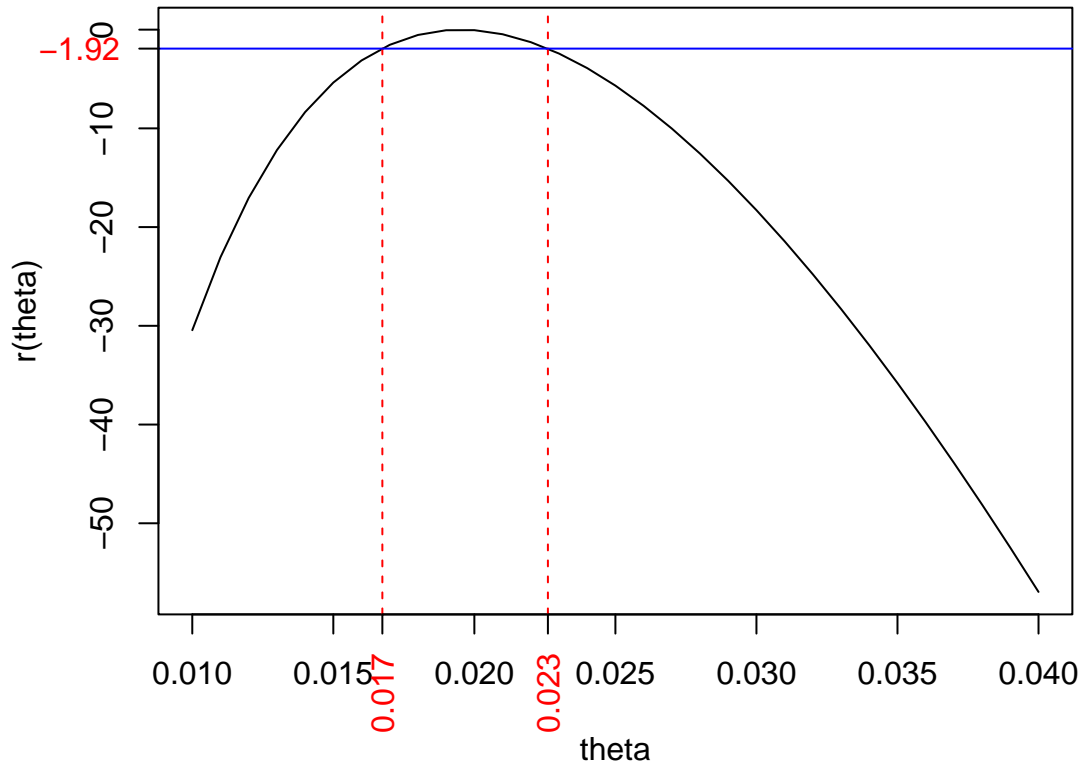
[1] 0.01673867

uniroot(LRCI, c(mle, 1), y = y, n = n)$root

[1] 0.02260709
```


- The likelihood ratio based 95 % CI is (0.017, 0.023).

$$-2r(\theta) < \chi_1^2(0.95) \iff r(\theta) > -\frac{1}{2}\chi_1^2(0.95) = -1.92.$$



Wald based 95 % CI: $\hat{\theta} \pm Z_{0.975}(I(\hat{\theta}))^{-1/2}$.

- For Binomial distribution $\hat{\theta} = y/n$ and

$$I(\hat{\theta}) = \frac{y}{\hat{\theta}^2} + \frac{n-y}{(1-\hat{\theta})^2} = n^2 \left(\frac{1}{y} + \frac{1}{n-y} \right).$$

- So we solve:

$$\begin{aligned} \hat{\theta} \pm 1.96(I(\hat{\theta}))^{-1/2} &= 0.0195 \pm 1.96(0.0015) \\ &= (0.017, 0.022). \end{aligned}$$

- The Wald based 95 % CI is: (0.017, 0.022).

Hypotheses Test

Suppose we are interested in testing hypotheses:

$$H_0: \theta = \theta_0 \text{ vs } H_A: \theta \neq \theta_0.$$

- **Likelihood ratio (LR) test:** $p\text{-value} = \mathbb{P}(\chi_1^2 > -2r(\theta_0))$.

- **Score test:** $p\text{-value} = \mathbb{P}(\chi_1^2 > (S(\theta))^2 / I(\theta_0))$.

- **Wald test:**

$$p\text{-value} = \mathbb{P}(\chi_1^2 > (\hat{\theta} - \theta_0)^2 I(\hat{\theta})), \text{ or } p\text{-value} = \mathbb{P}(|Z| > |\hat{\theta} - \theta_0| \sqrt{I(\hat{\theta})}).$$

Example: Hormone Therapy Data

Suppose we wish to test if women received EPT would have a risk of breast cancer same as that of the general population, say about 1.5 %.

$$H_0: \theta = 0.015 \text{ vs } H_A: \theta \neq 0.015.$$

- **Likelihood Ratio** based test:

$$\begin{aligned} r(\theta_0 = 0.015) &= \left(y \log(0.015) + (n - y) \log(1 - 0.015) \right) - \left(y \log\left(\frac{y}{n}\right) + (n - y) \log\left(1 - \frac{y}{n}\right) \right) \\ &= -5.3637. \end{aligned}$$

Thus, the p -value for the test is given by:

$$p = \mathbb{P}(\chi_{(1)}^2 > -2r(0.015)) = \mathbb{P}(\chi_{(1)}^2 > 10.7274) = 0.001.$$

Therefore, we *reject* H_0 and conclude that the risk of breast cancer for women received EPT is significantly different from 1.5 %.

Notes on Asymptotic Inference

- Asymptotic results: approximation improves as sample size increases.
- Results are exact for a Normal linear model if θ is the mean parameter and σ^2 is known.
- **LR approach:**
 - Need to evaluate (log) likelihood at two locations.
 - Not always a closed form solution for a CI.
 - Usually the best approach.
- **Score approach:**
 - Usually the least powerful test.
 - Don't actually need to find MLE to use.
- **Wald's approach:**
 - Always get a closed form solution for a CI.
 - May not behave well for skewed likelihoods (transform?).
- All three are asymptotically equivalent!

Likelihood Methods for Parameter Vectors

Suppose $\theta = (\theta_1, \theta_2, \dots, \theta_p)^\top \in \Omega$ is a continuous $p \times 1$ parameter vector indexing a probability density (or mass) function $f(y; \theta)$. The likelihood and log-likelihood functions are defined as before, but

- $S(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$ is the $p \times 1$ **Score vector**, i.e.,

$$S(\theta) = \begin{bmatrix} \frac{\partial \ell(\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\theta)}{\partial \theta_p} \end{bmatrix}.$$

- $I(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta^\top \partial \theta}$ is the $p \times p$ **Information matrix**, i.e.,

$$I(\theta) = \begin{bmatrix} -\frac{\partial^2 \ell(\theta)}{\partial \theta_1^2} & -\frac{\partial^2 \ell(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_1 \partial \theta_p} \\ -\frac{\partial^2 \ell(\theta)}{\partial \theta_2 \partial \theta_1} & -\frac{\partial^2 \ell(\theta)}{\partial \theta_2^2} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell(\theta)}{\partial \theta_p \partial \theta_1} & \frac{\partial^2 \ell(\theta)}{\partial \theta_p \partial \theta_2} & \cdots & \frac{\partial^2 \ell(\theta)}{\partial \theta_p^2} \end{bmatrix}.$$

- The Newton Raphson algorithm applies as before, but with vectors and matrices as follows:

$$\theta^{(i+1)} = \theta^{(i)} + I^{-1}(\theta^{(i)})S(\theta^{(i)}).$$

Again, we apply iteratively until we obtain convergence, but now check to see if $I(\hat{\theta})$ is a positive definite matrix.

Suppose we want to make inference about a specific parameter in θ , say we partition vector $\theta = (\alpha, \beta)^\top$ and α is the **parameter of interest**. Analogues to the LR, Score, and Wald results apply, e.g.,

- LR statistic: $-2[\ell(\alpha, \hat{\beta}_\alpha) - \ell(\hat{\alpha}, \hat{\beta})] \sim \chi_{(1)}^2$.
 - $\hat{\beta}_\alpha$ is the MLE of β given α is fixed.
 - $\hat{\alpha}$ and $\hat{\beta}$ are the joint MLE of α and β .
- Score statistic: $S_\alpha(\alpha, \hat{\beta}_\alpha)^2 I^{\alpha\alpha} \sim \chi_{(1)}^2$.
 - $S_\alpha = \frac{\partial \ell}{\partial \alpha}$.
 - $I^{\alpha\alpha}$ is the (α, α) element of $I(\alpha, \beta)^{-1}$ (inverse of Information matrix).
- Wald statistic: $(\hat{\alpha} - \alpha)^2 / I^{\alpha\alpha} \sim \chi_{(1)}^2$.

Topic 2a: Formulation of Generalized Linear Models

The Exponential Family

Definition (Exponential Family)

Consider a random variable Y with probability density (or mass) function $f(y; \theta, \phi)$, we say that the distribution is a member of the **exponential family** if we can write

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

for some functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$.

- The parameter θ is called the **canonical** parameter, and it is unknown.

- The parameter ϕ is called the **scale/dispersion** parameter, is constant, and assumed to be known.

Many well known distributions (continuous/discrete) can be shown to be a member of the exponential family.

Examples

- Poisson Distribution: $Y \sim \text{POI}(\lambda)$,

$$f(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad \lambda > 0, y = 0, 1, \dots$$

Show that Poisson is a member of exponential family and identify the canonical parameter and the functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$.

Solution. $f(y; \lambda) = \exp\{\log(f(y; \lambda))\} = \exp\left\{\frac{y \log(\lambda) - \lambda}{1} - \log(y!)\right\}$. Therefore,

$$\begin{aligned} \theta &= \log(\lambda) && \text{(canonical/natural parameter),} \\ b(\theta) &= \lambda = e^\theta, \\ \phi &= 1, \\ a(\phi) &= 1, \\ c(y; \phi) &= -\log(y!). \end{aligned}$$

- Normal Distribution: $Y \sim \mathcal{N}(\mu, \sigma^2)$ and σ^2 known,

$$f(y; \theta, \phi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y - \mu)^2}{2\sigma^2}\right\}.$$

Show that this Normal distribution is a member of the exponential family.

Solution.

$$\begin{aligned} f(y; \mu, \sigma^2) &= \exp\left\{-\frac{y^2 - 2\mu y + \mu^2}{\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2)\right\} \\ &= \exp\left\{\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2)\right\}. \end{aligned}$$

Therefore,

$$\begin{aligned} \theta &= \mu, \\ \phi &= \sigma^2, \\ a(\phi) &= \phi = \sigma^2, \\ b(\theta) &= \frac{\mu^2}{2} = \frac{\theta^2}{2}, \\ c(y; \phi) &= -\frac{y^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2). \end{aligned}$$

Properties of Exponential Family

Consider a single observation y from the exponential family.

$$L(\theta, \phi; y) = f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y; \phi)\right\}.$$

$$\ell(\theta, \phi; y) = \log(f(y; \theta, \phi)) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y; \phi).$$

$$S(\theta) = \frac{\partial \ell}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}.$$

$$I(\theta) = -\frac{\partial^2 \ell}{\partial \theta^2} = \frac{b''(\theta)}{a(\phi)}.$$

$$\mathcal{I}(\theta) = \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \theta^2}\right] = I(\theta).$$

Some General Results for Score and Information

Result # 1

The expectation of the score function is zero.

$$\mathbb{E}[S(\theta)] = 0.$$

Proof:

$$\begin{aligned} \int f(y; \theta, \phi) dy &= 1 \\ \frac{\partial}{\partial \theta} \int f(y; \theta, \phi) dy &= 0 \\ \int \frac{\partial}{\partial \theta} f(y; \theta, \phi) dy &= 0 \\ \int \left(\frac{\partial}{\partial \theta} \log(f(y; \theta, \phi)) \right) f(y; \theta, \phi) dy &= 0 \\ \int S(\theta) f(y; \theta, \phi) dy &= 0 \\ \mathbb{E}[S(\theta)] &= 0 \end{aligned} \tag{1}$$

Result # 2

The expectation of the score function squared is the expected information.

$$\mathbb{E}[S(\theta; y)^2] = \mathbb{E}[I(\theta; y)]$$

Proof: Differentiate (1) again,

$$\begin{aligned}
& \frac{\partial}{\partial \theta} \int \left(\frac{\partial}{\partial \theta} \log(f(y; \theta, \phi)) \right) f(y; \theta, \phi) dy = 0 \\
& \int \left(\frac{\partial^2}{\partial \theta^2} \log(f(y; \theta, \phi)) \right) f(y; \theta, \phi) dy + \int \left(\frac{\partial}{\partial \theta} \log(f(y; \theta, \phi)) \right) \frac{\partial}{\partial \theta} f(y; \theta, \phi) dy = 0 \\
& \int \frac{\partial^2}{\partial \theta^2} \log(f(y; \theta, \phi)) f(y; \theta, \phi) dy + \int \left(\frac{\partial}{\partial \theta} f(y; \theta, \phi) \right)^2 f(y; \theta, \phi) dy = 0 \\
& \int -I(\theta) f(y; \theta, \phi) dy + \int S(\theta)^2 f(y; \theta, \phi) dy = 0 \\
& \mathbb{E}[-I(\theta; y)] + \mathbb{E}[S(\theta; y)^2] = 0
\end{aligned}$$

Now for the exponential family, we apply above results and obtain:

$$\begin{aligned}
& \mathbb{E}[S(\theta)] = 0, \\
& \mathbb{E}\left[\frac{Y - b'(\theta)}{a(\phi)}\right] = 0, \\
& \mathbb{E}[Y] = b'(\theta), \\
& \mathbb{E}[S(\theta)^2] = \mathbb{E}[I(\theta)], \\
& \mathbb{E}\left[\left(\frac{Y - b'(\theta)}{a(\phi)}\right)^2\right] = \mathbb{E}\left[\frac{b''(\theta)}{a(\phi)}\right], \\
& \frac{1}{a(\phi)^2} \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \frac{b''(\theta)}{a(\phi)}, \\
& \text{Var}(Y) = b''(\theta)a(\phi).
\end{aligned}$$

Mean and Variance for the Exponential Family

- **Mean:** $\mathbb{E}[Y] = b'(\theta) = \mu$.
- **Variance:** $\text{Var}(Y) = b''(\theta)a(\phi)$.

Note that:

- $b'(\theta) = \mu$ tells the relationship between *canonical* parameter θ and μ .
- $b''(\theta)$ is a function of θ and hence can be also expressed as a function of μ .
- Thus, we write $b''(\theta) = V(\mu)$ and call $V(\mu)$ the **variance function**.
- Subsequently, we have:

$$\text{Var}(Y) = b''(\theta)a(\phi) = V(\mu)a(\phi),$$

which is the **mean-variance relationship** for the exponential family.

Link Functions

Definition (Link Function)

The **link function** relates the linear predictor $\eta = \mathbf{x}^\top \boldsymbol{\beta}$ to the expected value μ of the random variable Y , i.e.,

$$g(\mu) = \eta = \mathbf{x}^\top \boldsymbol{\beta},$$

where $g(\cdot)$ is the link function.

Definition (Canonical Link Function)

When Y is a member of the exponential family we define the **canonical link function** to be:

$$g(\mu) = \theta = \eta = \mathbf{x}^\top \boldsymbol{\beta}$$

(i.e., the choice of $g(\cdot)$ that sets canonical parameter = linear predictor).

Example

Recall that $\text{POI}(\lambda)$ is a member of exponential family,

$$f(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!} = \exp \left\{ \frac{y \log(\lambda) - \lambda}{1} - \log(y!) \right\}$$

where $\theta = \log(\lambda)$, $\phi = 1$, $b(\theta) = \lambda = e^\theta$, and $a(\phi) = 1$. Now to find the mean, variance function, and canonical link function:

- **Mean:** $\mathbb{E}[Y] = b'(\theta) = e^\theta = \mu \implies \theta = \log(\mu)$.
- **Variance Function:** $V(\mu) = b''(\theta) = e^\theta \implies V(\mu) = \mu$.
- **Variance:** $\text{Var}(Y) = V(\mu)a(\phi) = \mu$ (mean-variance relationship).
- **Canonical link:** set $\theta = \eta$ using $\theta = \log(\mu) = \eta = \mathbf{x}^\top \boldsymbol{\beta}$, i.e., $g(\mu) = \log(\mu)$ where $\log(\cdot)$ is the canonical link.

Moving forward, we consider a log-linear model: $\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$.

Remarks on Link Function

- We can choose any function $g(\cdot)$ as the link function in theory.
- The canonical link is a special link function, we often choose to use canonical link for its good statistical properties.
- Context and goodness of fit should motivate the choice of link function in practice.

Generalized Linear Models

Definition (Generalized Linear Model (GLM))

A **Generalized Linear Model (GLM)** is composed of three components:

- **Random Component:** The responses Y_1, \dots, Y_n are independent random variables and each Y_i is assumed to come from a parametric distribution that is a member of the exponential family.

- **Systematic Component** (or linear predictor):

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta},$$

a linear combination of explanatory variables \mathbf{x}_i and regression parameters $\boldsymbol{\beta}$.

- **Link function:**

$$g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta},$$

a function that relates the mean of response to the linear predictor.

Topic Summary

2a Formulation of Generalized Linear Models:

- Definition of the **Exponential Family**.
 - Exponential form of the probability density (or mass) function.
 - Derivation of Score and Information.
 - Properties of exponential family, mean-variance relationship.
 - Definition of canonical link.
- Definition of a **Generalized Linear Model**.

Next Topic: 2b Estimation for Generalized Linear Models.

WEEK 3
20th to 24th September

Topic 2b: Maximum Likelihood Estimation for Generalized Linear Models

Generalized Linear Models

Suppose for each subject $i = 1, \dots, n$ in a random sample:

- Y_i is the response variable.
- x_{i1}, \dots, x_{ip} are explanatory variables associated with Y_i .

We consider a **Generalized Linear Model** (GLM) for the data, by definition the GLM is composed following three components:

① Random Component:

$Y_i \sim \text{exponential family}$, Y_1, \dots, Y_n are independent.

② Systematic Component (or linear predictor):

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}.$$

- $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^\top$ is a covariate vector.
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^\top$ is a vector of regression coefficients.

③ Link function: a function $g(\cdot)$ links $\mathbb{E}[Y_i] = \mu_i$ to a linear prediction η_i :

$$g(\mu_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

Example: A Poisson Regression Model

Suppose $Y_i \stackrel{\text{ind}}{\sim} \text{POI}(\lambda_i)$ with mean $\mathbb{E}[Y_i] = \lambda_i, i = 1, \dots, n$:

$$f(y_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = \exp\{y_i \log(\lambda_i) - \lambda_i - \log(y_i!)\}.$$

Poisson distribution is a member of exponential family with:

- Canonical parameter: $\theta_i = \log(\lambda_i)$.
- Canonical link: $\theta_i = \eta_i \implies \log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$ (log link).

A Poisson regression model with the canonical link takes the form:

$$\log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} \quad (\text{log-linear model}).$$

Example: A Normal Regression Model

Assume $Y_i \stackrel{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$ and σ^2 is known, $i = 1, \dots, n$:

$$\begin{aligned} f(y_i) &= (2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right\} \\ &= \exp\left\{\frac{y_i \mu_i - \mu_i^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y_i^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right\}. \end{aligned}$$

A Normal distribution (σ^2 known) is a member of exponential family with:

- Canonical parameter: $\theta_i = \mu_i$.
- Canonical link: $\theta_i = \eta_i \implies \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$ (identity link).

A Normal regression model with the canonical link takes the form:

$$\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} \quad (\text{linear model}).$$

Linear regression model (STAT 331) is a Normal GLM using the canonical link!

Likelihood for Generalized Linear Models

We wish to use likelihood methods for the estimation of the regression parameter $\boldsymbol{\beta}$ from the GLM: $g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$. Consider the log-likelihood for a *single* observation from the exponential family:

$$\ell(\theta, \phi; y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y; \phi).$$

- ℓ is a function of θ (assume that ϕ is known).
- θ is related to μ through the result:

$$\mu = b'(\theta).$$

- η can be expressed in terms of μ through the link function:

$$g(\mu) = \eta.$$

- $\boldsymbol{\beta}$ can be expressed in terms of η through the linear predictor:

$$\eta = \mathbf{x}^\top \boldsymbol{\beta}.$$

Score Vector

To find the maximum likelihood estimator for β , we must solve $S(\beta) = \frac{\partial \ell}{\partial \beta} = \mathbf{0}$. Consider taking derivative with respect to β_j using the chain rule:

$$\frac{\partial \ell}{\partial \beta_j} = \frac{\partial \ell}{\partial \theta} \frac{\partial \theta}{\partial \mu} \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \beta_j},$$

where

$$\frac{\partial \ell}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)},$$

$$\frac{\partial \theta}{\partial \mu} = \left(\frac{\partial \mu}{\partial \theta} \right)^{-1} = \frac{1}{b''(\theta)}$$

since $\mu = b'(\theta)$,

$$\frac{\partial \mu}{\partial \eta} = \frac{\partial \mu}{\partial \eta},$$

$$\frac{\partial \eta}{\partial \beta_j} = x_j$$

since $\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_j x_j + \dots + \beta_p x_p$.

Hence, we have:

$$\frac{\partial \ell}{\partial \beta_j} = \frac{y - b'(\theta)}{a(\phi)} \frac{1}{b''(\theta)} \frac{\partial \mu}{\partial \eta} x_j$$

$$= \frac{y - \mu}{\text{Var}(Y)} \frac{\partial \mu}{\partial \eta} x_j$$

since $\mu = b'(\theta)$, $\text{Var}(Y) = a(\phi)b''(\theta)$

$$= \frac{y - \mu}{\text{Var}(Y)} \left(\frac{\partial \mu}{\partial \eta} \right)^2 \frac{\partial \eta}{\partial \mu} x_j$$

since $\frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \mu} = 1$

$$= (y - \mu) \left(\text{Var}(Y) \left(\frac{\partial \mu}{\partial \eta} \right)^2 \right)^{-1} \frac{\partial \eta}{\partial \mu} x_j$$

$$= (y - \mu) W \frac{\partial \eta}{\partial \mu} x_j,$$

where $W^{-1} = \text{Var}(Y) \left(\frac{\partial \eta}{\partial \mu} \right)^2$. Note that generally $\frac{\partial \eta}{\partial \mu}$ is easier to calculate than $\frac{\partial \mu}{\partial \eta}$ since we define the link as $\eta = g(\mu)$.

For a random sample Y_1, \dots, Y_n from exponential family and each Y_i has a probability density function

$$f(y_i; \theta, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}.$$

We write likelihood and log-likelihood functions as:

$$L = \prod_{i=1}^n f(y_i; \theta_i, \phi) = \prod_{i=1}^n \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\},$$

$$\ell = \sum_{i=1}^n \ell_i = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi).$$

The **element of the score vector** is:

$$[S(\beta)]_j = \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^n (y_i - \mu_i) W_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij}$$

where $W_i^{-1} = \text{Var}(Y_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2$, $g(\mu_i) = \eta_i = \mathbf{x}_i^\top \beta$. In vector and matrix form we can write:

$$S(\beta) = \mathbf{X} \mathbf{W} \mathbf{A}(\mathbf{y} - \boldsymbol{\mu}),$$

where

- $\mathbf{y} = (y_1, \dots, y_n)^\top$ and $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^\top$ are $n \times 1$ vectors,
- $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ is a $(p+1) \times n$ matrix,
- $\mathbf{W} = \text{diag}(W_1, \dots, W_n) = \begin{bmatrix} W_1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & W_n \end{bmatrix}$, and
- $\mathcal{A} = \text{diag}\left(\frac{\partial \eta_1}{\partial \mu_1}, \dots, \frac{\partial \eta_n}{\partial \mu_n}\right)$.

Example: Poisson Regression Model (Problem 1.4)

For a random sample from Poisson distribution, $Y_i \sim \text{POI}(\lambda_i)$, $i = 1, \dots, n$,

$$\ell_i = \log(f(y_i; \lambda_i)) = (y_i \log(\lambda_i) - \lambda_i - \log(y_i!)).$$

Poisson regression with a log-link:

$$\log(\lambda_i) = \eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

To write down the score vector for the regression coefficients $\boldsymbol{\beta}$, we may calculate the derivative using standard methods, i.e.,

$$\begin{aligned} [\mathbf{S}(\boldsymbol{\beta})]_j &= \sum_i \frac{\partial \ell_i}{\partial \beta_j} \\ &= \sum_i \frac{\partial}{\partial \beta_j} (y_i \log(\lambda_i) - \lambda_i - \log(y_i!)) \\ &= \sum_i (y_i x_{ij} - e^{\mathbf{x}_i^\top \boldsymbol{\beta}} x_{ij}). \end{aligned}$$

Or we can use the general results derived for the GLMs on the previous slides.

Solving $\mathbf{S}(\boldsymbol{\beta}) = \mathbf{0}$ for MLE

- 1 Newton Raphson update equation is:

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} + \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}^{(r)}) \mathbf{S}(\hat{\boldsymbol{\beta}}^{(r)}),$$

where \mathbf{I} is the observed information matrix.

- This requires us to find and repeatedly evaluate the information \mathbf{I} (possibly computationally intensive).
- Fisher suggested using the expected information matrix \mathcal{I} rather than the observed information matrix.

- 2 Fisher Scoring update equation is:

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \hat{\boldsymbol{\beta}}^{(r)} + \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}}^{(r)}) \mathbf{S}(\hat{\boldsymbol{\beta}}^{(r)}).$$

Information Matrix

Consider the (j, k) element of the Information matrix:

$$\begin{aligned}
 \mathbf{I}_{jk} &= -\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} \\
 &= -\frac{\partial}{\partial \beta_k} \frac{\partial \ell}{\partial \beta_j} \\
 &= \sum_i -\frac{\partial}{\partial \beta_k} \left[(y_i - \mu_i) W_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{ij} \right] \\
 &= \sum_i -(y_i - \mu_i) \left\{ \frac{\partial}{\partial \beta_k} \left[W_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{ij} \right] \right\} - W_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{ij} \left(\frac{\partial}{\partial \beta_k} (y_i - \mu_i) \right) \\
 &= \sum_i -(y_i - \mu_i) \left\{ \frac{\partial}{\partial \beta_k} \left[W_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{ij} \right] \right\} + W_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{ij} \frac{\partial \mu_i}{\partial \beta_k} \frac{\partial \eta_i}{\partial \mu_i} \\
 &= \sum_i -(y_i - \mu_i) \frac{\partial}{\partial \beta_k} \left[W_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{ij} \right] + x_{ij} W_i x_{ik}.
 \end{aligned}$$

Fisher Information

To get an element of the Expected/Fisher Information matrix:

$$\begin{aligned}
 \mathcal{I}_{jk} &= \sum_i \mathbb{E} \left[-\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} \right] \\
 &= \sum_i \mathbb{E} \left[-(Y_i - \mu_i) \frac{\partial}{\partial \beta_k} \left[W_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{ij} \right] + x_{ij} W_i x_{ik} \right] \\
 &= \sum_i -\mathbb{E}[(Y_i - \mu_i)] \frac{\partial}{\partial \beta_k} \left[W_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{ij} \right] + x_{ij} W_i x_{ik} \\
 &= \sum_i x_{ij} W_i x_{ik}.
 \end{aligned}$$

Therefore, we can write the (j, k) element of the Fisher information as:

$$\mathcal{I}_{jk} = \sum_{i=1}^n x_{ij} W_i x_{ik} = [\mathbf{X} \mathbf{W} \mathbf{X}^\top]_{jk}$$

where again, $\mathbf{W} = \text{diag}(W_1, \dots, W_n)$ and $W_i^{-1} = \text{Var}(Y_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2$.

When is Fisher Scoring Equivalent to Newton Raphson?

Recall information matrix:

$$\mathbf{I}_{jk} = \sum_i -(y_i - \mu_i) \frac{\partial}{\partial \beta_k} \left[W_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{ij} \right] + x_{ij} W_i x_{ik}.$$

Now examine:

$$\begin{aligned}
 W_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{ij} &= \left(\text{Var}(Y_i) \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 \right)^{-1} \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{ij} \\
 &= \left(a(\phi) b''(\theta_i) \frac{\partial \eta_i}{\partial \mu_i} \right)^{-1} x_{ij} && \text{since } \text{Var}(Y_i) = a_i(\phi) b''(\theta_i) \\
 &= \left(a(\phi) \frac{\partial \mu_i}{\partial \theta_i} \frac{\partial \eta_i}{\partial \mu_i} \right)^{-1} x_{ij} && \text{since } b'(\theta_i) = \mu_i, b''(\theta_i) = \frac{\partial \mu_i}{\partial \theta_i} \\
 &= (a(\phi))^{-1} x_{ij} && \text{under the canonical link } \theta_i = \eta_i.
 \end{aligned}$$

So under the **canonical link**,

$$\frac{\partial}{\partial \beta_k} \left[W_i \left(\frac{\partial \eta_i}{\partial \mu_i} \right) x_{ij} \right] = \frac{\partial}{\partial \beta_k} \left[(a(\phi))^{-1} x_{ij} \right] = 0,$$

therefore information matrix is same as the Fisher information:

$$\mathbf{I}_{jk} = \sum_i x_{ij} W_i x_{ij} = \mathbf{I}_{jk}$$

and Fisher Scoring is equivalent to Newton Raphson.

Iteratively Reweighted Least Squares

The Fisher Scoring is also called **iteratively reweighted least squares** (IRWLS). The reason is that the update equation can be rewritten as:

$$\hat{\beta}^{(r+1)} = \left(\mathbf{X} \mathbf{W}(\hat{\beta}^{(r)}) \mathbf{X}^\top \right)^{-1} \mathbf{X} \mathbf{W}(\hat{\beta}^{(r)}) \mathbf{Z}(\hat{\beta}^{(r)})$$

where \mathbf{Z} is a transformation of the response vector \mathbf{Y} such that:

$$\mathbf{Z} = \boldsymbol{\eta} + (\mathbf{Y} - \boldsymbol{\mu}) * \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}}$$

- See manipulation in Section 1.2.3 of course notes.
- Same form as the weighted LS estimate of β with dependent variable \mathbf{Z} and weight matrix \mathbf{W} .
- \mathbf{Z} and \mathbf{W} are updated at each iteration.

Topic Summary

2b Maximum Likelihood Estimation of Generalized Linear Models:

- When Y_i come from a distribution in the **exponential family**, we can use the theory of **Generalized Linear Models** to fit the regression equations of the form:

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

- The **link function** $g(\cdot)$ may be the canonical link, but its choice should come from model interpretation and fit.
- Can use Fisher Scoring (also known as IRWLS) to estimate the regression parameters β from any GLM based on general forms for $\mathbf{I}(\beta)$ and $\mathbf{S}(\beta)$.
- **PRACTICE**: Chapter 1 review problems.

Topic 3a: Binary Data and Odds Ratios

Binary Data Set-up

- Consider the simplest case with two *binary* variables:
 - COVID-19: infected or not infected (response).
 - Vaccination: yes or no (explanatory variable).
- Use a 2×2 table to summarize the data:

Vaccination	COVID-19		
	infected	not infected	
yes	y_1	$m_1 - y_1$	m_1
no	y_2	$m_2 - y_2$	m_2
Total	y_{\bullet}	$m_{\bullet} - y_{\bullet}$	m_{\bullet}

- Treat m_1 and m_2 as fixed, assume Y_1 and Y_2 are independent binomial r.v.'s

$$Y_k \sim \text{BIN}(m_k, \pi_k), \quad k = 1, 2,$$

where $\pi_k = \mathbb{P}(\text{infection} \mid \text{group } k)$.

- How do we measure the associate between COVID-19 infection and vaccination?

Measures of Association

Definition (Odds Ratio)

The **Odds Ratio** (OR) is the ratio of the odds of an event occurring in one group to the odds of the event in another group (e.g., not vaccinated):

$$\text{Odds Ratio} = \frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)}.$$

Interpretation of OR:

$$\begin{array}{llll} \pi_1 = \pi_2 & \implies & \text{OR} = 1 & \implies \text{equal risk (no association)} \\ \pi_1 > \pi_2 & \implies & \text{OR} > 1 & \implies \text{higher risk in group 1} \\ \pi_1 < \pi_2 & \implies & 0 < \text{OR} < 1 & \implies \text{higher risk in group 2} \end{array}$$

Relative Risk (RR)

The **Relative Risk** (RR) is the ratio of the probability of an event occurring in one group versus another group:

$$\text{Relative Risk} = \frac{\pi_1}{\pi_2}$$

In the case of a **rare disease** (i.e., when π_1 and π_2 are very small),

$$\text{OR} = \frac{\pi_1 / (1 - \pi_1)}{\pi_2 / (1 - \pi_2)} = \frac{\pi_1}{\pi_2} \underbrace{\left(\frac{1 - \pi_2}{1 - \pi_1} \right)}_{\approx 1} \approx \frac{\pi_1}{\pi_2} = \text{RR},$$

then

$$\text{OR} \approx \text{RR}.$$

Maximum Likelihood Estimation of Odds Ratio

- Goal: Estimate odds ratio $\psi = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$ using likelihood method. Based on “grouped” binomial data, $Y_k \sim \text{BIN}(m_k, \pi_k)$, $k = 1, 2$,

$$\begin{aligned} L(\pi_1, \pi_2) &= \binom{m_1}{y_1} \pi_1^{y_1} (1 - \pi_1)^{m_1 - y_1} \binom{m_2}{y_2} \pi_2^{y_2} (1 - \pi_2)^{m_2 - y_2} \\ &\propto \left(\frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)} \right)^{y_1} \left(\frac{\pi_2}{1 - \pi_2} \right)^{y_2 + y_1} (1 - \pi_1)^{m_1} (1 - \pi_2)^{m_2}. \end{aligned}$$

- Note that $\pi_1, \pi_2 \in [0, 1]$ and odds ratio $\psi \in (0, \infty)$ are restricted, we consider re-parameterize:

$$\theta_1 = \log\left(\frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}\right) = \log(\psi), \quad \theta_2 = \log\left(\frac{\pi_2}{1 - \pi_2}\right),$$

and now $\theta_1, \theta_2 \in (-\infty, \infty)$.

- Our re-parameterization implies:

$$\pi_1 = \frac{e^{\theta_1 + \theta_2}}{1 + e^{\theta_1 + \theta_2}}, \quad \pi_2 = \frac{e^{\theta_2}}{1 + e^{\theta_2}}.$$

Now the likelihood becomes:

$$\begin{aligned} L(\theta_1, \theta_2) &= (e^{\theta_1})^{y_1} (e^{\theta_2})^{y_1 + y_2} (1 + e^{\theta_1 + \theta_2})^{m_1} (1 + e^{\theta_2})^{-m_2}, \\ \ell(\theta_1, \theta_2) &= y_1 \theta_1 + (y_1 + y_2) \theta_2 - m_1 \log(1 + e^{\theta_1 + \theta_2}) - m_2 \log(1 + e^{\theta_2}). \end{aligned}$$

- The score vector is:

$$S(\theta_1, \theta_2) = \begin{pmatrix} \frac{\partial \ell}{\partial \theta_1} \\ \frac{\partial \ell}{\partial \theta_2} \end{pmatrix} = \begin{pmatrix} y_1 - m_1 \left(\frac{e^{\theta_1 + \theta_2}}{1 + e^{\theta_1 + \theta_2}} \right) \\ y_1 + y_2 - m_1 \left(\frac{e^{\theta_1 + \theta_2}}{1 + e^{\theta_1 + \theta_2}} \right) - m_2 \left(\frac{e^{\theta_2}}{1 + e^{\theta_2}} \right) \end{pmatrix}.$$

- Solving $S(\theta_1, \theta_2) = \mathbf{0}$ gives us the MLEs:

$$\hat{\theta}_1 = \log\left(\frac{y_1/(m_1 - y_1)}{y_2/(m_2 - y_2)}\right), \quad \hat{\theta}_2 = \log\left(\frac{y_2}{m_2 - y_2}\right).$$

- So by the invariance property of MLEs, we have:

$$\hat{\pi}_1 = \frac{y_1}{m_1}, \quad \hat{\pi}_2 = \frac{y_2}{m_2}, \quad \hat{\psi} = \frac{\hat{\pi}_1/(1 - \hat{\pi}_1)}{\hat{\pi}_2/(1 - \hat{\pi}_2)} = \frac{y_1/(m_1 - y_1)}{y_2/(m_2 - y_2)}.$$

Inference for Odds Ratio

- In order to do inference we will need the Information Matrix:

$$\mathbf{I}(\theta_1, \theta_2) = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \quad \text{where } I_{jk} = -\frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell(\theta_1, \theta_2).$$

Here, we have:

$$\begin{aligned} I_{11} &= m_1 \left(\frac{e^{\theta_1 + \theta_2}}{(1 + e^{\theta_1 + \theta_2})^2} \right), \\ I_{12} &= I_{21} = m_1 \left(\frac{e^{\theta_1 + \theta_2}}{(1 + e^{\theta_1 + \theta_2})^2} \right), \\ I_{22} &= m_1 \left(\frac{e^{\theta_1 + \theta_2}}{(1 + e^{\theta_1 + \theta_2})^2} \right) + m_2 \left(\frac{e^{\theta_2}}{(1 + e^{\theta_2})^2} \right). \end{aligned}$$

- We are interested in doing inference on $\theta_1 = \log(\psi)$ (while θ_2 is nuisance).
- Recall the asymptotic distribution result of a **Wald statistic**:

Wald Statistic

For a vector $\theta = (\theta_1, \theta_2)^\top$ where $\theta_1 = \log(\psi)$ is a scalar parameter of interest:

$$(\hat{\theta}_1 - \theta_1)^2 (I^{11}(\hat{\theta}_1, \hat{\theta}_2))^{-1} \sim \chi^2_{(1)}, \text{ or } (\hat{\theta}_1 - \theta)/\sqrt{I^{11}} \sim \mathcal{N}(0, 1),$$

where I^{11} is the $(1, 1)$ element of I^{-1} evaluated at MLE $\hat{\theta}_1$ and $\hat{\theta}_2$.

- Calculation of I^{11} by using a general result:

$$I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}, \quad I^{-1} = \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix}, \quad I^{11} = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}.$$

- We can use the Wald result to find a confidence interval for $\theta_1 = \log(\psi)$.

Confidence Interval for Odds Ratio

Here, we obtain:

$$I^{11}(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{y_1} + \frac{1}{m_1 - y_1} + \frac{1}{y_2} + \frac{1}{m_2 - y_2}.$$

Thus, a Wald-based 95 % confidence interval for $\theta_1 = \log(\psi)$ is:

$$\hat{\theta}_1 \pm 1.96 \sqrt{\frac{1}{y_1} + \frac{1}{m_1 - y_1} + \frac{1}{y_2} + \frac{1}{m_2 - y_2}} = (\hat{\theta}_{1L}, \hat{\theta}_{1U}).$$

A 95 % confidence interval for the Odds Ratio ψ is:

$$(\exp\{\hat{\theta}_{1L}\}, \exp\{\hat{\theta}_{1U}\}).$$

Example: Prenatal Care from Two Clinics

Consider the data below for the relationship between:

- **Response**: Fetal Mortality.
- **Explanatory variable**: Level of Care.

Level of Care	Fetal Mortality		Total
	Died	Survived	
Intensive	20	316	336
Regular	46	373	419
	66	689	755

- Using the above data, we obtain MLE of odds ratio ψ :

$$\hat{\psi} = \frac{y_1/(m_1 - y_1)}{y_2/(m_2 - y_2)} = \frac{20/316}{46/373} = 0.51.$$

$\hat{\psi} = 0.51 < 1$, the risk of mortality is lower with intensive care.

- A 95 % CI for $\theta_1 = \log(\psi)$:

$$\log(0.51) \pm 1.96 \sqrt{\frac{1}{20} + \frac{1}{316} + \frac{1}{46} + \frac{1}{373}} = (-1.219, -0.127).$$

- A 95 % CI for odds ratio ψ :

$$(\exp\{-1.219\}, \exp\{-0.127\}) = (0.30, 0.89).$$

Note that the CI does not cover the value $\psi = 1$ (no association), so we reject the null hypothesis of no association between fetal mortality and level of care. In other words, there is evidence of association.

Example: Prenatal Care from Two Clinics

There is an [additional explanatory variable](#): Clinic (A vs B).

Prenatal Care Data Stratified by Clinic

Level of Care	Clinic A			Clinic B		
	Died	Survived	Total	Died	Survived	Total
Intensive	16	293	309	4	23	27
Regular	12	176	188	34	197	231
	28	469	497	38	220	258

- $\hat{\psi}_A = 0.80$ (0.37, 1.73) and $\hat{\psi}_B = 1.01$ (0.33, 3.10). These cover value 1, different from the results from the pooled analysis on the previous slide.
- These results do NOT agree with the results from the pooled analysis on the previous slide.

Association Between Clinic and Level of Care

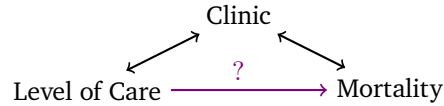
	A	B	
Intensive	309	27	336
Regular	118	231	419
	497	258	755

- $\hat{\psi} = 14.06$ (9.12, 21.76).

Association Between Clinic and Mortality

	A	B	
Died	28	38	66
Survived	469	220	689
	497	258	755

- $\hat{\psi} = 0.35$ (0.21, 0.58).
- The initial strong association between Level of Care and Infant Mortality disappeared when we stratified by clinic.
- Instead of having to examine multiple 2×2 tables we'd like to estimate the OR and compute associations using a multiple regression model.
- One way to do this is by fitting a Binomial GLM to the data.



Topic 3b: Binomial Regression Models for Binary Data

Recall Topic 3a: Binary Data and Odds Ratios

Last week, we introduce a simple method for association between two binary variables, 2×2 contingency table

analysis: Measure of Association: $OR = \psi = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)}$,

Level of Care	Mortality		
	Died	Survived	
Intensive	y_1	$m_1 - y_1$	$Y_1 \sim \text{BIN}(m_1, \pi_1)$
Regular	y_2	$m_2 - y_2$	$Y_2 \sim \text{BIN}(m_2, \pi_2)$

- $OR = 1$ (equal risk).
- $0 < OR < 1$ (lower risk in group 1).
- $OR > 1$ (higher risk in group 1).

Maximum likelihood estimator for OR is:

$$\hat{\psi} = \frac{y_1/(m_1 - y_1)}{y_2/(m_2 - y_2)},$$

and a Wald-based 95 % CI is:

$$\exp\left\{\log(\hat{\psi}) \pm 1.96 \underbrace{\sqrt{\frac{1}{y_1} + \frac{1}{m_1 - y_1} + \frac{1}{y_2} + \frac{1}{m_2 - y_2}}}_{\text{se}(\log(\hat{\psi}))}\right\}.$$

Prenatal Care Data Example

OR (Mortality and Care)	Est.	95 % CI
Intensive vs Regular	0.51	(0.30, 0.89)

Table 1: $1 \notin (0.30, 0.89) \implies$ evidence of association between Mortality and Care.

However, Mortality and Care are also related to another variable, Clinic:

OR (Mortality and Clinic)	Est.	95 % CI
Clinic A vs Clinic B	0.35	(0.12, 0.58)

Table 2: Association between Mortality and Clinic.

- Therefore, we wish to consider how a variable, e.g., Mortality (Y), is related to multiple explanatory variables together, e.g., Care (x_1) and Clinic (x_2).
- This can be done using [multiple regression methodology](#) for binary data \implies Topic 3b: Binomial Regression Models for Binary Data.

OR (Care and Clinic)	Est.	95 % CI
Clinic A vs Clinic B	14.06	(9.12, 21.76)

Table 3: Association between Care and Clinic.

Multiple Regression for Binary Data

- Often we need to consider the relationship between a binary outcome and multiple explanatory variables, using multiple regression methodology.
- This is because we may want to:
 - control for confounding variables and hence want to examine the effect of several variables simultaneously;
 - examine the effect of categorical variables (> 2 levels) or continuous covariates;
 - develop sophisticated models that describe complex relationship.
- Suppose *subject level data* is binary with a value of 1 indicating that an event of interest occurs and a value of 0 indicating that event doesn't occur.
- Subjects can be classified according to the values of explanatory variables into n groups (i.e., common covariates values within each group), so we have *grouped data* such that:
 - m_i denotes number of subjects in group i ;
 - Y_i denotes number of subjects experienced the event in group i ;
 - x_{i1}, \dots, x_{ip} denote the covariates values associated with group i where $i = 1, \dots, n$.

Set-up of a Binomial Regression Model

- ① **Response Variable:** $Y_i \sim \text{BIN}(m_i, \pi_i)$, $i = 1, \dots, n$, and Binomial distribution is a member of Exponential family!

$$\begin{aligned}
 f(y_i) &= \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i} \\
 &= \exp \left\{ y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) + m_i \log(1 - \pi_i) + \log \left(\binom{m_i}{y_i} \right) \right\},
 \end{aligned}$$

where

$$\begin{aligned}
 \theta_i &= \log \left(\frac{\pi_i}{1 - \pi_i} \right), \\
 a(\phi) &= \phi = 1, \\
 b(\theta_i) &= -m_i \log(1 - \pi_i) = m_i \log(1 + e^{\theta_i}). \\
 c(y_i; \phi) &= \log \left(\binom{m_i}{y_i} \right).
 \end{aligned}$$

- ② **Linear Predictor:**

$$\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}.$$

- ③ **Link Function:** Recall that for Binomial distribution, we have $\mathbb{E}[Y_i] = \mu_i = m_i \pi_i$, therefore we typically re-write the link function in terms of π_i ,

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

As $\pi_i \in (0, 1)$, any function $g: (0, 1) \rightarrow (-\infty, \infty)$ may work, and here are some link functions we can consider:

log-log	$g(\pi) = \log(-\log(\pi))$
complementary log-log	$g(\pi) = \log(-\log(1 - \pi))$
Probit ^a	$g(\pi) = \Phi^{-1}(\pi)$
Logit (canonical)	$g(\pi) = \log(\pi/(1 - \pi))$

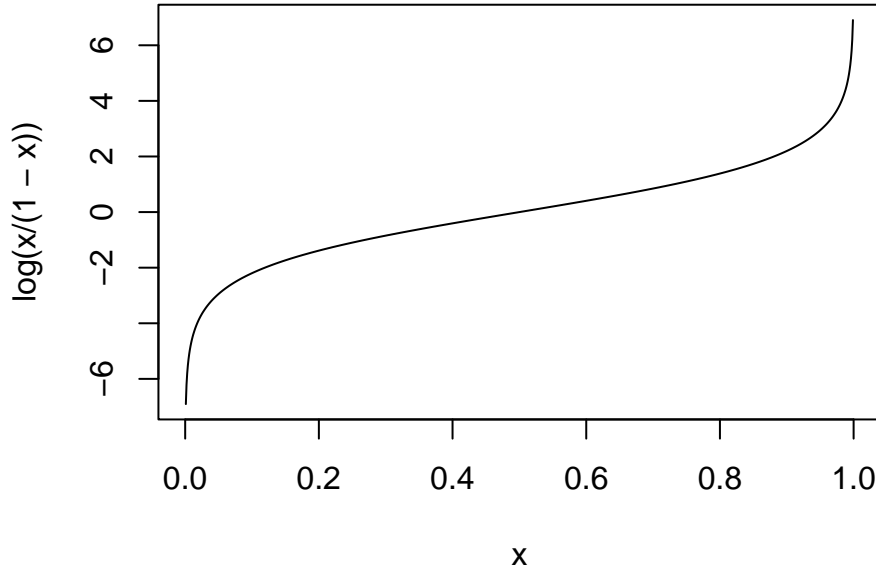
^aFor the Probit link, $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0, 1)$.

Canonical Link and Logistic Regression

Recall for Binomial distribution $\theta_i = \log\left(\frac{\pi_i}{1 - \pi_i}\right)$, and by setting $\theta_i = \eta_i$, we have:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \eta_i.$$

The [Logit link](#), $g(\pi_i) = \log(\pi_i/(1 - \pi_i))$, is the canonical link for the Binomial!



This leads to a Logistic Regression Model:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

Prediction from Logistic Regression

Aside: The inverse of the logit function is called the expit function:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i^\top \boldsymbol{\beta} \iff \pi_i = \frac{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}{1 + \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}} = \text{expit}(\mathbf{x}_i^\top \boldsymbol{\beta}).$$

Suppose we have found MLE $\hat{\boldsymbol{\beta}}$ using Fisher scoring, then the fitted value for the [probability of response](#) π_i given explanatory variables \mathbf{x}_i is:

$$\hat{\pi}_i = \frac{\exp\{\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\}}{1 + \exp\{\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\}}.$$

The predicted number of responses are: $\hat{Y}_i = m_i \hat{\pi}_i$.

Interpretation of β in Logistic Regression

- Consider a simple logistic model with a single binary explanatory variable:

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1},$$

where $x_{i1} = 0$ (group 0) and $x_{i1} = 1$ (group 1).

- Let's compare the model when $x_{i1} = 1$ vs $x_{i1} = 0$.

Group	\mathbf{x}_i^\top	$\eta_i = \log(\pi_i/(1 - \pi_i))$
1	$(1, 1)^\top$	$\beta_0 + \beta_1 = \log(\pi_1/(1 - \pi_1))$
0	$(1, 0)^\top$	$\beta_0 = \log(\pi_0/(1 - \pi_0))$
		$\beta_1 = \log\left(\frac{\pi_1/(1 - \pi_1)}{\pi_0/(1 - \pi_0)}\right) = \log(\text{OR})$

- We subtract line 2 from line 1 to isolate β_1 and find its interpretation.
- $\beta_1 = \log$ odds ratio of response for subjects with $x_{i1} = 1$ vs $x_{i1} = 0$.
- Please see Section 2.4.2 for general interpretations of β 's in multiple logistic regression models.

Logistic Regression for Prenatal Care Example

- Response:** Fetal Mortality, that is,

$$Y_i \sim \text{BIN}(m_i, \pi_i), \quad i = 1, 2, \dots$$

- Explanatory Variables:

$$x_{i1} = \begin{cases} 1 & \text{Intensive Care} \\ 0 & \text{Regular Care} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{Clinic A} \\ 0 & \text{Clinic B} \end{cases}$$

$$x_{i3} = x_{i1}x_{i2} = \begin{cases} 1 & \text{Intensive care and Clinic A} \\ 0 & \text{Otherwise} \end{cases}$$

- We will use the context of this example to illustrate how to:
 - fit (simple and multiple) logistic regression models using R, and
 - interpret regression parameters.

Model 1: Level of Care only model

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1}.$$

- $\beta_0 = \log$ odds of mortality for babies born to mothers treated with regular care.
- $\beta_1 = \log$ odds ratio of mortality for babies born to mothers treated with intensive vs regular care.

Level of Care	Clinic	\mathbf{x}_i^\top	$\log(\pi_i/(1 - \pi_i))$
Intensive	—	$(1, 1)^\top$	$\beta_0 + \beta_1$
Regular	—	$(1, 0)^\top$	β_0

Model 2: Main effects model

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

Level of Care	Clinic	\mathbf{x}_i^\top	$\log(\pi_i/(1 - \pi_i))$
Intensive	A	$(1, 1, 1)^\top$	$\beta_0 + \beta_1 + \beta_2$
Regular	A	$(1, 0, 1)^\top$	$\beta_0 + \beta_2$
Intensive	B	$(1, 1, 0)^\top$	$\beta_0 + \beta_1$
Regular	B	$(1, 0, 0)^\top$	β_0

- β_0 = **log odds** of mortality with regular care at Clinic B.
- β_1 = **log odds ratio** of mortality for babies born to mothers treated with **intensity vs regular** care at the *same clinic*.
- β_2 = **log odds ratio** of mortality for babies born to mothers treated at **Clinic A vs Clinic B** at the *same level of care*.

Model 3: Interaction model

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

Level of Care	Clinic	\mathbf{x}_i^\top	$\log(\pi_i/(1 - \pi_i))$
Intensive	A	$(1, 1, 1)^\top$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$
Regular	A	$(1, 0, 1)^\top$	$\beta_0 + \beta_2$
Intensive	B	$(1, 1, 0)^\top$	$\beta_0 + \beta_1$
Regular	B	$(1, 0, 0)^\top$	β_0

- β_1 = **log odds ratio** of mortality for babies born to mothers treated with **intensity vs regular** care at *Clinic B*.
- $\beta_1 + \beta_3$ = **log odds ratio** of mortality for babies born to mothers treated with **intensity vs regular** care at *Clinic A*.
- β_2 = **log odds ratio** of mortality for babies born to mothers treated at **Clinic A vs Clinic B** with *regular* care.
- $\beta_2 + \beta_3$ = **log odds ratio** of mortality for babies born to mothers treated at **Clinic A vs Clinic B** with *intensive* care.
- β_3 represents the **difference in log odds ratios**.
- If $\beta_3 = 0$ then association between mortality and level of care does not depend on clinic.
- Equivalently, if $\beta_3 = 0$ then the association between mortality and clinic does not depend on level of care.

Data file prenatal.dat

	clinic	loc	y	m
1	0	0	34	231
2	0	1	4	27
3	1	0	12	188
4	1	1	16	309

- The first line contains the variable names/labels.
- We are using indicator variables for the explanatory variables:

$$\begin{aligned}x_{i1} &= \text{loc} && (1 \text{ for Intensive, } 0 \text{ for Regular}) \\x_{i2} &= \text{clinic} && (1 \text{ for Clinic A, } 0 \text{ for Clinic B})\end{aligned}$$

- The variable y records the number of deaths (events).

Fit GLMs using R

The `glm()` function in R is used to fit the generalized linear models:

```
fit = glm(formula, family = (link = ), data = ).
```

- formula: a linear formula describing the model, e.g.,

```
resp ~ loc + clinic.
```

- family: a description of the exponential family distribution and link function to be used in the model, e.g.,

```
family = binomial, gaussian, poisson, Gamma, etc..
```

```
link = logit, log, loglog, cloglog, identity, probit, etc..
```

- The default is the canonical link.

R Code and Output for Analysis of Prenatal Care data

For binomial data, we need to construct “resp” variable as the pair $(y_i, m_i - y_i)$.

```
# read file prenatal.data
prenatal.dat = read.table("prenatal.dat", header = T)
# construct the binomial response for the logistic regression
# analysis
prenatal.dat$resp = cbind(prenatal.dat$y, prenatal.dat$m - prenatal.dat$y)
prenatal.dat
```

	clinic	loc	y	m	resp.1	resp.2
1	0	0	34	231	34	197
2	0	1	4	27	4	23
3	1	0	12	188	12	176
4	1	1	16	309	16	293

The logistic regression models are fit using the `glm()` commands like:

```
# fit the logistic model using the glm function
model1 = glm(resp ~ loc, family = binomial(link = logit), data = prenatal.dat)
summary(model1)
```

Fit of Model 1: Level of Care Model

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1}.$$

```
# fit the logistic model using the glm function
model1 = glm(resp ~ loc, family = binomial(link = logit), data = prenatal.dat)
summary(model1)$coefficients
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.0929370	0.1562692	-13.393150	6.630754e-41
loc	-0.6670729	0.2785400	-2.394891	1.662530e-02

Components of the `summary()` output for `glm` objects:

- **Estimate**: the maximum likelihood estimates of the regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.
- **Std. Error**: estimated standard errors, the square root of the diagonal of the inverse of the Information matrix:

$$\text{se}(\hat{\beta}_j) = \sqrt{[I^{-1}(\hat{\beta})]_{jj}} = \sqrt{I^{jj}(\hat{\beta})}.$$

- **z value**: Wald-type test statistics for testing the hypotheses:

$$H_0: \beta_j = 0 \text{ vs } H_A: \beta_j \neq 0.$$

- **Pr(>|z|)**: p -value for above Wald test.

For this model:

- β_1 is the log odds ratio of mortality for infants born to mothers treated with intensive versus regular care.

Hypothesis test for β_j

- We may wish to test:

$$H_0: \beta_j = \beta^* \text{ versus } H_A: \beta_j \neq \beta^*.$$

- The general **Wald** result for a single parameter β_j is:

$$(\hat{\beta}_j - \beta^*)^2 (I^{jj}(\hat{\beta}))^{-1} \sim \chi_1^2,$$

equivalently $\frac{\hat{\beta}_j - \beta^*}{\text{se}(\hat{\beta}_j)} \sim \mathcal{N}(0, 1)$ where $\text{se}(\hat{\beta}_j) = \sqrt{I^{jj}(\hat{\beta})}$.

- We can find the p -value of this test using:

$$p = 2 \mathbb{P}\left(Z > \frac{|\hat{\beta}_j - \beta^*|}{\text{se}(\hat{\beta}_j)}\right).$$

- The `summary()` output gives the test statistics and p -values for testing

$$H_0: \beta_j = 0 \text{ vs } H_A: \beta_j \neq 0.$$

Hypothesis test for β_1 from Model 1: Level of Care Model

```
summary(model1)$coefficients
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.0929370	0.1562692	-13.393150	6.630754e-41
loc	-0.6670729	0.2785400	-2.394891	1.662530e-02

- We wish to test:

$$H_0: \beta_1 = 0 \text{ vs } H_A: \beta_1 \neq 0$$

- Wald test:

$$z = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_1)} = \frac{-0.6671}{0.2785} = -2.3949$$

- p -value:

$$p = 2\mathbb{P}(Z > |-2.3949|) = 0.0166 < 0.05$$

- Therefore, we reject the null hypothesis that $\beta_1 = 0$.

- Estimate of OR for Mortality for Intensive vs Regular Care:

$$\hat{\psi} = \exp\{\hat{\beta}_1\} = \exp\{-0.6670729\} = 0.51.$$

- Confidence Interval for OR:

$$\begin{aligned}\exp\{\hat{\beta}_1 \pm 1.96 \text{se}(\hat{\beta}_1)\} &= \exp\{-0.6671 \pm 1.96(0.2785)\} \\ &= (\exp\{-1.2130\}, \exp\{-0.1211\}) \\ &= (0.30, 0.89)\end{aligned}$$

- The estimate and Wald 95 % CI here match those found previously from the 2×2 table analysis. That is, the 2×2 table analysis is equivalent to a simple logistic regression with a single binary covariate.

Fit of Model 2: Main Effects Model

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

```
model2 <- glm(resp ~ loc + clinic, family = binomial(link = logit),  
  data = prenatal.dat)  
summary(model2)$coefficients
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.7410476	0.1784691	-9.7554560	1.748132e-22
loc	-0.1503053	0.3301670	-0.4552402	6.489365e-01
clinic	-0.9862793	0.3089322	-3.1925427	1.410261e-03

- What is the OR for mortality for Intensive vs Regular Care, now controlling for Clinic?

$$\widehat{\text{OR}} = \hat{\psi} = \exp\{-0.1503\} = 0.86.$$

- 95 % CI:

$$\exp\{-0.1503 \pm 1.96 \times 0.3302\} = (0.4505, 1.6436).$$

Fit of Model 3: Interaction Model

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

```
model3 <- glm(resp ~ loc + clinic + loc * clinic, family = binomial(link = logit),
  data = prenatal.dat)
summary(model3)$coefficients
```

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-1.756843204	0.1857092	-9.46018403	3.074017e-21
loc	0.007643349	0.5726827	0.01334657	9.893513e-01
clinic	-0.928734141	0.3514300	-2.64272868	8.224091e-03
loc:clinic	-0.229649891	0.6949054	-0.33047646	7.410400e-01

Level of Care	Clinic	\mathbf{x}_i^\top	$\log(\pi_i/(1 - \pi_i))$
Intensive	A	$(1, 1, 1)^\top$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$
Regular	A	$(1, 0, 1)^\top$	$\beta_0 + \beta_2$
Intensive	B	$(1, 1, 0)^\top$	$\beta_0 + \beta_1$
Regular	B	$(1, 0, 0)^\top$	β_0

- What is the OR for Mortality for Intensive vs Regular Care at Clinic A?

$$\text{OR} = \psi = \exp\{\beta_1 + \beta_3\} \implies \hat{\psi} = \exp\{0.0076 - 0.2296\} = 0.8.$$

- $\text{se}(\hat{\beta}_1 + \hat{\beta}_3)$ is required for calculation of 95 % CI.
 - Recall $\text{Var}(\hat{\beta}) = \mathbf{I}^{-1}(\hat{\beta})$, now for any linear function of β 's, e.g., $\mathbf{c}\beta$ where \mathbf{c} is a row vector of constants, then MLE of $\mathbf{c}\beta$ is $\mathbf{c}\hat{\beta}$, and $\text{se}(\mathbf{c}\hat{\beta}) = \sqrt{\mathbf{c}\mathbf{I}^{-1}(\hat{\beta})\mathbf{c}^\top}$.
- Therefore, $\log(\psi) = \beta_1 + \beta_3 = \mathbf{c}\beta$, $\mathbf{c} = (0, 1, 0, 1)$. In R, `vcov(model3)` gives $\mathbf{I}^{-1}(\hat{\beta})$.
- What is OR for Mortality for Intensive vs Regular Care at Clinic B?

$$\text{OR} = \psi = \exp\{\beta_1\} \implies \hat{\psi} = \exp\{0.0076\} = 1.01.$$

Topic 3c: Likelihood Ratio Test for Logistic Regression Models

Logistic Regression Models

Recall major developments of Binomial logistic regression from last topic 3b: $Y_i \sim \text{BIN}(m_i, \pi_i)$, $i = 1, \dots, n$ independently, with covariate vector \mathbf{x}_i and

$$\log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

- Estimation: $\hat{\beta}$ come from Fisher scoring using R function `glm()`.
- Interpretation: $\exp\{\beta_j\}$ has OR interpretation.
- Hypothesis tests of $H_0: \beta_j = 0$ using Wald statistic.
- Confidence Intervals: $\hat{\beta}_j \pm z_{1-\alpha/2} \text{se}(\hat{\beta}_j)$.

Likelihood for Logistic Regression Models

- Log-likelihood for Binomial Distribution:

$$\begin{aligned}\ell &= \log \left(\prod_{i=1}^n \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i} \right) \\ &= \sum_{i=1}^n y_i \log \left(\frac{\pi_i}{1 - \pi_i} \right) + m_i \log(1 - \pi_i).\end{aligned}$$

- Using logit link we can re-parameterize the log-likelihood in terms of β :

$$\log \left(\frac{\pi_i}{1 - \pi_i} \right) = \mathbf{x}_i^\top \beta, \quad \pi_i = \frac{\exp\{\mathbf{x}_i^\top \beta\}}{1 + \exp\{\mathbf{x}_i^\top \beta\}}.$$

- Log likelihood for logistic regression:

$$\ell = \sum_{i=1}^n y_i \mathbf{x}_i^\top \beta - m_i \log(1 + \exp\{\mathbf{x}_i^\top \beta\}).$$

- Maximization of log-likelihood $\ell(\beta)$ gives MLE $\hat{\beta}$, and

- estimated probability of response:

$$\hat{\pi}_i = e^{\mathbf{x}_i^\top \hat{\beta}} / (1 + e^{\mathbf{x}_i^\top \hat{\beta}}) = \text{expit}(\mathbf{x}_i^\top \hat{\beta}),$$

- estimated number of responses: $\hat{y}_i = m_i \hat{\pi}_i$.

- Questions:

- How good is the model? How well do the estimated number of events \hat{y}_i approximate the observed data y_i ? (**goodness of fit**).
- How much worse is the fit of a model when several of the covariates are excluded? (**nested models**):

$$H_0: \beta_k = \beta_{k+1} = 0 \text{ vs } H_A: \beta_k \neq 0 \text{ or } \beta_{k+1} \neq 0.$$

Likelihood Ratio Test (General Setting)

- Suppose $\ell(\theta)$ is the likelihood for a q -dimension parameter vector θ and let

- $\tilde{\theta}$ be the q -dim MLE of θ (unconstrained/**saturated**, $q = n$),
- $\hat{\theta}$ be the p -dim MLE of θ (constrained/**unsaturated**, $p < q$).

- Hypotheses:

- H_0 : the constrained model is adequate (i.e., as good as the unconstrained).
- H_A : constrained model is not adequate.

- Recall the Likelihood Ratio (LR) result:

$$\text{Under } H_0: \quad -2 \log \left(\frac{L(\hat{\theta})}{L(\tilde{\theta})} \right) = -2[\ell(\hat{\theta}) - \ell(\tilde{\theta})] \sim \chi_{q-p}^2.$$

- Reject H_0 at θ if

$$p\text{-value} = \mathbb{P}(\chi_{q-p}^2 > -2[\ell(\hat{\theta}) - \ell(\tilde{\theta})]) < \alpha.$$

Likelihood Ratio Test (Logistic Regression Model)

- **Saturated** (unconstrained) model MLEs:

$$\tilde{\pi}_i = \frac{y_i}{m_i}, \quad i = 1, \dots, n.$$

- Binomial MLE without imposing any constraint.
- We will have $\tilde{y}_i = m_i \tilde{\pi}_i = y_i$, **a perfect fit!**

- **Unsaturated** (constrained) model MLEs:

$$\hat{\pi}_i = \text{expit}(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}).$$

- Regression models are a way of imposing constraints on the estimation of π_i through p -dim regression coefficients $\boldsymbol{\beta}$.
- We will have fitted number of responses $\hat{y}_i = m_i \hat{\pi}_i = m_i \text{expit}(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}})$.

- **Hypotheses:**

- H_0 : the p -dim model, e.g., $\text{logit}(\pi_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$ is adequate.
- H_A : the p -dim model, e.g., $\text{logit}(\pi_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$ is not adequate *compared to the n -dim saturated model*.

- **Likelihood Ratio Statistic** (also referred to as the **Deviance**):

$$\begin{aligned} D &= -2[\ell(\hat{\boldsymbol{\pi}}) - \ell(\tilde{\boldsymbol{\pi}})] \\ &= -2\left(\sum_{i=1}^n \left(y_i \log(\hat{\pi}_i) + (m_i - y_i) \log(1 - \hat{\pi}_i)\right) - \sum_{i=1}^n \left(y_i \log(\tilde{\pi}_i) + (m_i - y_i) \log(1 - \tilde{\pi}_i)\right)\right) \\ &= -2\sum_{i=1}^n \left(y_i \log\left(\frac{y_i}{m_i \hat{\pi}_i}\right) + (m_i - y_i) \log\left(\frac{m_i - y_i}{m_i(1 - \hat{\pi}_i)}\right)\right). \end{aligned}$$

- The **LR/Deviance** can also be written in a general form as:

$$D = 2 \sum_{i=1}^n \sum_{j=1}^2 \left(O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right) \right).$$

- $O_{i1} = y_i$, $E_{i1} = m_i \hat{\pi}_i$ (observed and expected # of events).
- $O_{i2} = m_i - y_i$, $E_{i2} = m_i(1 - \hat{\pi}_i)$ (observed and expected # of non-events).

- We expect $D \sim \chi_{n-p}^2$ under H_0 , and reject H_0 if $\mathbb{P}(\chi_{n-p}^2 > D) < \alpha$.
 - Unfortunately, this is not a great approximation.
 - Approximation is much better for testing nested unsaturated models though.

Example: Prenatal Care Data

- Model 2: Main Effects Model,

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

- H_0 : Model 2 is adequate.
- H_A : Model 2 is not adequate compared to the saturated model.

- In R, the `summary()` output D is reported as the **Residual Deviance**.

```

model2 = glm(resp ~ loc + clinic, family = binomial(link = logit),
  data = prenatal.dat)
summary(model2)

Call:
glm(formula = resp ~ loc + clinic, family = binomial(link = logit),
  data = prenatal.dat)

Deviance Residuals:
    1      2      3      4 
-0.08521  0.25805  0.13909 -0.11719

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7410     0.1785  -9.755  < 2e-16 ***
loc           -0.1503     0.3302  -0.455  0.64894
clinic        -0.9863     0.3089  -3.193  0.00141 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16.91763  on 3  degrees of freedom
Residual deviance:  0.10693  on 1  degrees of freedom
AIC: 23.262

Number of Fisher Scoring iterations: 3

```

- Deviance: $D = 0.10693$.
- p -value: $\mathbb{P}(\chi_{n-p}^2 > D) = \mathbb{P}(\chi_1^2 > D) = 0.7436689 \gg 0.05$.
- Do not reject the null hypothesis that Model 2 is adequate.

Pearson Statistic

- The Pearson statistic is another statistic that can be used for assessing “overall” fit (or goodness of fit) of a Binomial model:

$$P = \sum_{i=1}^n \frac{(y_i - m_i \hat{\pi}_i)^2}{m_i \hat{\pi}_i (1 - \hat{\pi}_i)}.$$

- As with LR/Deviance statistic, $P \sim \chi_{n-p}^2$ under H_0 : the model is adequate.
- Note that P has the general form:

$$P = \sum_i \frac{(O_i - E_i)^2}{V_i}.$$

- The χ^2 approximation is a bit better than for deviance statistic D .
- Both are poor if the sample size (m_i) is small though.

Testing Nested Non-saturated Models

- The previous LR/Deviance test was for an unsaturated model vs a saturated model.
- Now consider two unsaturated models ($p < q < n$).

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} \quad (1)$$

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \cdots + \beta_{q-1} x_{iq-1} \quad (2)$$

- Model (1) is *nested* within Model (2).
- H_0 : Model (1) fits the data as well as Model (2).
 - H_0 : $\beta_p = \cdots = \beta_{q-1} = 0$.
- H_A : Model (1) is inadequate compared to Model (2).
 - H_A : at least one of $\beta_p, \dots, \beta_{q-1} \neq 0$.

Model	Dimension	MLEs
(1) Reduced model	p	$\hat{\pi}_i$
(2) Full model	q	$\tilde{\pi}_i$
Saturated model	n	$\tilde{\tilde{\pi}}_i$

- LR/Deviance test of Model (1) vs Saturated Model:

$$D_0 = -2(\ell(\hat{\pi}) - \ell(\tilde{\tilde{\pi}})).$$

- LR/Deviance test of Model (2) vs Saturated Model:

$$D_A = -2(\ell(\tilde{\pi}) - \ell(\tilde{\tilde{\pi}})).$$

- Now, we wish to conduct LR test of Model (1) vs Model (2):

$$\Delta D = D_0 - D_A = -2(\ell(\hat{\pi}) - \ell(\tilde{\pi})).$$

- It can be shown that under H_0 : Model (1) is as adequate as Model (2),

$$\Delta D \sim \chi_{q-p}^2.$$

- This approximation is much better than when testing an unsaturated model vs the saturated model.
- If $p = \mathbb{P}(\chi_{q-p}^2 > \Delta D) < \alpha$, reject H_0 .
 - Reduced model does not fit the data as well as Full model.
 - One or more of covariates x_{ip}, \dots, x_{iq-1} is important (i.e., associated with the response).

Example: Prenatal Care Data

- Summary of Deviance (“residual deviance”) from R output:
- Compare nested models:
 - Model 2: $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$.
 - Model 4: $\text{logit}(\pi_i) = \beta_0 + \beta_2 x_{i2}$.

Model	Covariates	Deviance	Parameters	$n - p$
1	loc	10.814378	2	2
2	loc + clinic	0.106928	3	1
3	loc + clinic + loc*clinic	0	4	0
4	clinic	0.314841	2	2

- Is level of care associated with fetal mortality after accounting for clinic?
 - H_0 : Model 4 is as adequate as Model 2 (e.g., $\beta_1 = 0$).
 - H_A : Model 4 is inadequate compared to Model 2 (e.g., $\beta_1 \neq 0$).
- LR test for comparing Model 4 vs Model 2, or equivalently testing hypotheses:

$$H_0: \beta_1 = 0 \text{ vs } H_A: \beta_1 \neq 0.$$

- We do not reject H_0 of no association between level and care and fetal mortality after controlling for Clinic.

```
model2 = glm(resp ~ loc + clinic, family = binomial, data = prenatal.dat)
model4 = glm(resp ~ clinic, family = binomial, data = prenatal.dat)
D = model4$deviance - model2$deviance
1 - pchisq(D, 2 - 1)

[1] 0.6484081
```

- This implies that level of care is no longer important when clinic is included in the model.
- It also implies that Model 4 is as adequate compared to Model 2.
- Finally, when testing a single parameter, e.g., $H_0: \beta_1 = 0$, LR/Deviance test result is consistent with the Wald test result provided in the R output:

```
model2 = glm(resp ~ loc + clinic, family = binomial, data = prenatal.dat)
summary(model2)

Call:
glm(formula = resp ~ loc + clinic, family = binomial, data = prenatal.dat)

Deviance Residuals:
    1      2      3      4 
-0.08521  0.25805  0.13909 -0.11719

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.7410     0.1785  -9.755  < 2e-16 ***
loc           -0.1503     0.3302  -0.455  0.64894
clinic        -0.9863     0.3089  -3.193  0.00141 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 16.91763  on 3  degrees of freedom
```

```
Residual deviance: 0.10693 on 1 degrees of freedom
AIC: 23.262
```

```
Number of Fisher Scoring iterations: 3
```

Summary of LR/Deviance Test for Logistic Regression

- For Binomial GLM with logit link the LR/Deviance test statistic is:

$$D = \sum_{i=1}^n 2 \left(y_i \log \left(\frac{y_i}{m_i \hat{\pi}_i} \right) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i (1 - \hat{\pi}_i)} \right) \right).$$

- This is reported as the “Residual Deviance” in R glm summary output.
- Deviance statistic D can be used to:
 - Test adequacy/goodness of fit of a non-saturated logistic model:

$$D \stackrel{H_0}{\sim} \chi_{n-p}^2.$$

- Compare the fit of two nested-non saturated logistic models:

$$\Delta D = D_0 - D_A \stackrel{H_0}{\sim} \chi_{q-p}^2.$$

WEEK 5
3rd to 8th October

Topic 3d: Residuals for Binomial Data and Neuroblastoma Example

Recall: Residuals in Linear Regression Models

- Normal linear regression models (STAT 331),

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2).$$

- Fitted values:

$$\hat{y}_i = \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}.$$

- Residuals:

$$r_i = y_i - \hat{y}_i.$$

- The overall fit of the model and validity of the model assumptions are assessed using various [residual plots](#), e.g.,
 - Residuals r_i vs fitted value \hat{y}_i plot (check normality and constant variance).
 - QQ plot of residuals r_i 's (check normality).

Residuals for Binomial Data

- When fit a logistic regression model to Binomial data, we evaluate the adequacy of the model by using the LR deviance test statistic:

$$D = \sum_{i=1}^n 2 \left(y_i \log \left(\frac{y_i}{m_i \hat{\pi}_i} \right) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i (1 - \hat{\pi}_i)} \right) \right) \\ = \sum_{i=1}^n d_i.$$

- Deviance Residual:

$$r_i^D = \text{sign}(y_i - m_i \hat{\pi}_i) \sqrt{|d_i|}.$$

- Under H_0 : the model is adequate:

$$D = \sum_{i=1}^n d_i \stackrel{\text{approx}}{\sim} \chi_{n-p}^2 \implies r_i^D \stackrel{\text{approx}}{\sim} \mathcal{N}(0, 1).$$

- We can use the plots of deviance residuals to assess whether r_i^D 's look independent observations from $\mathcal{N}(0, 1)$.

Example: Prenatal Care Data

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \text{clinic}_i$$

```
model4 <- glm(resp ~ clinic, family = binomial(link = logit), data = prenatal.dat)
summary(model4)$deviance.resid
```

```
      1      2      3      4
-0.004318004  0.012618764  0.436709170 -0.352063013
```

- Pearson Residual:

$$r_i^P = \frac{y_i - m_i \hat{\pi}_i}{\sqrt{m_i \hat{\pi}_i (1 - \hat{\pi}_i)}} = \frac{O_i - E_i}{\sqrt{V_i}}.$$

- Under H_0 : the model is adequate,

$$r_i^P \sim \mathcal{N}(0, 1).$$

- Note: if $m_i \hat{\pi}_i < 5$ (or $m_i (1 - \hat{\pi}_i) < 5$) for one or more cases, we should be concerned about the validity of the approximation (χ^2 or $\mathcal{N}(0, 1)$) and hence our conclusions.

Prognosis for Children with Neuroblastoma

- A study is conducted to investigate the probability of *disease-free survival* (surviving 2 years free of disease) following the treatment for neuroblastoma.
- Associated risk factors include *age at diagnosis* and *stage of disease at diagnosis*.
 - Cell entries are of the form y/m with y representing the number of patients surviving 2 years, and m representing the number of patients in that age-stage combination at the start of the study.
- As an initial look at the data, consider the marginal distributions.
 - Higher chance of survival at younger age at diagnosis.
 - Higher chance of survival with lower stage of disease at diagnosis.

Age (months)	Stage				
	I	II	III	IV	V
0-11	11/12	15/16	2/4	5/18	18/19
12-23	3/4	3/7	5/8	0/25	1/3
24+	4/5	4/12	3/15	3/93	2/5

Age (months)	Stage					Total
	I	II	III	IV	V	
0-11	11/12	15/16	2/4	5/18	18/19	51/69
12-23	3/4	3/7	5/8	0/25	1/3	12/47
24+	4/5	4/12	3/15	3/93	2/5	16/130
Total	18/21	22/35	10/27	8/136	21/27	79/246

Setup Regression Models for Neuroblastoma Data

- Response Variable:

- Y_i is the number of 2-yr disease-free survivors out of m_i total children in group i , assume $Y_i \sim \text{BIN}(m_i, \pi_i)$, $i = 1, \dots, 15$, and

$$\pi_i = \mathbb{P}(\text{2-yr disease-free survival in group } i).$$

- Explanatory Variables:

- **Age** (0-11, 12-23, 24+ months); age 0-11 month is the baseline/reference,

$$x_{i1} = \begin{cases} 1 & \text{if age 12-23 months} \\ 0 & \text{o.w.} \end{cases} \quad x_{i2} = \begin{cases} 1 & \text{if age 24+ months} \\ 0 & \text{o.w.} \end{cases}$$

- **Stage** (I, II, III, IV, V); stage 1 is the baseline/reference,

$$x_{i3} = \begin{cases} 1 & \text{stage II} \\ 0 & \text{o.w.} \end{cases} \quad x_{i4} = \begin{cases} 1 & \text{if stage III} \\ 0 & \text{o.w.} \end{cases}$$

$$x_{i5} = \begin{cases} 1 & \text{if stage IV} \\ 0 & \text{o.w.} \end{cases} \quad x_{i6} = \begin{cases} 1 & \text{if stage V} \\ 0 & \text{o.w.} \end{cases}$$

- Consider the following logistic regression models:

- Model 1: Age & Stage

$$\text{logit}(\pi_i) = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2}}_{\text{Age}} + \underbrace{\beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}}_{\text{Stage}}.$$

- Model 2: Age only

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

- Model 3: Stage only

$$\text{logit}(\pi_i) = \beta_0 + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}.$$

Fitting Logistic Regression Models Using R

```

neuro.dat = read.table("neuro.dat", header = T)
neuro.dat

  age stage  y  m
1   1     1  1 11 12
2   1     2  2 15 16
3   1     3  2  4
4   1     4  5 18
5   1     5 18 19
6   2     1  3  4
7   2     2  3  7
8   2     3  5  8
9   2     4  0 25
10  2     5  1  3
11  3     1  4  5
12  3     2  4 12
13  3     3  3 15
14  3     4  3 93
15  3     5  2  5

# here we construct the response variable for logistic
# regression
neuro.dat$resp = cbind(neuro.dat$y, neuro.dat$m - neuro.dat$y)
neuro.dat

  age stage  y  m resp.1 resp.2
1   1     1  1 11 12     11     1
2   1     2  2 15 16     15     1
3   1     3  2  4      2     2
4   1     4  5 18      5     13
5   1     5 18 19     18     1
6   2     1  3  4      3     1
7   2     2  3  7      3     4
8   2     3  5  8      5     3
9   2     4  0 25      0     25
10  2     5  1  3      1     2
11  3     1  4  5      4     1
12  3     2  4 12      4     8
13  3     3  3 15      3     12
14  3     4  3 93      3     90
15  3     5  2  5      2     3

neuro.dat$age <- factor(neuro.dat$age, levels = c(1, 2, 3), labels = c("0-11",
"12-23", "24+"))
neuro.dat$stage <- factor(neuro.dat$stage, levels = c(1, 2, 3, 4,
5), labels = c("I", "II", "III", "IV", "V"))

```

Summary of Model 1: Age & Stage

$$\text{logit}(\pi_i) = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2}}_{\text{Age}} + \underbrace{\beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}}_{\text{Stage}}.$$

```
Call:
glm(formula = resp ~ age + stage, family = binomial(link = logit),
    data = neuro.dat)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.47408  -0.61913  -0.09643   0.53163   1.52114
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.3175     0.7721   4.297 1.73e-05 ***
age12-23     -2.1181     0.5736  -3.693 0.000222 ***
age24+       -2.6130     0.5017  -5.208 1.91e-07 ***
stageII      -1.2529     0.7837  -1.599 0.109860
stageIII     -1.7759     0.8003  -2.219 0.026478 *
stageIV      -4.3678     0.7902  -5.528 3.25e-08 ***
stageV       -1.0222     0.8644  -1.183 0.236980
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 162.832 on 14 degrees of freedom
Residual deviance: 9.625 on 8 degrees of freedom
AIC: 55.382
```

```
Number of Fisher Scoring iterations: 4
```

- Before interpreting these results too much, we should look to see how good the fit is to the data.

- fv1: $\hat{\pi}_i = \text{expit}(\mathbf{x}_i^\top \hat{\beta})$.
- yhat: $\hat{y}_i = m_i \hat{\pi}_i$.
- rd1: r_i^D (deviance residual).
- rp1: r_i^P (Pearson residual).

```
y = neuro.dat$y
m = neuro.dat$m
fv1 = model1$fitted.values
yhat = m * fv1
rd1 = residuals.glm(model1, "deviance")
rp1 = (y - m * fv1)/sqrt(m * fv1 * (1 - fv1))
cbind(rd1, rp1, yhat, y)
```

```
      rd1      rp1      yhat y
1 -0.77808711 -0.91184050 11.580304 11
2  0.68559153  0.63381666 14.198641 15
3 -1.47407847 -1.69888561  3.294804  2
4  0.17884403  0.18019371  4.665014  5
5  0.63431439  0.58779486 17.261237 18
6 -0.08658336 -0.08736144  3.073705  3
7 -0.30801258 -0.30734393  3.406432  3
```

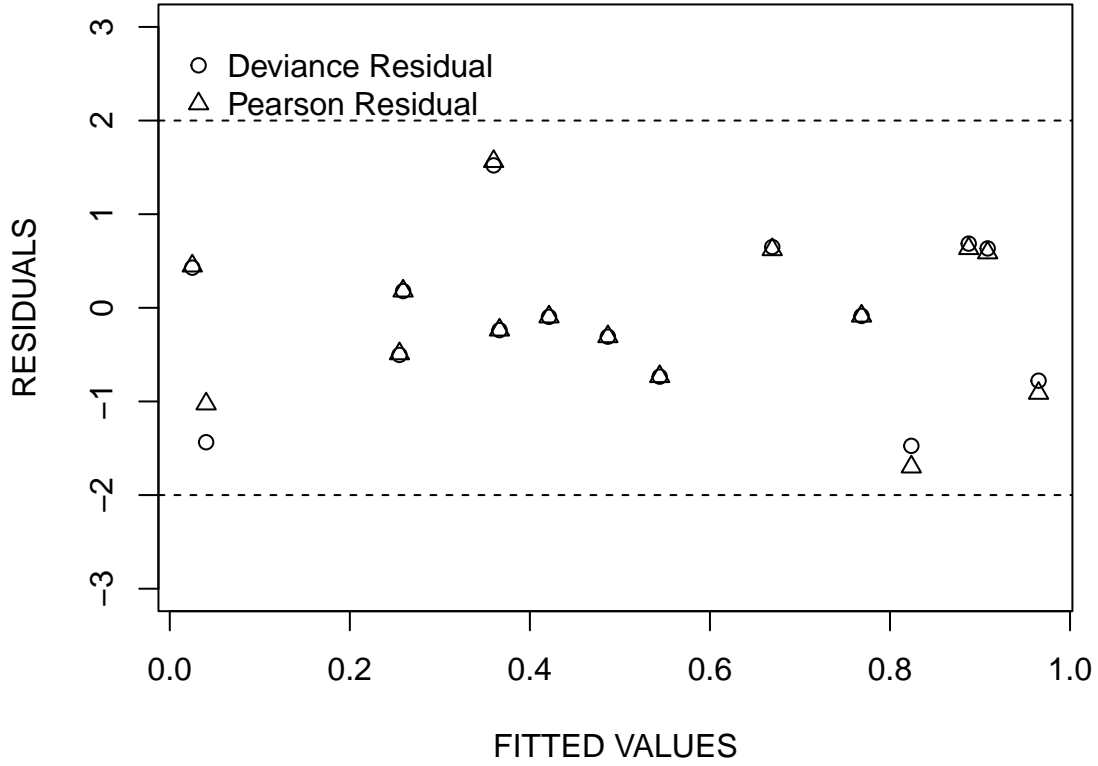


Figure 1: Plot of Residuals by Fitted Values for Neuroblastoma Data based on Logistic Regression Model with main effects of Age and Stage.

8	1.52114028	1.56325351	2.877982	5
9	-1.43545385	-1.02556686	1.009324	0
10	-0.73520283	-0.73328264	1.632557	1
11	0.64949774	0.62163765	3.345991	4
12	-0.23825133	-0.23663531	4.394927	4
13	-0.50305728	-0.48993834	3.827214	3
14	0.42894854	0.44782015	2.325662	3
15	-0.09643089	-0.09619454	2.106206	2

- Residuals are a random scatter around 0 and $\in (-2, 2)$ therefore r_i^D (or r_i^P) $\sim \mathcal{N}(0, 1)$. Therefore, model 1 is adequate.
- We can test H_0 : model 1 is adequate using LR/D statistic $p\text{-value} = \mathbb{P}(\chi_8^2 > 9.625) < 0.05$, do not reject H_0 .

Summary of Model 2: Age only

- Now we consider simplifying the model further by examining the decrease in the quality of the fit that results from dropping the stage variable(s).

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

```
model3 = glm(resp ~ age, family = binomial(link = logit), data = neuro.dat)
summary(model3)
```

Call:
glm(formula = resp ~ age, family = binomial(link = logit), data = neuro.dat)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.0853	-0.3591	1.5613	2.0684	3.4667

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	1.0415	0.2742	3.799	0.000145	***
age12-23	-2.1119	0.4325	-4.883	1.05e-06	***
age24+	-3.0051	0.3827	-7.853	4.06e-15	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 162.832 on 14 degrees of freedom
Residual deviance: 83.583 on 12 degrees of freedom
AIC: 121.34

Number of Fisher Scoring iterations: 5

Summary of Model 3: Stage only

- Now we fit the model excluding the age variable to examine the drop in the quality of fit from model 1.

$$\text{logit}(\pi_i) = \beta_0 + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}.$$

```
model2 = glm(resp ~ stage, family = binomial(link = logit), data = neuro.dat)
summary(model2)
```

Call:
glm(formula = resp ~ stage, family = binomial(link = logit),
data = neuro.dat)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-----	----	--------	----	-----

```

-2.0699  -1.5375  -0.5639   1.0444   2.9391

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7918     0.6236   2.873  0.00406 **
stageII      -1.2657     0.7150  -1.770  0.07671 .
stageIII     -2.3224     0.7401  -3.138  0.00170 **
stageIV      -4.5643     0.7223  -6.319 2.63e-10 ***
stageV       -0.5390     0.7766  -0.694  0.48768

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 162.832  on 14  degrees of freedom
Residual deviance:  42.446  on 10  degrees of freedom
AIC: 84.203

Number of Fisher Scoring iterations: 5

```

Testing Nested Models

- Now we can compare nested models using [LR/Deviance](#) Tests:

Model	Covariates	Deviance (D)	Parameters (p)	DF ($n - p$)
M1	Age & Stage	9.625	7	8
M2	Age	83.583	3	12
M3	Stage	42.446	5	10

- Recall:

$$\Delta D = D_0 - D_A = -2(\ell(\hat{\pi}) - \ell(\tilde{\pi})) \sim \chi_{q-p}^2$$

- D_0 and D_A are deviances from the [reduced](#) and [full](#) models respectively.
- $\hat{\pi}$ and $\tilde{\pi}$ represents the MLEs from the [reduced](#) and [full](#) models respectively.

Objective: Pick the model that best represents the important associations between the outcome and explanatory variables.

1. Is Stage important?

$H_0: \beta_3 = \dots = \beta_6 = 0$ (Model 2 is as adequate as Model 1)

H_A : at least one of them is not 0 (Model 2 is not adequate)

$$\Delta D = D_2 - D_1 = 83.583 - 9.625 = 73.958$$

$$p = \mathbb{P}(\chi_{7-3}^2 > 73.958) < 0.001$$

```
1 - pchisq(83.583 - 9.625, 7 - 3)
```

```
[1] 3.330669e-15
```

We reject H_0 and conclude that there is evidence that Stage is important.

2. Is Age important?

$$H_0: \beta_1 = \beta_2 = 0$$

(Model 3 is as adequate as Model 1)

$$H_A: \text{at least one of them is not 0}$$

(Model 3 is not adequate)

$$\Delta D = D_3 - D_1 = 42.446 - 9.625 = 32.821$$

$$p = \mathbb{P}(\chi_{7-5}^2 > 32.821) < 0.001$$

```
1 - pchisq(42.446 - 9.625, 7 - 5)
```

```
[1] 7.464666e-08
```

We reject H_0 and conclude that there is evidence that Age is important.

3. Do we need an Age*Stage interaction?

```
1 - pchisq(model1$deviance, model1$df.residual)
```

```
[1] 0.292341
```

$$p\text{-value} = \mathbb{P}(\chi_8^2 > 9.625) = 0.292 > 0.05$$

- Model with age, stage, and age*stage is the saturated model!
- Do not reject H_0 : model 1 is as adequate as the saturated model (interaction model).
- Do not need to consider age*stage.

Interpret the Selected Model

So we select [Model 1](#) for interpretation.

```
model1 = glm(resp ~ age + stage, family = binomial(link = logit),
  data = neuro.dat)
summary(model1)
```

Call:

```
glm(formula = resp ~ age + stage, family = binomial(link = logit),
  data = neuro.dat)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-1.47408	-0.61913	-0.09643	0.53163	1.52114

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	3.3175	0.7721	4.297	1.73e-05	***
age12-23	-2.1181	0.5736	-3.693	0.000222	***
age24+	-2.6130	0.5017	-5.208	1.91e-07	***
stageII	-1.2529	0.7837	-1.599	0.109860	


```
stageIII    -1.7759    0.8003   -2.219  0.026478 *
stageIV     -4.3678    0.7902   -5.528  3.25e-08 ***
stageV      -1.0222    0.8644   -1.183  0.236980
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 162.832 on 14 degrees of freedom
Residual deviance: 9.625 on 8 degrees of freedom
AIC: 55.382

Number of Fisher Scoring iterations: 4

Q1: What is the **odds ratio** of 2 yr disease-free survival for a child **aged 24+ months** versus **aged < 12 months**?

Age	Stage	\mathbf{x}_i^\top	$\log(\pi_i/(1 - \pi_i))$
0-11	—	$(1, 0, 0, x_{i3}, x_{i4}, x_{i5}, x_{i6})^\top$	$\beta_0 + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}$
24+	—	$(1, 0, 1, x_{i3}, x_{i4}, x_{i5}, x_{i6})^\top$	$\beta_0 + \beta_2 + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}$

- The odds ratio is therefore $\psi = \exp\{\beta_2\}$, its MLE is:

$$\hat{\psi} = \exp\{\hat{\beta}_2\} = \exp\{-2.614\} = 0.0733.$$

- The **95 % CI** for this odds ratio is:

$$\exp\{\hat{\beta}_2 \pm 1.96 \text{se}(\hat{\beta}_2)\} = \exp\{-2.613 \pm 1.96 \times 0.5017\} = (0.0274, 0.1960).$$

- When controlling for stage at the diagnosis, the odds of 2-yr DFS for children aged 24+ months is only about 7 % [95 % CI: (0.0274, 0.1960)] of that for those aged less than 12 months.

Q2: What is the **odds ratio** of 2 yr disease-free survival for a child with **stage V** versus **stage II** cancer?

Age	Stage	\mathbf{x}_i^\top	$\log(\pi_i/(1 - \pi_i))$
—	V	$(1, x_{i1}, x_{i2}, 0, 0, 0, 1)^\top$	$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_6$
—	II	$(1, x_{i1}, x_{i2}, 1, 0, 0, 0)^\top$	$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3$

- The odds ratio is therefore $\psi = \exp\{\beta_6 - \beta_3\}$, its MLE is:

$$\hat{\psi} = \exp\{\hat{\beta}_6 - \hat{\beta}_3\} = \exp\{-1.022 + 1.253\} = 1.26.$$

- When controlling for age at the diagnosis, the odds of a 2-yr DFS for those diagnosed in stage V is 1.26 times of that for those diagnosed in stage II.

Q3: What is the **95 % CI** for OR $\psi = \exp\{\beta_6 - \beta_3\}$?

- Finding the 95 % CI for $\eta = \beta_6 - \beta_3 = C\beta$, where

$$C = \begin{bmatrix} 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix}.$$

Standard error for $\hat{\eta} = \hat{\beta}_6 - \hat{\beta}_3 = C^\top \hat{\beta}$:

$$\widehat{\text{Var}}(\hat{\beta}) = \mathbf{I}^{-1}(\hat{\beta})$$

$$\text{se}(C\beta) = \sqrt{C\mathbf{I}^{-1}(\hat{\beta})C^\top}.$$

```
C = c(0, 0, 0, -1, 0, 0, 1)
se = sqrt(C %*% vcov(model1) %*% C)
se
      [,1]
[1,] 0.6729361
```

The 95 % CI for $\eta = \beta_6 - \beta_3$ is:

$$\hat{\eta} \pm 1.96 \text{se}(\hat{\eta}) = (-1.0222 + 1.2529) \pm 1.96 \times 0.6729 = (-1.0882, 1.5496).$$

2. Exponentiate it to obtain the 95 % CI for $\psi = \exp\{\eta\} = \exp\{\beta_6 - \beta_3\}$:

$$\exp\{\hat{\eta} \pm 1.96 \text{se}(\hat{\eta})\} = (0.3368, 4.7098).$$

Topic 3e: Dose-Response Models

Bioassay Experiments

- **Bioassay experiment:** Several groups of subjects are exposed to varying levels of a drug/toxin to determine how many responses within a fixed period of time.
- **Stimulus:** Each group is subjected to a particular dose of the drug/toxin:

$$\text{dose} = \log(\text{concentration})$$

- **Response:** As a result of the stimulus, subjects will often manifest a binary response indicating the occurrence of an adverse event (e.g., death).
- **Tolerance:** We assume that for each subject there is a certain dose level above which the response will always occur.
 - This level is called the tolerance or threshold.
 - The tolerance varies from one individual to another in the population and therefore from subject to subject in the sample.
 - We can therefore ascribe a distribution to it.

The Tolerance Distribution

- z = concentration of the stimulus (toxin/drug).
- $x = \log(z)$ = dose/intensity of the stimulus.
- $f(x)$ = pdf for the distribution of the tolerance in the population (*i.e., the distribution for the stimulus/dose at which response occurs*).
- Suppose a dose of x_0 were applied to the population. What proportion would respond?

$$\pi_0 = \int_{-\infty}^{x_0} f(s) \, ds = F(x_0)$$

- If $x_0 < x_1$, then $\pi_0 < \pi_1$.

Modelling the Dose-Response Relationship

For each group $i = 1, \dots, n$ let:

- x_i = dose applied to subjects in group i ,
- m_i = number of subjects in group i ,
- y_i = the number of subjects with response in group i .

Dose	Responders	Total	
x_i	y_i	m_i	y_i/m_i
1.6907	6	59	0.10
1.7242	13	60	0.22
1.7552	18	62	0.29
\vdots	\vdots	\vdots	\vdots

- Assume

$$Y_i \sim \text{BIN}(m_i, \pi_i), \quad i = 1, \dots, n,$$

π_i = probability of response in group i with dose x_i .

- **Objective:** To model probability of response π_i as a function of dose x_i .
- Binomial Regression Models:

$$g(\pi_i) = \eta_i = \beta_0 + \beta_1 x_i,$$

where $g(\cdot)$ is a choice of link function.

- Then we have:

$$\pi_i = g^{-1}(\beta_0 + \beta_1 x_i),$$

that is, the probability of response as a function of dose x_i via $g^{-1}(\cdot)$.

- Question: What link function should we select?
- Realize that:

- If we assume a tolerance distribution $f(x)$, the probability of response to dose x_i is:

$$\pi_i = \int_{-\infty}^{x_i} f(x) dx = F(x_i).$$

- With a Binomial regression model and a link function $g(\cdot)$, we have:

$$\pi_i = g^{-1}(\beta_0 + \beta_1 x_i).$$

- These suggest that the choice of the tolerance distribution determines the form of the link function, i.e., selecting $g(\cdot)$ such that $g^{-1}(\cdot)$ is a cdf:

$$\pi_i = g^{-1}(\beta_0 + \beta_1 x_i) = F^*(\beta_0 + \beta_1 x_i).$$

Some Choices for the Tolerance Distribution

① Normal Tolerance Distribution:

$$\begin{aligned}\pi(x) &= \int_{-\infty}^x f(s) \, ds \\ &= \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2\right\} \, ds \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right)\end{aligned}$$

where Φ is the $\mathcal{N}(0, 1)$ cdf. This implies that

$$\Phi^{-1}(\pi) = \frac{x - \mu}{\sigma},$$

i.e., the [Probit link](#) s.t.,

$$g(\pi) = \Phi^{-1}(\pi) = -\frac{\mu}{\sigma} + \frac{1}{\sigma}x = \beta_0 + \beta_1 x.$$

A Binomial Probit Model:

$$\Phi^{-1}(\pi) = \beta_0 + \beta_1 x.$$

How do we interpret β_0 and β_1 ?

- They are no longer log odds ratios (as with logistic link)
- Interpretation is in terms of μ and σ the parameters of the Normal distribution for tolerance, i.e.,

$$\beta_0 = -\frac{\mu}{\sigma}, \quad \beta_1 = \frac{1}{\sigma}.$$

② Logistic Distribution:

$$f(x; \mu, s) = \frac{\exp\left\{-\frac{x-\mu}{s}\right\}}{s \left[1 + \exp\left\{-\frac{x-\mu}{s}\right\}\right]^2}, \quad s > 0, \mathbb{E}[X] = \mu.$$

The probability of response:

$$\begin{aligned}\pi(x) &= \int_{-\infty}^x f(x; \mu, s) \, ds = \left[1 + \exp\left\{-\frac{x-\mu}{s}\right\}\right]^{-1} \\ 1 - \pi(x) &= \frac{\exp\left\{-\frac{x-\mu}{s}\right\}}{1 + \exp\left\{-\frac{x-\mu}{s}\right\}} \\ \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) &= \frac{x - \mu}{s}.\end{aligned}$$

This implies the [Logit link](#) s.t.,

$$g(\pi) = \text{logit}(\pi) = -\frac{\mu}{s} + \frac{1}{s}x = \beta_0 + \beta_1 x.$$

③ Extreme Value Distribution:

$$f(x; \mu, s) = \frac{1}{s} \exp\left\{\frac{x-\mu}{s} - \exp\left\{\frac{x-\mu}{s}\right\}\right\}, \quad s > 0.$$

The probability of response:

$$\begin{aligned}\pi(x) &= \int_{-\infty}^x f(x; \mu, s) \, ds \\ &= 1 - \exp\left\{-\exp\left\{-\frac{x - \mu}{s}\right\}\right\} \\ \log(-\log(1 - \pi(x))) &= \frac{x - \mu}{s}.\end{aligned}$$

This implies the [Complementary log-log link](#) s.t.,

$$g(\pi) = \log(-\log(1 - \pi)) = -\frac{\mu}{s} + \frac{1}{s}x = \beta_0 + \beta_1 x.$$

Tolerance Distribution	Link Function	Dose-Response Model
Normal	Probit	$\Phi^{-1}(\pi) = \beta_0 + \beta_1 x$
Logistic	Logit	$\text{logit}(\pi) = \beta_0 + \beta_1 x$
Extreme Value	Complementary log-log	$\log(-\log(1 - \pi)) = \beta_0 + \beta_1 x$

Median Lethal/Effective Dose

- The [median lethal/effective dose](#) (ED50) is the dose at which 50 % of the population has the response.
- That is, if we let δ be the ED50, then by definition:

$$\pi(\delta) = \int_{-\infty}^{\delta} f(x) \, dx = 0.50.$$

- How do we find the expression of δ given a Dose-Response model? Suppose we fit a Binomial Probit model (i.e., Normal tolerance distribution):

$$\Phi^{-1}(\pi) = \beta_0 + \beta_1 x.$$

Note that at dose δ (ED50), $\pi = 0.50$.

$$\begin{aligned}\Phi^{-1}(0.50) &= \beta_0 + \beta_1 \delta \\ 0 &= \beta_0 + \beta_1 \delta \\ \delta &= -\frac{\beta_0}{\beta_1}\end{aligned}$$

A Dose-Response Example — Beetle Mortality

Beetle Mortality

Consider an experiment by Bliss (Annals of Applied Biology, 1935) in which groups of beetles were exposed to varying concentrations of carbon disulphide (CS₂) gas.

Dose (x_i)	# of insects killed (x_i)	# of insects m_i	y_i/m_i
1.6907	6	59	0.10
1.7242	13	60	0.22
1.7552	18	62	0.29
1.7842	28	56	0.50
1.8113	52	63	0.83
1.8369	53	59	0.89
1.8610	61	62	0.98
1.8839	60	60	1.00

- **Objective:** modelling the dose-response relationship.
- We will fit several binomial regression models to this data:

$$g(\pi_i) = \beta_0 + \beta_1 x_i,$$

where x_i = dose in group i , $i = 1, \dots, 8$.

- Various link functions will be used to find the best fitted model:
 - Logistic link.
 - Probit link.
 - Cloglog link.

Dose-Response Analysis using R

```
# read beetle data
beetle.dat = read.table("beetle.dat", header = T)
# here we construct the response variable for Binomial
# regression
beetle.dat$resp <- cbind(beetle.dat$y, beetle.dat$m - beetle.dat$y)
beetle.dat
```

	dose	y	m	resp.1	resp.2
1	1.6907	6	59	6	53
2	1.7242	13	60	13	47
3	1.7552	18	62	18	44
4	1.7842	28	56	28	28
5	1.8113	52	63	52	11
6	1.8369	53	59	53	6
7	1.8610	61	62	61	1
8	1.8839	60	60	60	0

Fit of the Logistic Model

```
model1 = glm(resp ~ dose, family = binomial(link = logit), data = beetle.dat)
summary(model1)
```

```

Call:
glm(formula = resp ~ dose, family = binomial(link = logit), data = beetle.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5941  -0.3944   0.8329   1.2592   1.5940

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717     5.181  -11.72  <2e-16 ***
dose          34.270     2.912   11.77  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4

```

Fit of the Probit Model

```

model2 = glm(resp ~ dose, family = binomial(link = probit), data = beetle.dat)
summary(model2)

Call:
glm(formula = resp ~ dose, family = binomial(link = probit),
    data = beetle.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5714  -0.4703   0.7501   1.0632   1.3449

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -34.935     2.648  -13.19  <2e-16 ***
dose          19.728     1.487   13.27  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.20  on 7  degrees of freedom
Residual deviance:  10.12  on 6  degrees of freedom
AIC: 40.318

Number of Fisher Scoring iterations: 4

```

Fit of the Complementary Log-log Model

```
model3 = glm(resp ~ dose, family = binomial(link = cloglog), data = beetle.dat)
summary(model3)
```

Call:

```
glm(formula = resp ~ dose, family = binomial(link = cloglog),
    data = beetle.dat)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.80329	-0.55135	0.03089	0.38315	1.28883

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-39.572	3.240	-12.21	<2e-16 ***
dose	22.041	1.799	12.25	<2e-16 ***

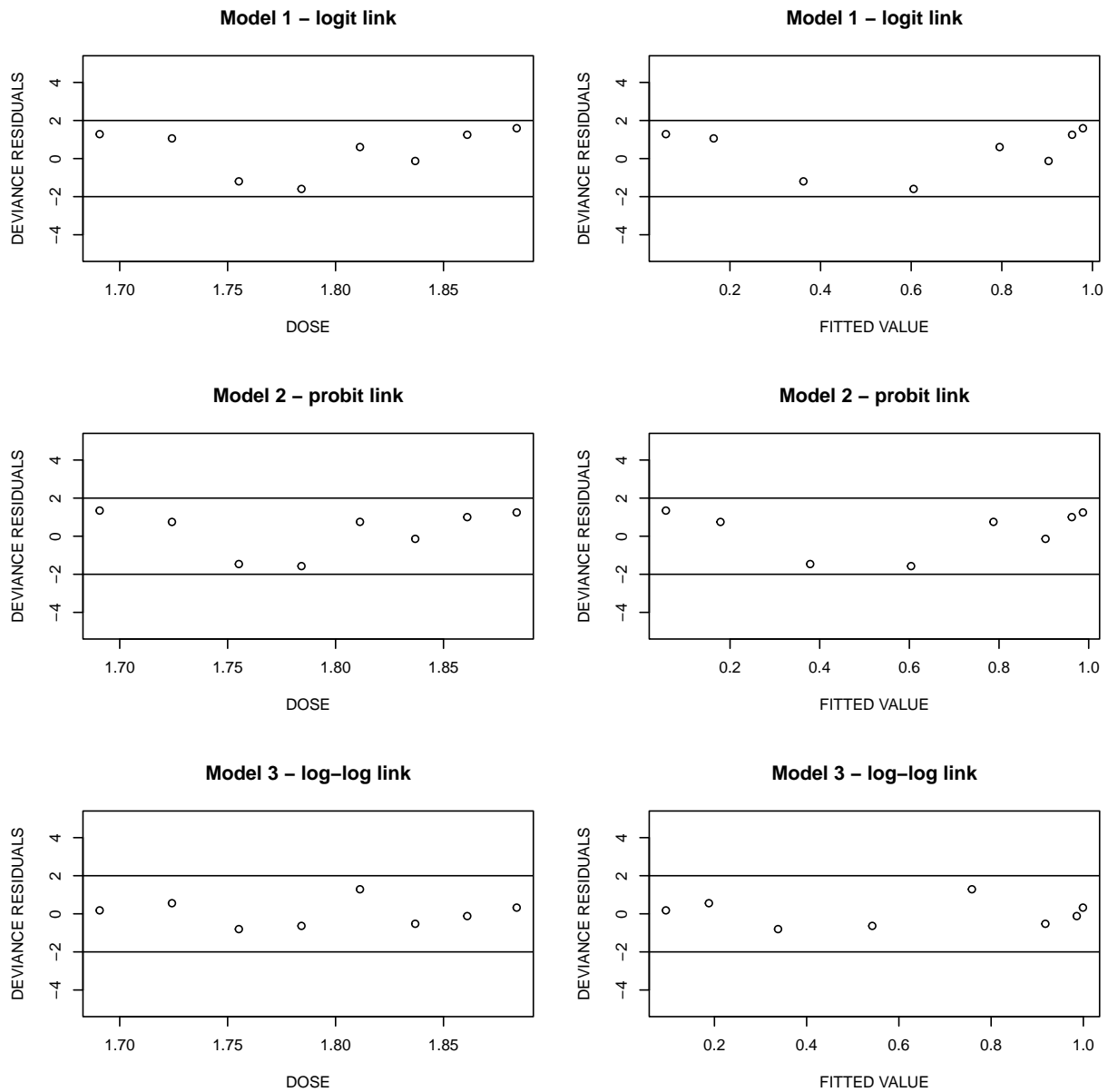
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.2024 on 7 degrees of freedom
Residual deviance: 3.4464 on 6 degrees of freedom
AIC: 33.644

Number of Fisher Scoring iterations: 4

Deviance Residual Plots



Choice of Tolerance Distribution or Binomial Model

- Observed probability of response:

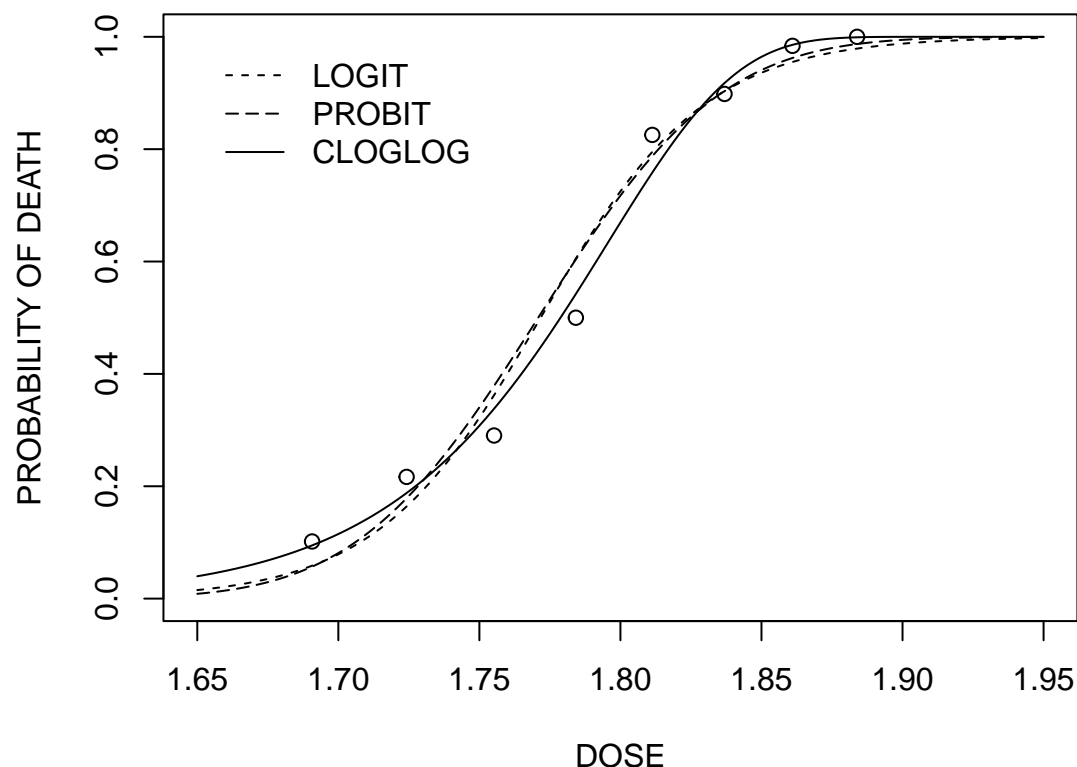
$$\tilde{\pi}_i = \frac{y_i}{m_i}.$$

- Fitted probability of response:

$$\hat{\pi}_i = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_i).$$

- The tolerance distribution (or the Binomial model) that provides the “best” agreement between the observed and fitted probability of response is the one that fits the data the “best.”
- We can check this by plotting the observed and fitted probability of response $\tilde{\pi}_i$ and $\hat{\pi}_i$, against dose x_i .

Fitted Dose-Response Curves



- Note that the curve for the complementary log-log link fits the data better than the other two, as expected from the residual plots and the deviance statistics.
- (The R code for generating above plot see course notes, 2.10.3, page 47).

Interpretation of Dose-Response Models

- Interpretation of regression parameter β_1 will depend on the link function.
 - Logistic model: $\text{logit}(\pi) = \beta_0 + \beta_1 x$.
 - * $\beta_1 = \log$ odds ratio for response associated with a one unit increase in dose.
 - Probit model: $\Phi^{-1}(\pi) = \beta_0 + \beta_1 x$, or Complementary log-log model $\log(-\log(1 - \pi)) = \beta_0 + \beta_1 x$, interpretation of β parameters is not as natural as in logistic models.
- Estimation of δ (ED50) from a Binomial model $g(\pi) = \beta_0 + \beta_1 x$:

$$g(\pi = 0.5) = \beta_0 + \beta_1 \delta \implies \hat{\delta} = \frac{g(0.5) - \hat{\beta}_0}{\hat{\beta}_1}.$$

- **Exercise:** What is $\delta_{0.25}$, the dose at which 25 % of the population has the response?

Topic 3f: Summary of Binomial Regression Models

Binomial GLM Specification

- $Y_i \sim \text{BIN}(m_i, \pi_i)$, $i = 1, \dots, n$ independently and

$$g(\pi_i) = \mathbf{x}_i^\top \boldsymbol{\beta},$$

where

- \mathbf{x}_i is a vector of explanatory variables,
 - π_i is the probability of event of interest,
 - $g(\cdot)$ is a link function that relates explanatory variables \mathbf{x}_i to probability π_i , and
 - $\boldsymbol{\beta}$ is a vector of regression parameters.
- When using the canonical link of Binomial distribution, i.e., $g(\cdot) = \text{logit}(\cdot)$, we have

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i^\top \boldsymbol{\beta},$$

which is called a *logistic regression model* which is commonly used in practice.

Parameters Estimation

- Likelihood methods are used for parameter estimating and inference.
- MLE $\hat{\boldsymbol{\beta}}$ come from Fisher Scoring using R function `glm()`.
- Interpretation: β_k has a *log OR interpretation* for logistic models.
- Variance covariance estimate for $\hat{\boldsymbol{\beta}} = \widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$, where \mathbf{I}^{-1} is the inverse of the information matrix evaluated at MLE $\hat{\boldsymbol{\beta}}$.

- Standard error: $\text{se}(\hat{\beta}_k) = \sqrt{[\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})]_{kk}} = \sqrt{I^{kk}(\hat{\boldsymbol{\beta}})}$.

- Wald-test of a single parameter: $H_0: \beta_k = \beta^*$ vs $H_A: \beta_k \neq \beta^*$:

$$\frac{(\hat{\beta}_k - \beta^*)^2}{I^{kk}(\hat{\boldsymbol{\beta}})} \stackrel{H_0}{\sim} \chi_1^2,$$

or

$$\frac{\hat{\beta}_k - \beta^*}{\text{se}(\hat{\beta}_k)} \sim \mathcal{N}(0, 1) \text{ under } H_0.$$

For testing $H_0: \beta_k = 0$, we have $\frac{\hat{\beta}_k}{\text{se}(\hat{\beta}_k)}$, reported as “z-value” in `glm()` summary.

- Confidence interval for a single β_k :

$$\hat{\beta}_k \pm Z_{1-\alpha/2} \text{se}(\hat{\beta}_k).$$

- Confidence interval for $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta}$:

$$\mathbf{x}_i^\top \hat{\boldsymbol{\beta}} \pm Z_{1-\alpha/2} \sqrt{\mathbf{x}_i^\top \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}) \mathbf{x}_i},$$

or equivalently

$$\mathbf{x}_i^\top \hat{\boldsymbol{\beta}} \pm Z_{1-\alpha/2} \sqrt{\mathbf{x}_i^\top \widehat{\text{Var}}(\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}) \mathbf{x}_i}.$$

How about CI for $\pi_i = \text{expit}(\mathbf{x}_i^\top \boldsymbol{\beta})$?

Model Checking and Selection

- LR/Deviance: Recall $LR = -2 \log \left(\frac{L(\hat{\pi})}{L(\tilde{\pi})} \right) = -2(\ell(\hat{\pi}) - \ell(\tilde{\pi}))$.

$$\begin{aligned} D &= -2[\ell(\hat{\pi}) - \ell(\tilde{\pi})] \\ &= -2 \sum_{i=1}^n \left(y_i \log \left(\frac{y_i}{m_i \hat{\pi}_i} \right) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i(1 - \hat{\pi}_i)} \right) \right) \\ &= \sum_{i=1}^n d_i. \end{aligned}$$

- LR/Deviance test for *adequacy of a model* (H_0 : fitted model is as adequate as the saturated model):

$$D \sim \chi_{n-p}^2 \text{ under } H_0.$$

- LR/Difference in Deviance test for *comparing nested models* (H_0 : reduced/simpler model is as adequate as the fitted model):

$$\Delta D = D_0 - D_A \sim \chi_{p-q}^2 \text{ under } H_0.$$

- Deviance Residuals:

$$r_i^D = \text{sign}(y_i - m_i \hat{\pi}_i) \sqrt{|d_i|},$$

where r_i^D 's should behave like an iid sample from $\mathcal{N}(0, 1)$ for a well-fitted model.

- Residuals plots:
 - *deviance residual vs fitted value* (i.e., r_i^D vs $\hat{\pi}_i$),
 - *deviance residual vs covariate* (i.e., r_i^D vs x_i).
 - In both cases, we expect a pattern of random scatter around 0, within $(-2, 2)$.
- Residual plots can be used to evaluate the fit of a model or compare multiple models in general.
 - For example, non-nested models, using different link functions.

Binomial Model for Dose-Response Relationship

- Dose: $X = \log(\text{concentration})$.
- Tolerance distribution is $f(x)$ and probability of responding to dose x is:

$$\pi(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(x) dx = F(x).$$

- Binomial GLMs can be utilized to evaluate the dose-response relationship:

$$g(\pi) = \beta_0 + \beta_1 x$$

In every case,

Link	Tolerance Distribution
Logit	Logistic(μ, s)
Probit	Normal(μ, s)
Cloglog	Extreme Value(μ, s)

$$\beta_0 = -\frac{\mu}{s}, \quad \beta_1 = \frac{1}{s}.$$

- Estimation of the *median lethal/effective dose* ($\delta_{0.5}$):

$$g(0.5) = \hat{\beta}_0 + \hat{\beta}_1 \delta_{0.5} \implies \delta_{0.5} = \frac{g(0.5) - \hat{\beta}_0}{\hat{\beta}_1}.$$

- Calculation of dose related to q^{th} percentile of response.

Topic 4a: Poisson GLMs for Count Data

The Poisson Distribution

- Recall for $Y \sim \text{POI}(\mu)$,

$$f(y) = \frac{\mu^y e^{-\mu}}{y!} = \exp\{y \log(\mu) - \mu - \log(y!)\}, \quad \mu > 0, y = 0, 1, 2, \dots$$

Examples of count data:

- Health service, # of emergency visits, # of hospitalizations.
- Insurance, # of claims.
- Engineering/manufacturing, # of defects.
- The Poisson is a member of the *exponential family* with

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

- Mean and variance:

$$\begin{aligned} \mathbb{E}[Y] &= b'(\theta) = e^\theta = \mu, \\ \text{Var}(Y) &= b''(\theta)a(\phi) = e^\theta = \mu. \end{aligned}$$

Therefore, $\mathbb{E}[Y] = \text{Var}(Y)$.

- The *Canonical link*:

$$\theta = \eta \implies \log(\mu) = \eta = x^\top \beta,$$

the log link, $g(\mu) = \log(\mu)$, is the canonical link.

Poisson Log Linear Model and Likelihood Function

- Now, suppose we have a random sample of size n :

$$Y_i \sim \text{POI}(\mu_i), \quad i = 1, 2, \dots, n,$$

and association with each y_i there is a covariate vector $\mathbf{x}_i^\top = (1, x_{i1}, \dots, x_{ip-1})^\top$.

- The likelihood and log-likelihood are then

$$\begin{aligned} L(\boldsymbol{\mu}) &= \prod_{i=1}^n \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, \\ \ell(\boldsymbol{\mu}; \mathbf{y}) &= \sum_{i=1}^n (y_i \log(\mu_i) - \mu_i - \log(y_i!)). \end{aligned}$$

- Using the **Canonical link** (i.e., log link):

$$\log(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} = \sum_{j=0}^{p-1} x_{ij} \beta_j,$$

which is referred to as **log linear regression** because the use of the log link.

- We can obtain the log-likelihood in terms of $\boldsymbol{\beta}$ by substitution:

$$\begin{aligned} \ell(\boldsymbol{\mu}; \mathbf{y}) &= \sum_{i=1}^n (y_i \log(\mu_i) - \mu_i - \log(y_i!)) \\ &= \sum_{i=1}^n (y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\} - \log(y_i!)). \end{aligned}$$

Estimation of $\boldsymbol{\beta}$ from log linear regression

- The j^{th} contribution to the Score vector is:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^n (y_i x_{ij} - x_{ij} \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}).$$

- The (j, k) element of the Information Matrix is:

$$-\frac{\partial^2 \ell}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^n (x_{ij} x_{ik} \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}).$$

- These can also be found using general exponential family results.
- Use the above to estimate $\hat{\boldsymbol{\beta}}$ via Fisher Scoring.

Poisson Deviance/LR Tests

- Let $\tilde{\mu}_i$ be the MLE under the **saturated model** (i.e., $\tilde{\mu}_i = y_i$ which is the Poisson MLE for μ_i).
- Let $\hat{\mu}_i$ be the MLE under a p -dimensional **constrained model** (e.g., $\hat{\mu}_i \exp\{\mathbf{x}_i^\top \hat{\boldsymbol{\beta}}\}$).
- Recall the Likelihood Ratio or Deviance Statistic has the form:

$$D = -2 \log \left(\frac{L(\hat{\boldsymbol{\mu}})}{L(\tilde{\boldsymbol{\mu}})} \right) = 2(\ell(\tilde{\boldsymbol{\mu}}) - \ell(\hat{\boldsymbol{\mu}})).$$

- Under H_0 : constrained model is as adequate as saturated model, we have the following asymptotic distribution result:

$$D \sim \chi_{n-p}^2.$$

- For the Poisson we have:

$$\begin{aligned} D &= 2 \sum_{i=1}^n \left((y_i \log(\tilde{\mu}_i) - \tilde{\mu}_i) - (y_i \log(\hat{\mu}_i) - \hat{\mu}_i) \right) \\ &= 2 \sum_{i=1}^n \left(y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right) \\ &= 2 \sum_{i=1}^n \left(O_i \log \left(\frac{O_i}{E_i} \right) - (O_i - E_i) \right). \end{aligned}$$

- Question: does the Deviance Statistic have the form as the Binomial case, i.e.,

$$D = 2 \sum O_i \log\left(\frac{O_i}{E_i}\right) ?$$

When there is an intercept included in the Poisson log-linear model:

$$\frac{\partial \ell}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \left[\sum_{i=1}^n y_i \mathbf{x}_i^\top \boldsymbol{\beta} - \exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\} \right] = \sum_{i=1}^n (y_i - \mu_i) \implies \sum_{i=1}^n (y_i - \hat{\mu}_i) = 0,$$

then the Deviance takes the form

$$D = 2 \sum_{i=1}^n y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) = 2 \sum_{i=1}^n O_i \log\left(\frac{O_i}{E_i}\right).$$

- Use the Deviance to test nested models:
 - H_0 : the reduced model with p parameters is adequate versus

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1}$$

- H_A : the full model with q parameters ($p < q$)

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + \cdots + \beta_{q-1} x_{iq-1}.$$

- LR/Difference in Deviance test statistic:

$$\Delta D = D_0 - D_A \sim \chi_{q-p}^2 \text{ under } H_0.$$

- The p -value for this test is given by:

$$p\text{-value} = \mathbb{P}(\chi_{q-p}^2 > \Delta D).$$

Deviance Residuals

- We can write the Deviance as a sum:

$$D = 2 \sum_{i=1}^n \left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right) = \sum_{i=1}^n d_i.$$

- The **Deviance Residuals** are given by:

$$r_i^D = \text{sign}(y_i - \hat{\mu}_i) \sqrt{|d_i|},$$

and are approximately $\mathcal{N}(0, 1)$ if H_0 holds.

- We can use the residual plots to evaluate the fit of a model.

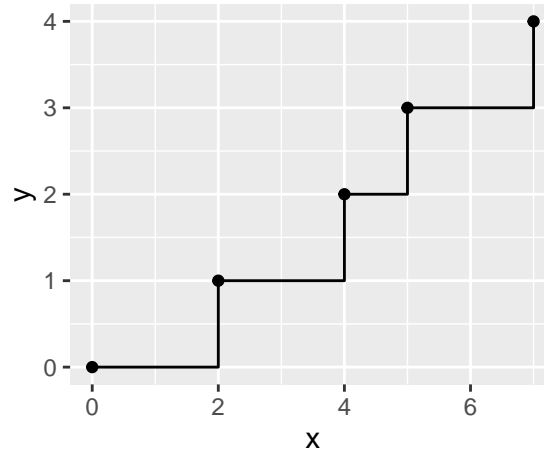
Regression for Poisson Processes

- The Poisson distribution assumes a **common observation period** for all individuals, so that the number of event does not depend on the time at risk.
- However, this may not be the case for many situations in practice.

Counting Process $N(t)$

A **counting process** $N(t)$ is any non-decreasing integer function of time such that $N(0) = 0$ and $N(t)$ is the number of events occurring in $(0, t]$.

- **Example:** Suppose events occurred at times $(2, 4, 5, 7)$.
- Draw a plot of $N(t)$ versus t :



Poisson Process $N(t)$

A counting process $N(t)$ is a **Poisson process** if it satisfies:

1. **Independent increments:** For $s_1 < t_1 < s_2 < t_2$:

$$N(t_1) - N(s_1) = \# \text{ events in } (s_1, t_1],$$

is independent of

$$N(t_2) - N(s_2) = \# \text{ events in } (s_2, t_2].$$

2. The number of events over $(0, t]$ has a Poisson distribution, i.e.,

$$\mathbb{P}(N(t) = n; \lambda) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, \quad \lambda > 0, n = 0, 1, 2, \dots$$

- Expected number of events in $(0, t]$ is

$$\mathbb{E}[N(t)] = \mu(t) = \lambda t.$$

Parameter λ is a constant representing the *rate of occurrence of the event per unit of time*:

$$\lambda = \text{rate parameter},$$

$$t = \text{length of observation period}.$$

- Since λ is constant (not a function of t) we call this a **time homogeneous Poisson process**.
- Use the log link to do regression:

$$\log(\mu(t)) = \log(\lambda t) = \log(\lambda) + \log(t).$$

For each subject $i = 1, \dots, n$ we observe:

- $N_i(t_i)$ = the number of events observed over $(0, t_i]$.
- Explanatory variables: $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip-1})^\top$.

Log Linear Regression Model for a Time Homogeneous Poisson Process

$$\begin{aligned}\log(\mu_i(t_i)) &= \log(\lambda_i) + \log(t_i) \\ &= \mathbf{x}_i^\top \boldsymbol{\beta} + \log(t_i)\end{aligned}\quad \text{e.g., } \log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

- The term $\log(t_i)$ is called an “*offset term*”.
- It accounts for different lengths of observation.
- It *explains* some variation in the event counts across subjects, but does so in a deterministic way.

Next week: an example of fitting Poisson GLM using R.