Generalized Linear Models and their Applications

STAT 431/STAT 831*

Fall 2021 (1219)†

Cameron Roopnarine‡ Leilei Zeng§

27th September 2021

*STAT 431 ≡ STAT 831

†Online Course

‡LaTeXer

§Instructor

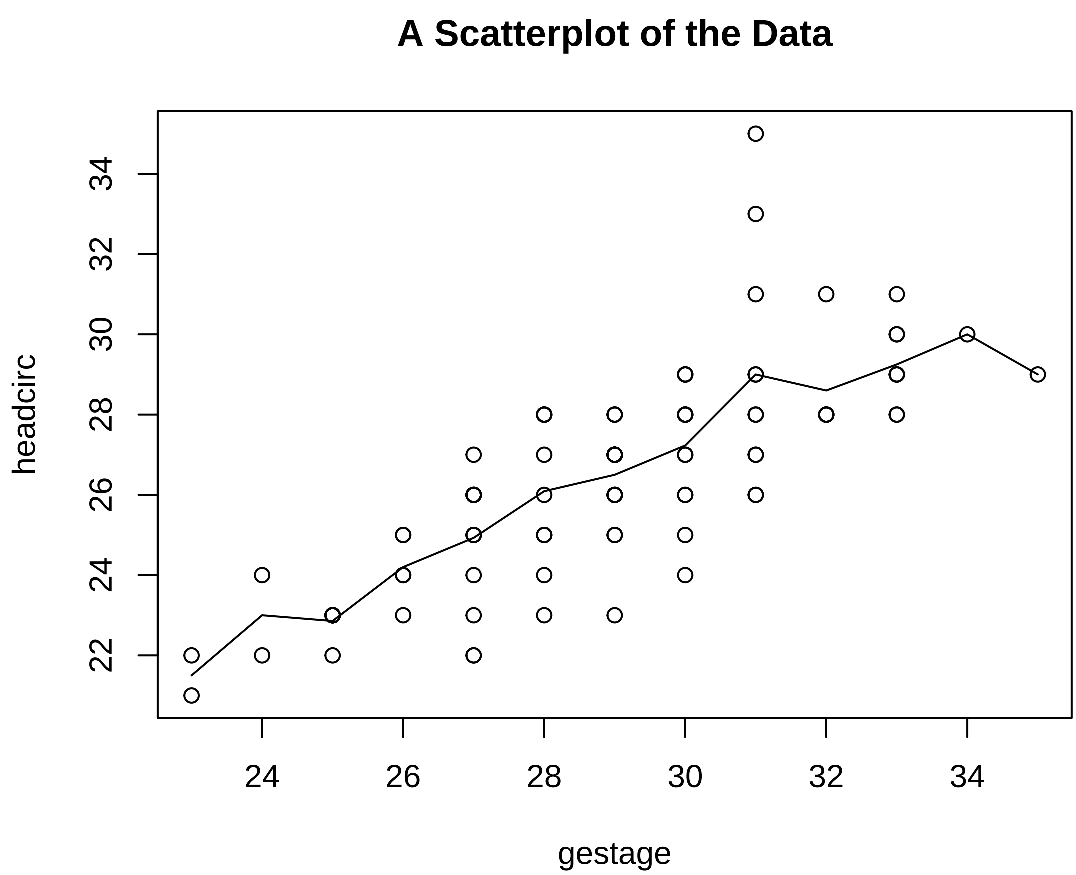
Contents

Topic 1a: Review of Linear Regression

Example: low birthweight infants study[1]

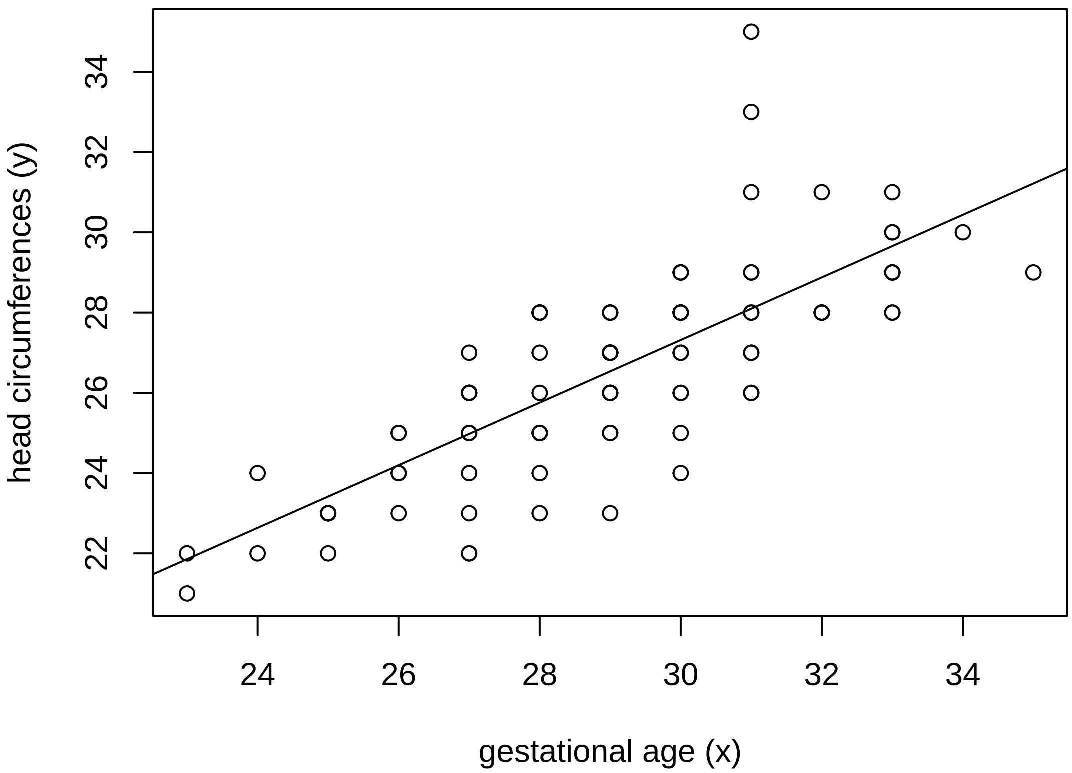
A study was conducted at two teaching hospitals in Boston, Massachusetts, where the head circumference, gestational age and some other variables are recorded for 100 low birth weight infants.

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



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

[1] Principles 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
- $\boldsymbol{x}_i = (1, x_{i1}, \dots, x_{ip})^\top$ a vector of explanatory variables.

Specification for Multiple Linear Regression

- Linear regression equation:

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

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

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

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

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

where

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

and

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

2 Estimation

Least Squares Method

We wish to minimize a loss function:

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

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

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

Maximum Likelihood Method

The probability density function for Y_i is:

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

The log-likelihood function is therefore:

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

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

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

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

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

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

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

Low Birthweight Infant Data Example

- For $n = 100$ infants, we have observed Y_i = head circumference and x_i = gestational age for baby i, $i = 1, \ldots, 100$.
- 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:

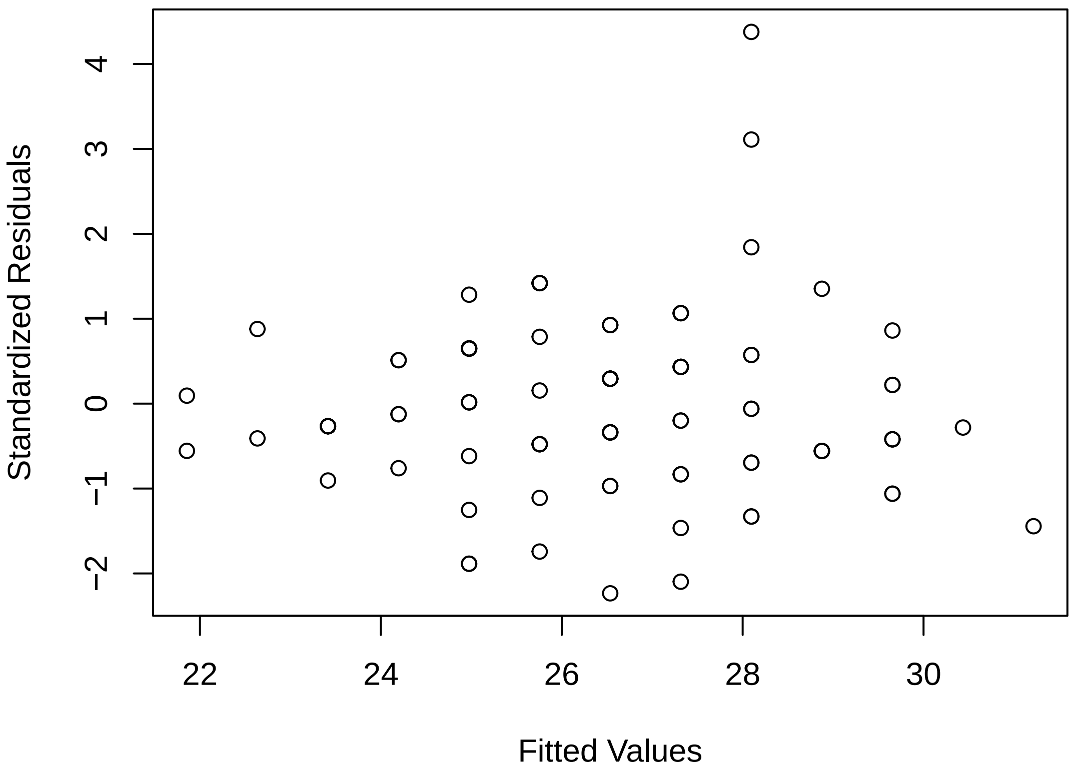
$$d_i = \frac{r_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}},$$

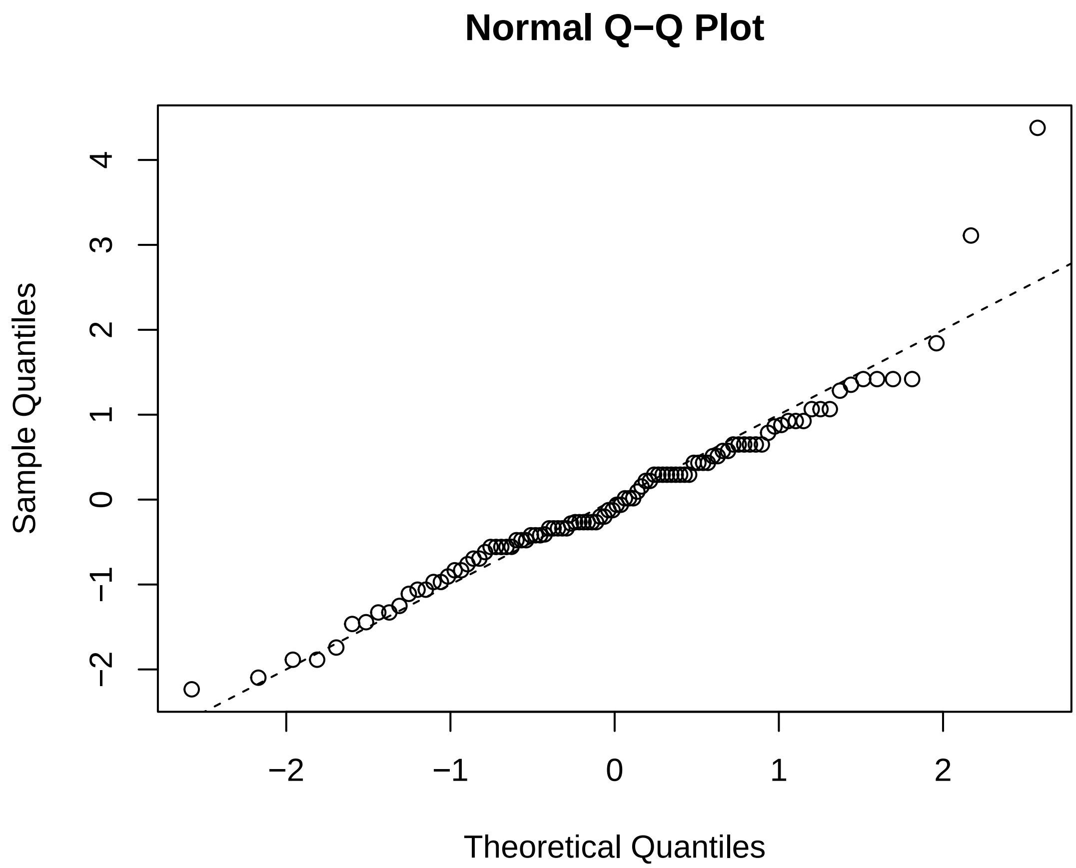
where h_{ii} is the (i, i) element of $\boldsymbol{H} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$. By asymptotic theory, if the model provides a good fit to the data then we should expect that:

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

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.


Standardized Residuals
−2
−1
0
1
2
3
4
22
24
26
28
30
Fitted Values



4 Inference

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

$$\hat{\boldsymbol{\beta}} \sim \mathrm{MVN}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}),$$
$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 v_{jj}), \qquad \text{where } v_{jj} = \left[(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right]_{(j,j)}.$$

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

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

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: $\beta_1 \neq 0$.

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

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?

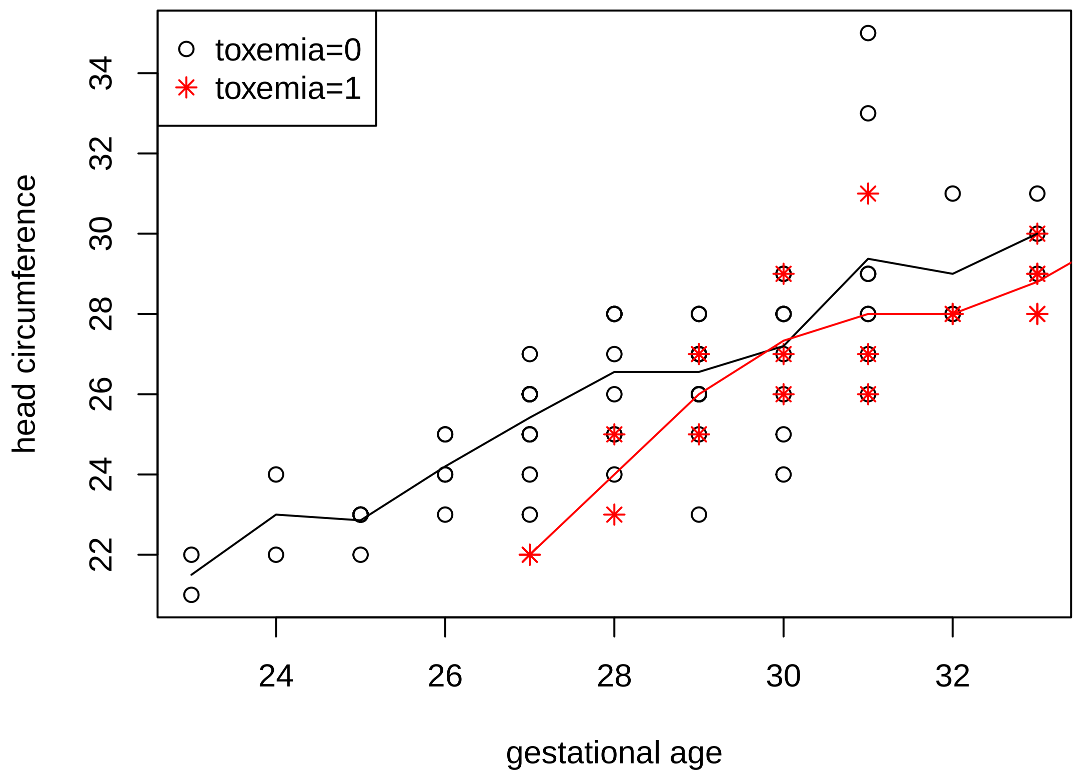
  A 95 % CI for β_1:

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

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?

```
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-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 ($Y = 0, 1, 2, 3, \ldots$).

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

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

- 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: $\ell(\theta) = \log\big(L(\theta)\big)$.
- 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: $\mathcal{I}(\theta) = \mathrm{E}\big[I(\theta)\big]$.
- Relative likelihood function: $R(\theta) = L(\theta)/L(\hat{\theta})$.
- Log relative likelihood function: $r(\theta) = \log\big(L(\theta)/L(\hat{\theta})\big) = \ell(\theta) - \ell(\hat{\theta})$.

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 followup.
- 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 $Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(\pi)$ where $\pi = \mathrm{P}(Y_i = 1)$, then the total number of woman developed breast cancer is:

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

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

- Likelihood function:
$$L(\pi; y) = c\,\mathrm{P}(Y = y; \pi) = \pi^y(1-\pi)^{n-y},$$
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:
$$I(\pi) = -\ell''(\pi) = \frac{y}{\pi^2} + \frac{n-y}{(1-\pi)^2} > 0\ \forall \pi \in (0,1).$$

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 followup.
- Assume $Y \sim \text{Binomial}(n, \pi)$ with unknown parameter π.
- The maximum likelihood estimate of π is:
$$\hat{\pi} = \bar{y} = \frac{y}{n} = \frac{166}{8506} = 0.0195.$$

Example: Poisson Distribution

Suppose $y_1, \ldots, y_n$ is an iid sample from a Poisson distribution with probability mass function:
$$f(y; \lambda) = \mathrm{P}(Y = y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!},\ \lambda > 0,\ y = 0, 1, 2, \ldots.$$

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

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

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

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:

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

- 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 $\theta^{(0)}$:

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

- 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: $-2\log\big(R(\theta)\big) = -2r(\theta) \sim \chi^2_{(1)}$.
- Score statistic: $\big(S(\theta)\big)^2/I(\theta) \sim \chi^2_{(1)}$.
- 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:

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

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:

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

or equivalently

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

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: $\big\{\theta : -2r(\theta) < \chi^2_1(0.95)\big\}$ where $r(\theta) = \ell(\theta) - \ell(\hat\theta)$.

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

$$r(\theta) = \big(y\log(\theta) + (n-y)\log(1-\theta)\big) - \left(y\log\Big(\frac{y}{n}\Big) + (n-y)\log\Big(1-\frac{y}{n}\Big)\right).$$

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

```
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 $(0.017, 0.023)$.

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

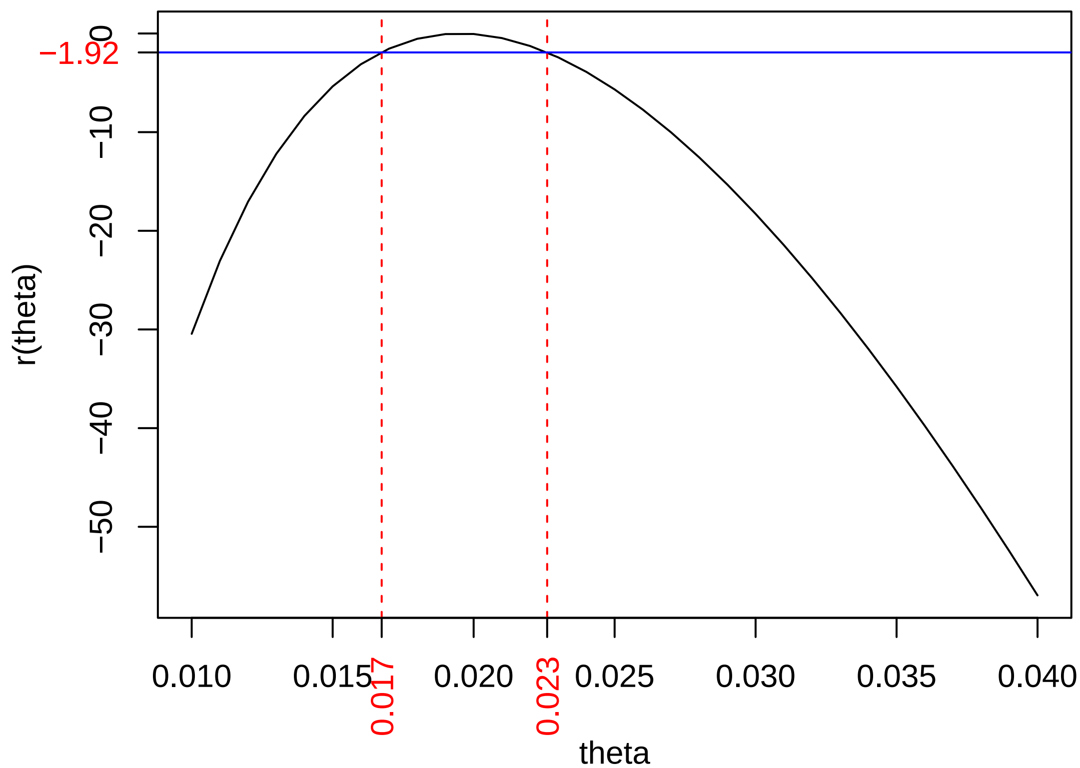
- For Binomial distribution $\hat{\theta} = y/n$ and

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

- So we solve:

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

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



Hypotheses Test

Suppose we are interested in testing hypotheses:

$$\mathrm{H}_0\colon\ \theta = \theta_0 \text{ vs } \mathrm{H}_\mathrm{A}\colon\ \theta \neq \theta_0.$$

- Likelihood ratio (LR) test: p-value $= \mathrm{P}\big(\chi^2_1 > -2r(\theta_0)\big)$.

- Score test: p-value $= \mathrm{P}\Big(\chi_1^2 > \big(S(\theta)\big)^2 / I(\theta_0)\Big)$.
- Wald test:

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

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 %.

$$\mathrm{H}_0\colon\ \theta = 0.015 \text{ vs } \mathrm{H}_\mathrm{A}\colon\ \theta \neq 0.015.$$

- Likelihood Ratio based test:

$$\begin{aligned} r(\theta_0 = 0.015) &= \Big(y\log(0.015) + (n-y)\log(1-0.15)\Big) - \left(y\log\Big(\frac{y}{n}\Big) + (n-y)\log\Big(1-\frac{y}{n}\Big)\right) \\ &= -5.3637. \end{aligned}$$

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

$$p = \mathrm{P}\Big(\chi^2_{(1)} > -2r(0.015)\Big) = \mathrm{P}\Big(\chi^2_{(1)} > 10.7274\Big) = 0.001.$$

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 from 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 $\boldsymbol{\theta} \in \Omega$ is a continuous $p \times 1$ parameter vector indexing a probability density (or mass) function $f(\boldsymbol{y};\boldsymbol{\theta})$. The likelihood and log-likelihood functions are defined as before, but

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

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

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

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

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

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

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

- Analogs to the LR, Score and Wald results apply based on partitioning the Information matrix by $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})^\top$, where $\boldsymbol{\alpha}$ is a $p \times 1$ vector of nuisance parameters and $\boldsymbol{\beta}$ is a $q \times 1$ vector of parameters of interest:

$$\boldsymbol{I} = \boldsymbol{I}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \begin{pmatrix} \boldsymbol{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) & \boldsymbol{I}_{\boldsymbol{\alpha}\boldsymbol{\beta}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) & \boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \end{pmatrix},$$

  where $\boldsymbol{I}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^\top}$ is $p \times p$, $\boldsymbol{I}_{\boldsymbol{\alpha}\boldsymbol{\beta}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\beta}^\top}$ is $p \times q$, $\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\alpha}^\top}$ is $q \times p$, and $\boldsymbol{I}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}$ is $q \times q$.

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

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

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

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

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

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

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

  Solution.

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

  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.

$$\begin{aligned}
L(\theta,\phi;y) &= f(y;\theta,\phi) = \exp\bigg\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\bigg\}.\\
\ell(\theta,\phi;y) &= \log\big(f(y;\theta,\phi)\big) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi).\\
S(\theta) &= \frac{\partial \ell}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}.\\
I(\theta) &= -\frac{\partial^2 \ell}{\partial \theta^2} = \frac{b''(\theta)}{a(\phi)}.\\
\mathcal{I}(\theta) &= \mathrm{E}\bigg[-\frac{\partial^2 \ell}{\partial \theta^2}\bigg] = I(\theta).
\end{aligned}$$

Some General Results for Score and Information

Result # 1

The expectation of the score function is zero.

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

Proof:

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

Result # 2

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

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

Proof: Differentiate (1) again,

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

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

$$\begin{aligned}
\mathrm{E}\big[S(\theta)\big] &= 0, \\
\mathrm{E}\left[\frac{Y - b'(\theta)}{a(\phi)}\right] &= 0, \\
\mathrm{E}[Y] &= b'(\theta),
\end{aligned}$$

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

Mean and Variance for the Exponential Family

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

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 $b''(\theta) = \mathrm{V}(\mu)$ and call $\mathrm{V}(\mu)$ the variance function.
- Subsequently, we have:
$$\mathrm{Var}(Y) = b''(\theta)a(\phi) = \mathrm{V}(\mu)a(\phi),$$
which is the mean-variance relationship for the exponential family.

Link Functions

Definition (Link Function)

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

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

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:

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

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

Examples

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

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

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 $Y_1, \ldots, Y_n$ 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):

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

  a linear combination of explanatory variables $\boldsymbol{x}_i$ and regression parameters $\boldsymbol{\beta}$.
- Link function:

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

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

Topic Summary

1. 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.
2. 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:

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

2 Systematic Component (or linear predictor):

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

- $\boldsymbol{x}_i = (1, x_{i1}, \dots, x_{ip})^\top$ 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 $\mathrm{E}[Y_i] = \mu_i$ to a linear prediction η_i:

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

Example: A Poisson Regression Model

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

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

Poisson distribution is a member of exponential family with:

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

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

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

Example: A Normal Regression Model

Assume $Y_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$ 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: $\theta_i = \eta_i \Longrightarrow \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ (identity link).

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

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

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

Score Vector

To find the maximum likelihood estimator for $\boldsymbol{\beta}$, we must solve $\boldsymbol{S}(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{0}$. 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

$$\begin{aligned}
\frac{\partial \ell}{\partial \theta} &= \frac{y - b'(\theta)}{a(\phi)}, \\
\frac{\partial \theta}{\partial \mu} &= \left(\frac{\partial \mu}{\partial \theta}\right)^{-1} = \frac{1}{b''(\theta)} && \text{since } \mu = b'(\theta), \\
\frac{\partial \mu}{\partial \eta} &= \frac{\partial \mu}{\partial \eta}, \\
\frac{\partial \eta}{\partial \beta_j} &= x_j && \text{since } \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \cdots + \beta_p x_p.
\end{aligned}$$

Hence, we have:

$$
\begin{aligned}
\frac{\partial \ell}{\partial \beta_j} &= \frac{y - b'(\theta)}{a(\phi)} \frac{1}{b''(\theta)} \frac{\partial \mu}{\partial \eta} x_j \\
&= \frac{y-\mu}{\operatorname{Var}(Y)} \frac{\partial \mu}{\partial \eta} x_j && \text{since } \mu = b'(\theta),\ \operatorname{Var}(Y) = a(\phi) b''(\theta) \\
&= \frac{y-\mu}{\operatorname{Var}(Y)} \left(\frac{\partial \mu}{\partial \eta}\right)^2 \frac{\partial \eta}{\partial \mu} x_j && \text{since } \frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \mu} = 1 \\
&= (y-\mu) \left(\operatorname{Var}(Y) \left(\frac{\partial \mu}{\partial \eta}\right)^2 \right)^{-1} \frac{\partial \eta}{\partial \mu} x_j \\
&= (y-\mu) W \frac{\partial \eta}{\partial \mu} x_j,
\end{aligned}
$$

where $W^{-1} = \operatorname{Var}(Y) \left(\frac{\partial \eta}{\partial \mu}\right)^2$. 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:

$$
\begin{aligned}
L &= \prod_{i=1}^n f(y_i; \theta_i, \phi) = \prod_{i=1}^n \exp\left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}, \\
\ell &= \sum_{i=1}^n \ell_i = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi).
\end{aligned}
$$

The element of the score vector is:

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

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

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

where

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

Example: Poisson Regression Model (Problem 1.4)

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

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

Poisson regression with a log-link:

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

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

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

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

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

1. Newton Raphson update equation is:

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

   where $\boldsymbol{I}$ is the observed information matrix.

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

2. Fisher Scoring update equation is:

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

Information Matrix

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

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

Fisher Information

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

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

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

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

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

When is Fisher Scoring Equivalent to Newton Raphson?

Recall information matrix:

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

Now examine:

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

So under the canonical link,

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

therefore information matrix is same as the Fisher information:

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

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:

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

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

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

- See manipulation in Section 1.2.3 of course notes.
- Same form as the weighted LS estimate of $\boldsymbol{\beta}$ with dependent variable $\boldsymbol{Z}$ and weight matrix $\boldsymbol{\mathcal{W}}$.
- $\boldsymbol{Z}$ and $\boldsymbol{\mathcal{W}}$ 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:

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

- 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 $\boldsymbol{\beta}$ from any GLM based on general forms for $\boldsymbol{I}(\boldsymbol{\beta})$ and $\boldsymbol{S}(\boldsymbol{\beta})$.
- 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:

	COVID		
Vaccination	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

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

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

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:

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

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:

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

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

Our re-parameterization implies:

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

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

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

So by the invariance property of MLEs, we have:

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

Inference for Odds Ratio

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

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

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

$$(\hat{\theta}_1 - \theta_1)^2\left(I^{11}(\hat{\theta}_1,\hat{\theta}_2)\right)^{-1} \sim \chi^2_{(1)},$$

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

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

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

- 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:

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

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

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

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* Died	Survived	Total
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:

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

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

- $\hat{\psi}_{\mathrm{A}} = 0.80\ (0.37, 1.73)$ and $\hat{\psi}_{\mathrm{B}} = 1.01\ (0.33, 3.10)$. 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 Morality disappeared when we stratified by clinic.

$$\begin{tikzcd} & \text{Clinic} \arrow[dl, <->] \arrow[dr, <->] & \\ \text{Level of Care} \arrow[rr, "?"] & & \text{Mortality} \end{tikzcd}$$

- 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.