Generalized Linear Models and their Applications

STAT 431/STAT 831[*]

Fall 2021 (1219)[†]

Cameron Roopnarine[‡] Leilei Zeng[§]

12th November 2021

[*] STAT 431 ≡ STAT 831

[†] Online Course

[‡] LaTeXer

[§] Instructor

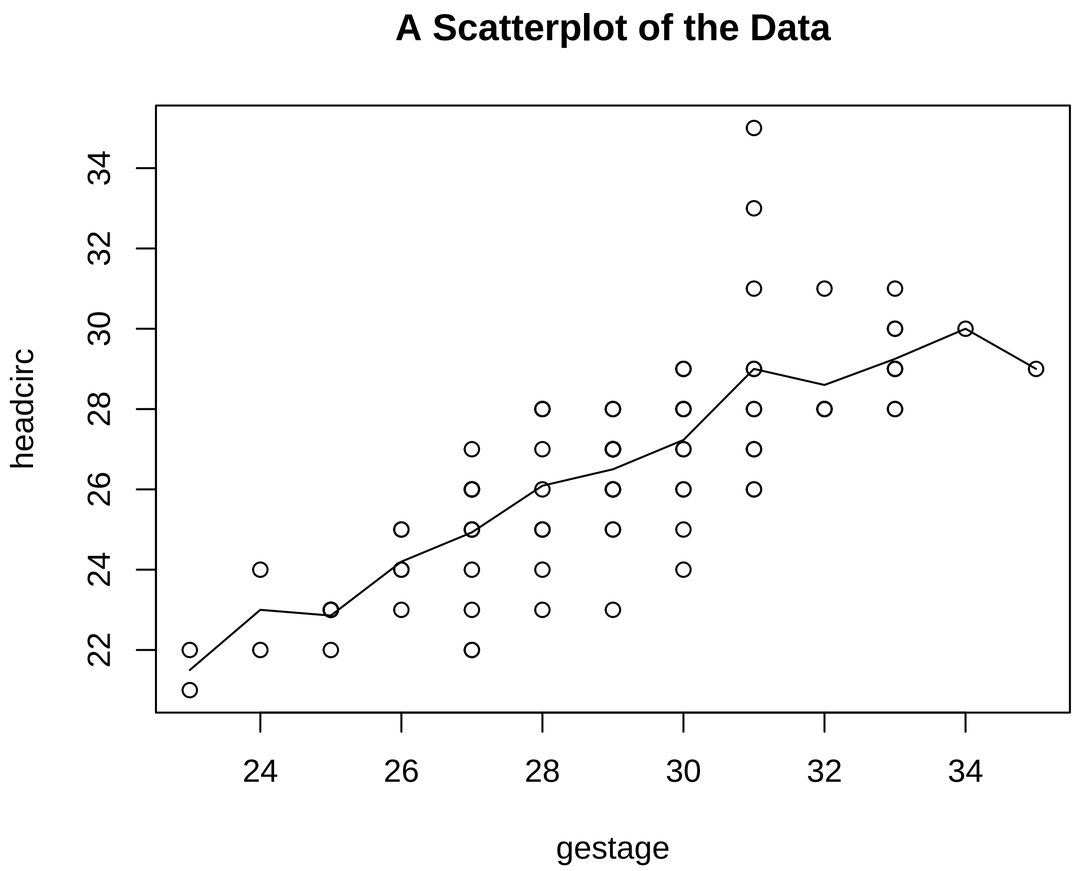
Contents

Topic 1a: Review of Linear Regression

Example: low birthweight infants study[1]

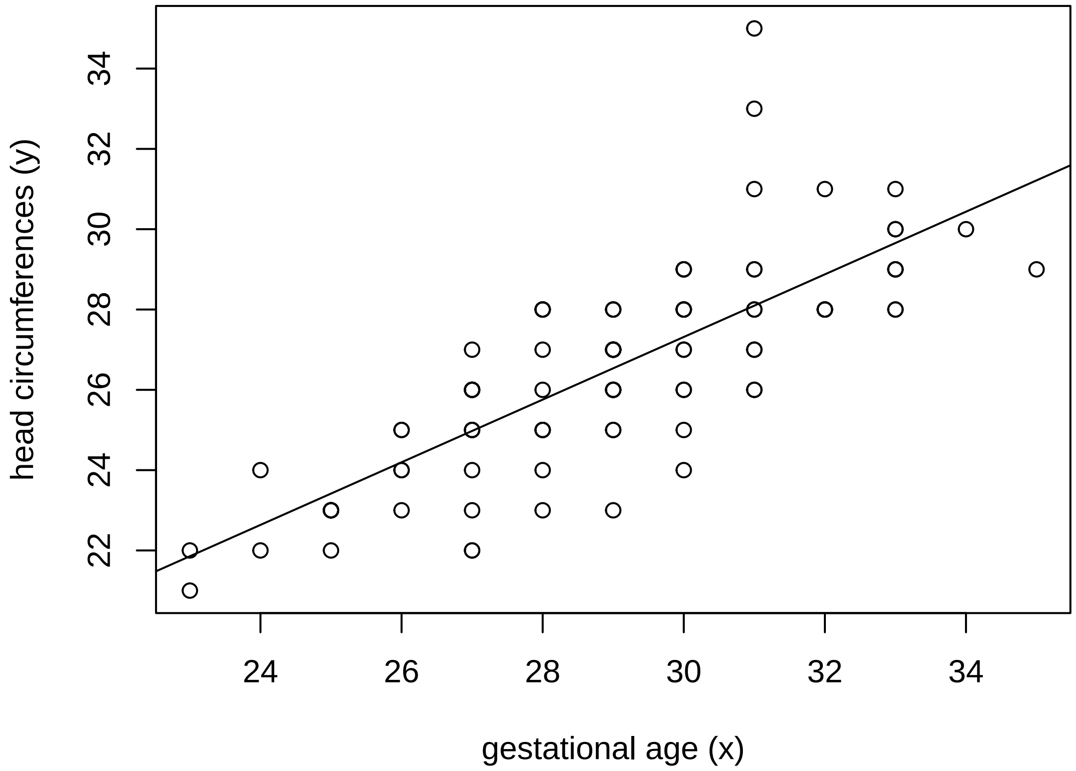
A study was conducted at two teaching hospitals in Boston, Massachusetts, where the head circumference, gestational age and some other variables are recorded for 100 low birth weight infants.

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 $i = 1, \ldots, n$ we have:

- Y_i = random variable representing the response, and
- $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$, 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

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

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

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

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

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

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:

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

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:

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

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:

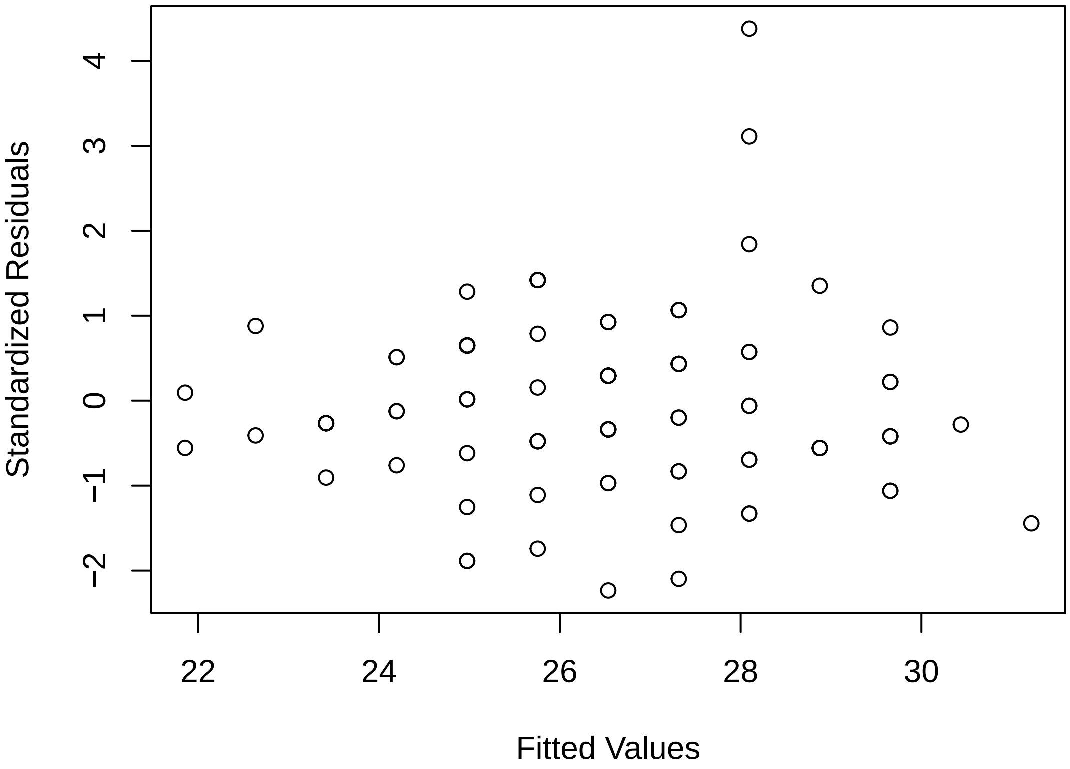
$$d_i = \frac{r_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}},$$

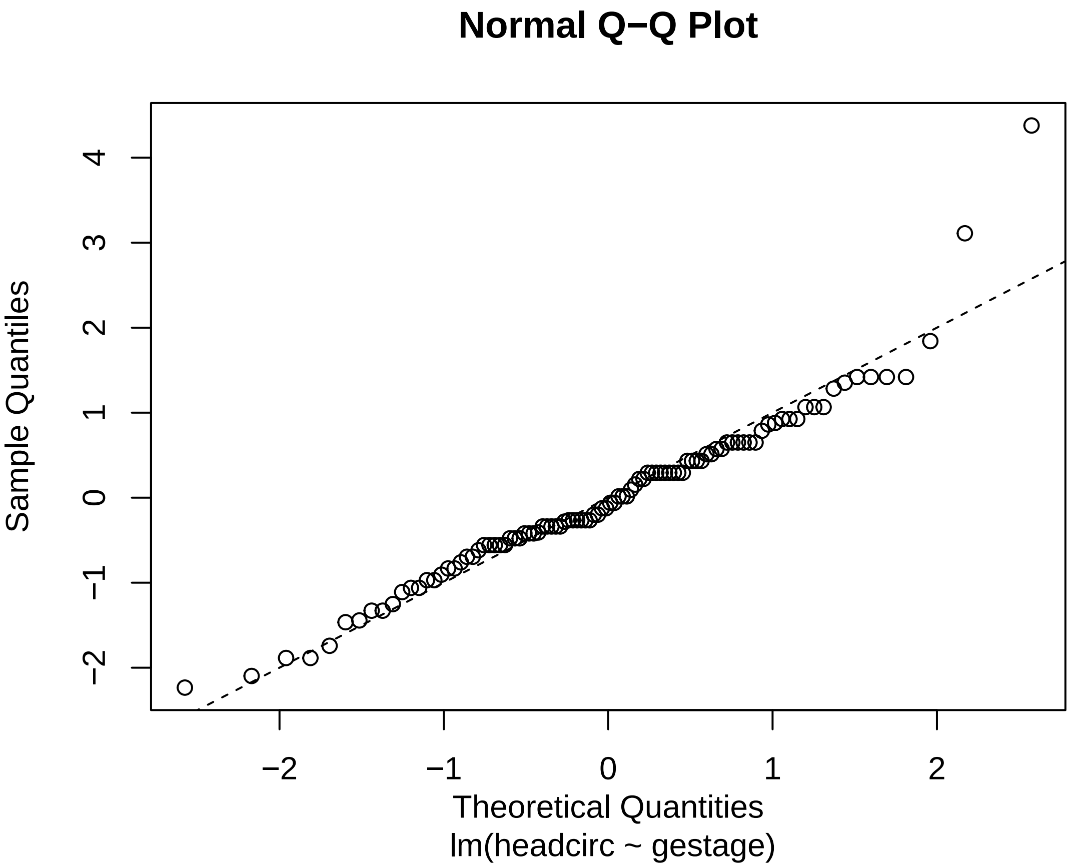
where h_{ii} is the (i, i) element of $\boldsymbol{H} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$. By asymptotic theory, if the model provides a good fit to the data then we should expect that:

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

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:

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

- 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 *H₀*: $\beta_1 = 0$ 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?

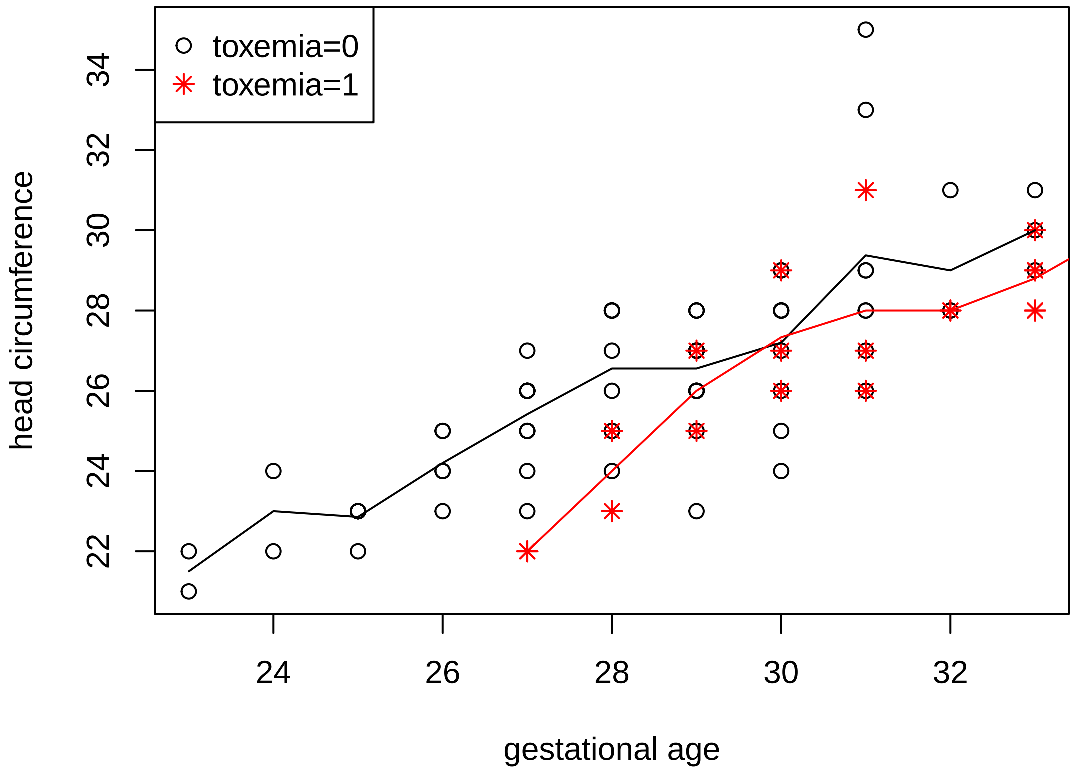
  A $95\,\%$ CI for β_1:

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

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 $1.41\,\mathrm{cm}$) 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?

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

```r
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.,
$$L(\theta; y) = cf(y; \theta) = c\,\mathbb{P}(Y = y; \theta),$$
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
$$\ell(\theta; y) = \log\big(L(\theta; y)\big).$$
- 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: $\mathcal{I}(\theta) = \mathbb{E}\big[I(\theta)\big]$.
- Relative likelihood function: $R(\theta) = L(\theta)/L(\hat{\theta})$, $0 \le R(\theta) \le 1$.
- Log relative likelihood function: $r(\theta) = \log\big(L(\theta)/L(\hat{\theta})\big) = \ell(\theta) - \ell(\hat{\theta})$, $r(\theta) \le 0$.

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 $\mathrm{e}^{\hat{\theta}}$ 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 $Y_i \overset{\text{iid}}{\sim} \text{BERN}(\pi)$ 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:

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

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

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

- Likelihood function:

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

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

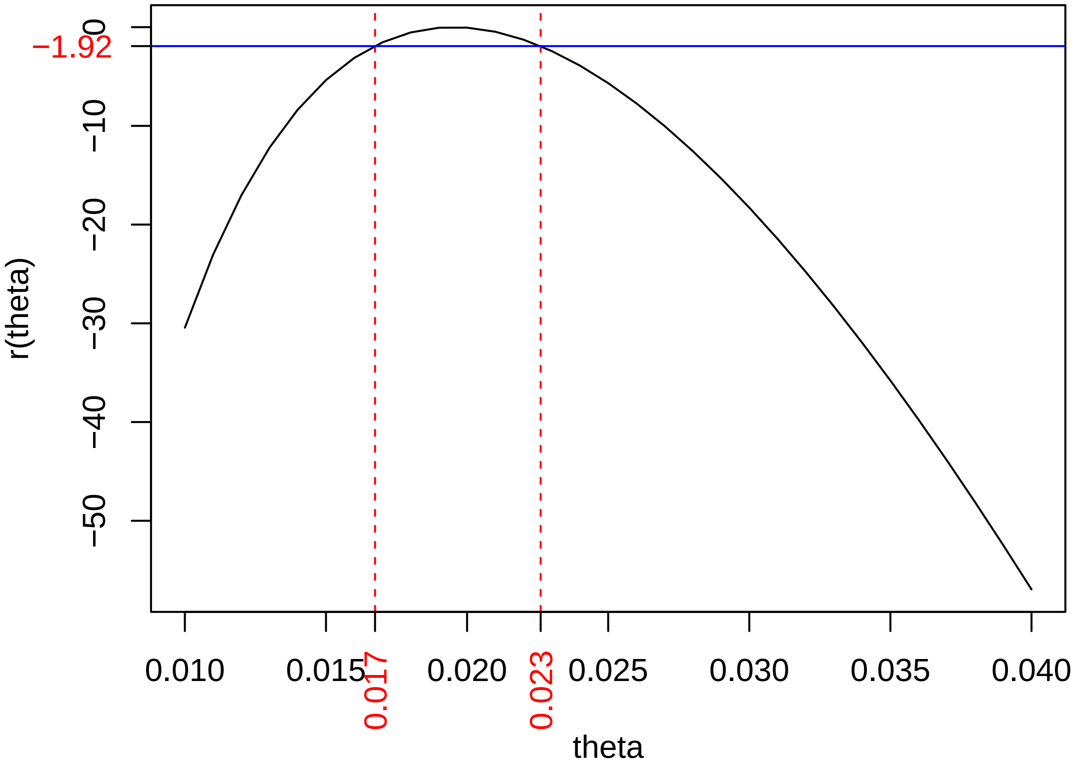
$$r(\theta) = \underbrace{\big(y\log(\theta) + (n-y)\log(1-\theta)\big)}_{\ell(\theta)} - \underbrace{\left(y\log\left(\frac{y}{n}\right) + (n-y)\log\left(1-\frac{y}{n}\right)\right)}_{\ell(\hat\theta)}.$$

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

```r
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 $(0.017, 0.023)$.

$$-2r(\theta) < \chi^2_1(0.95) \iff r(\theta) > -\frac{1}{2}\chi^2_1(0.95) = -1.92.$$



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:

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

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

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

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 $1.5\,\%$.

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

- Likelihood Ratio based test:

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

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

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

  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 $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^\top \in \Omega$ 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.

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

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

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

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

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

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

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

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

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

Solution.

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

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.

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

Some General Results for Score and Information

Result # 1

The expectation of the score function is zero.

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

Proof:

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

Result # 2

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

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

Proof: Differentiate (1) again,

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

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

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

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

Mean and Variance for the Exponential Family

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

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 $b''(\theta)=\mathsf{V}(\mu)$ and call $\mathsf{V}(\mu)$ the variance function.
- Subsequently, we have:

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

  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).

Example

Recall that POI(λ) is a member of exponential family,

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

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

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

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

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

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 $i = 1, \ldots, n$ in a random sample:

- Y_i is the response variable.
- $x_{i1}, \ldots, x_{ip}$ 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:

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

2. Systematic Component (or linear predictor):

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

   - $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ is a covariate vector.
   - $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ 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:

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

Example: A Poisson Regression Model

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

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

Poisson distribution is a member of exponential family with:

- Canonical parameter: $\theta_i = \log(\lambda_i)$.
- Canonical link: $\theta_i = \eta_i \implies \log(\lambda_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ (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, $i = 1, \ldots, n$:

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

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

- Canonical parameter: $\theta_i = \mu_i$.
- Canonical link: $\theta_i = \eta_i \implies \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ (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 $\boldsymbol{S}(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \boldsymbol{0}$. 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:

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

where $W^{-1} = \mathsf{Var}(Y)\left(\frac{\partial \eta}{\partial \mu}\right)^2$. 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 $Y_1, \ldots, Y_n$ 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:

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

where $W_i^{-1} = \mathsf{Var}(Y_i)(\frac{\partial \eta_i}{\partial \mu_i})^2$, $g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top\boldsymbol{\beta}$. 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,
- $\boldsymbol{\mathcal{W}} = \mathsf{diag}(W_1, \dots, W_n) = \begin{bmatrix} W_1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & W_n \end{bmatrix}$, and
- $\boldsymbol{\mathcal{A}} = \mathsf{diag}\left(\frac{\partial \eta_1}{\partial \mu_1}, \dots, \frac{\partial \eta_n}{\partial \mu_n}\right)$.

Example: Poisson Regression Model (Problem 1.4)

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

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

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.,

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

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

Solving $\boldsymbol{S}(\boldsymbol{\beta}) = 0$ 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:

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

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

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:

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

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-19		
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 $\pi_k = \mathbb{P}(\text{infection} \mid \text{group } k)$.

- 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:

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

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:

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

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

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

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

- 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.

	Fetal Mortality		
Level of Care	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

	Clinic A			*Clinic B*		
Level of Care	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

- $\hat{\psi}_{\text{A}} = 0.80\ (0.37, 1.73)$ and $\hat{\psi}_{\text{B}} = 1.01\ (0.33, 3.10)$. 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

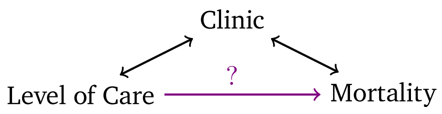
	A	B	
Intensive	309	27	336
Regular	118	231	419
	497	258	755

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

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.
- 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 $= \psi = \dfrac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$,

	Mortality		
Level of Care	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:

$$\exp\left\{\log(\hat{\psi}_1) \pm 1.96 \underbrace{\sqrt{\frac{1}{y_1} + \frac{1}{m_1 - y_1} + \frac{1}{y_2} + \frac{1}{m_2 - y_2}}}_{\text{se}(\log(\hat{\psi}))}\right\}.$$

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 cofounding 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:

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

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,

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

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

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

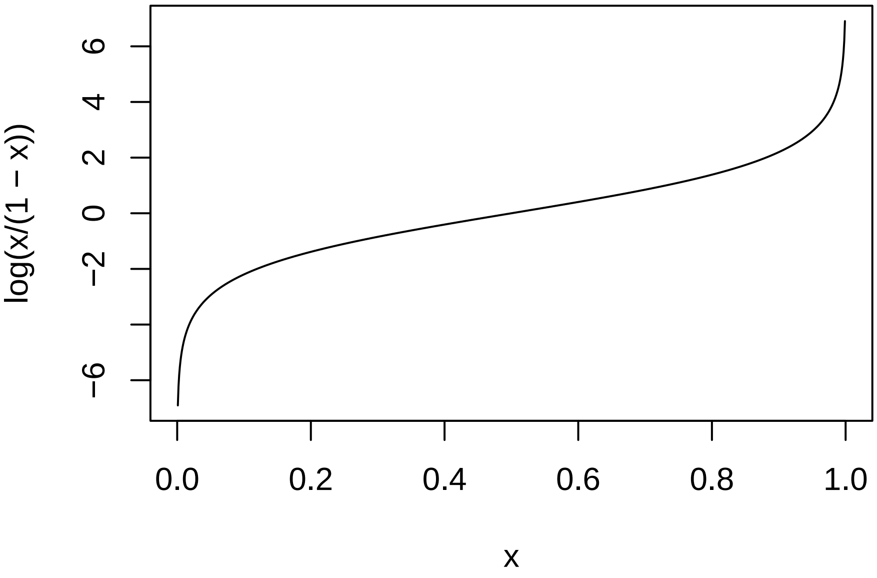
[a] For the Probit link, $\Phi(\,\cdot\,)$ is the *CDF* of $\mathcal{N}(0,1)$.

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, $g(\pi_i) = \log\big(\pi_i/(1-\pi_i)\big)$, is the canonical link for the Binomial!



This leads to a Logistic Regression Model:

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

Prediction from Logistic Regression

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

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

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 $\boldsymbol{x}_i$ is:

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

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	$\boldsymbol{x}_i^\top$	η_i	$= \log\big(\pi_i/(1-\pi_i)\big)$
1	$(1,1)^\top$	$\beta_0 + \beta_1$	$= \log\big(\pi_1/(1-\pi_1)\big)$
0	$(1,0)^\top$	β_0	$= \log\big(\pi_0/(1-\pi_0)\big)$
		β_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,

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

- 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	$\boldsymbol{x}_i^\top$	$\log(\pi_i/(1-\pi_i))$
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	$\boldsymbol{x}_i^\top$	$\log(\pi_i/(1-\pi_i))$
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

- $\beta_0 =$ log odds of mortality with regular care at Clinic B.
- $\beta_1 =$ log odds ratio of mortality for babies born to mothers treated with intensity vs regular care at the *same clinic*.
- $\beta_2 =$ 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	$\boldsymbol{x}_i^\top$	$\log(\pi_i/(1-\pi_i))$
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

- $\beta_1 =$ 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*.
- $\beta_2 =$ 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 dependent 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:

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

- 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:

```r
# 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

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

```r
# 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:

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

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

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

- `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:

$$H_0\colon \beta_j = \beta^\star \text{ versus } H_A\colon \beta_j \neq \beta^\star.$$

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

$$(\hat{\beta}_j - \beta^\star)^2 \left(I^{jj}(\hat{\boldsymbol{\beta}})\right)^{-1} \sim \chi^2_1,$$

equivalently $\dfrac{\hat{\beta}_j - \beta^\star}{\mathrm{se}(\hat{\beta}_j)} \sim \mathcal{N}(0,1)$ where $\mathrm{se}(\hat{\beta}_j) = \sqrt{I^{jj}(\hat{\boldsymbol{\beta}})}$.

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

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

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

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

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:

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

- 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:

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

- Confidence Interval for OR:

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

- 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

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

```
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?

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

- 95 % CI:

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

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	$\boldsymbol{x}_i^\top$	$\log(\pi_i/(1-\pi_i))$
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 $\text{Var}(\hat{\beta}) = \boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})$, now for any linear function of $\boldsymbol{\beta}$'s, e.g., $\boldsymbol{c\beta}$ where $\boldsymbol{c}$ is a row vector of constants, then MLE of $\boldsymbol{c\beta}$ is $\boldsymbol{c}\hat{\boldsymbol{\beta}}$, and $\text{se}(\widehat{\boldsymbol{c\beta}}) = \sqrt{\boldsymbol{cI}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{c}^\top}$.
- Therefore, $\log(\psi) = \beta_1 + \beta_3 = \boldsymbol{c\beta}$, $\boldsymbol{c} = (0,1,0,1)$. In R, `vcov(model3)` gives $\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})$.
- 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: $Y_i \sim \text{BIN}(m_i, \pi_i)$, $i = 1, \ldots, n$ independently, with covariate vector $\boldsymbol{x}_i$ and

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

- 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 $\boldsymbol{\beta}$:

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

- Log likelihood for logistic regression:

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

- Maximization of log-likelihood $\ell(\boldsymbol{\beta})$ gives MLE $\hat{\boldsymbol{\beta}}$, and
  - estimated probability of response:

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

  - 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):

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

Likelihood Ratio Test (General Setting)

- Suppose $\ell(\boldsymbol{\theta})$ is the likelihood for a q-dimension parameter vector $\boldsymbol{\theta}$ and let
  - $\tilde{\boldsymbol{\theta}}$ be the q-dim MLE of $\boldsymbol{\theta}$ (unconstrained/saturated, $q = n$),
  - $\hat{\boldsymbol{\theta}}$ be the p-dim MLE of $\boldsymbol{\theta}$ (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:

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

- Reject H_0 at θ if

$$p\text{-value} = \mathbb{P}\Bigl(\chi^2_{q-p} > -2\bigl[\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})\bigr]\Bigr) < \alpha.$$

Likelihood Ratio Test (Logistic Regression Model)

- Saturated (unconstrained) model MLEs:

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

  - 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:

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

  - 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 $\hat{y}_i = m_i \hat{\pi}_i = m_i \,\mathsf{expit}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})$.

- Hypotheses:

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

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

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

- 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 $D \sim \chi^2_{n-p}$ under H_0, and reject H_0 if $\mathbb{P}(\chi^2_{n-p} > D) < \alpha$.

  - 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,

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

  - 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.

```r
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: $\mathbb{P}(\chi^2_{n-p} > D) = \mathbb{P}(\chi^2_1 > D) = 0.7436689 \gg 0.05$.

- 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, $P \sim \chi^2_{n-p}$ 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 ($p < q < n$).

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

- 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 $\beta_p, \ldots, \beta_{q-1} \neq 0$.

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:

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

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

$$D_{\text{A}} = -2\big(\ell(\tilde{\boldsymbol{\pi}}) - \ell(\tilde{\tilde{\boldsymbol{\pi}}})\big).$$

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

$$\Delta D = D_0 - D_{\text{A}} = -2\big(\ell(\hat{\boldsymbol{\pi}}) - \ell(\tilde{\boldsymbol{\pi}})\big).$$

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

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

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

Example: Prenatal Care Data

- Summary of Deviance ("`residual deviance`") from R output:

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

- Compare nested models:
  - Model 2: $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$.
  - Model 4: $\text{logit}(\pi_i) = \beta_0 + \beta_2 x_{i2}$.
- 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:

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

  - 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:

$$D \overset{H_0}{\sim} \chi^2_{n-p}.$$

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

$$\Delta D = D_0 - D_{\text{A}} \overset{H_0}{\sim} \chi^2_{q-p}.$$

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),

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

- Fitted values:

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

- 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:

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

- Deviance Residual:

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

	Stage				
Age (months)	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

- Under H_0: the model is adequate:

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

- 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

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

```
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.

	Stage					
Age (months)	I	II	III	IV	V	Total
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 $Y_i \sim$ BIN(m_i, π_i), $i = 1, \ldots, 15$, 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,

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

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

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

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

- 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 11 12
2    1     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 11 12     11      1
2    1     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: $\hat{\pi}_i = \text{expit}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})$.
  - 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
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
```

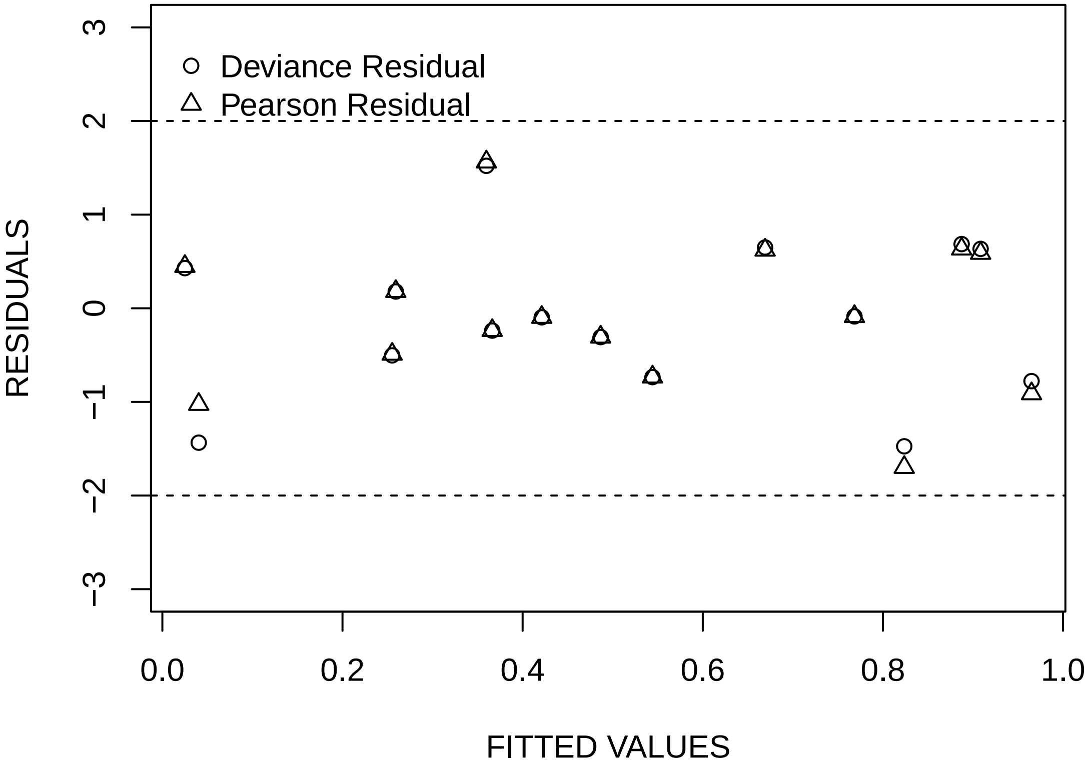


Figure 1: Plot of Residuals by Fitted Values for Neuroblastoma Data based on Logistic Regression Model with main effects of Age and Stage.

- 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-value $= \mathbb{P}(\chi^2_8 > 9.625) > 0.05$, 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:

$$\Delta D = D_0 - D_{\text{A}} = -2\big(\ell(\hat{\boldsymbol{\pi}}) - \ell(\tilde{\boldsymbol{\pi}})\big) \sim \chi^2_{q-p}$$

  - D_0 and D_{A} are deviances from the reduced and full models respectively.
  - $\hat{\boldsymbol{\pi}}$ and $\tilde{\boldsymbol{\pi}}$ 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: $\beta_3 = \cdots = \beta_6 = 0$ (Model 2 is as adequate as Model 1)
H_A: at least one of them is not 0 (Model 2 is not adequate)

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

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

```
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?

$$\begin{aligned} &H_0\colon \beta_1 = \beta_2 = 0 && \text{(Model 3 is as adequate as Model 1)} \\ &H_A\colon \text{at least one of them is not 0} && \text{(Model 3 is not adequate)} \end{aligned}$$

$$\Delta D = D_3 - D_1 = 42.446 - 9.625 = 32.821$$
$$p = \mathbb{P}(\chi^2_{7-5} > 32.821) < 0.001$$

```
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
```

$$p\text{-value} = \mathbb{P}(\chi^2_8 > 9.625) = 0.292 > 0.05$$

- 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	$\boldsymbol{x}_i^\top$	$\log(\pi_i/(1-\pi_i))$
0-11	—	$(1,0,0,x_{i3},x_{i4},x_{i5},x_{i6})^\top$	$\beta_0+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}+\beta_6x_{i6}$
24+	—	$(1,0,1,x_{i3},x_{i4},x_{i5},x_{i6})^\top$	$\beta_0+\beta_2+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}+\beta_6x_{i6}$

- 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:

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

- *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	$\boldsymbol{x}_i^\top$	$\log(\pi_i/(1-\pi_i))$
—	V	$(1,x_{i1},x_{i2},0,0,0,1)^\top$	$\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_6$
—	II	$(1,x_{i1},x_{i2},1,0,0,0)^\top$	$\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3$

- 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 $\psi = \exp\{\beta_6 - \beta_3\}$?

1. Finding the $95\,\%$ CI for $\eta = \beta_6 - \beta_3 = C\boldsymbol{\beta}$, where

$$C = \begin{bmatrix} 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_6 \end{bmatrix}.$$

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

$$\begin{aligned} \widehat{\mathsf{Var}}(\hat{\boldsymbol{\beta}}) &= \boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}) \\ \mathsf{se}(C\boldsymbol{\beta}) &= \sqrt{C\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})C^\top}. \end{aligned}$$

```
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:

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

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

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

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?

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

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

Modelling the Dose-Response Relationship

For each group $i = 1, \ldots, n$ 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 x_i	Responders y_i	Total m_i	 y_i/m_i
1.6907	6	59	0.10
1.7242	13	60	0.22
1.7552	18	62	0.29
⋮	⋮	⋮	⋮

- Assume

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

$$\pi_i = \text{probability of response in group } i \text{ with dose } x_i.$$

- 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:

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

  - 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:

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

Some Choices for the Tolerance Distribution

(1) Normal Tolerance Distribution:

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

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.,

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

(2) Logistic Distribution:

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

The probability of response:

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

This implies the Logit link s.t.,

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

③ Extreme Value Distribution:

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

The probability of response:

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

This implies the Complementary log-log link s.t.,

$$g(\pi)=\log\big(-\log(1-\pi)\big)=-\frac{\mu}{s}+\frac{1}{s}x=\beta_0+\beta_1x.$$

Tolerance Distribution	Link Function	Dose-Response Model
Normal	Probit	$\Phi^{-1}(\pi)=\beta_0+\beta_1x$
Logistic	Logit	$\operatorname{logit}(\pi)=\beta_0+\beta_1x$
Extreme Value	Complementary log-log	$\log\big(-\log(1-\pi)\big)=\beta_0+\beta_1x$

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:

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

- 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):

$$\Phi^{-1}(\pi)=\beta_0+\beta_1x.$$

  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 (CS_2) gas.

	# of insects	# of insects	
Dose (x_i)	killed (x_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
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, $i = 1, \ldots, 8$.
- 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
```

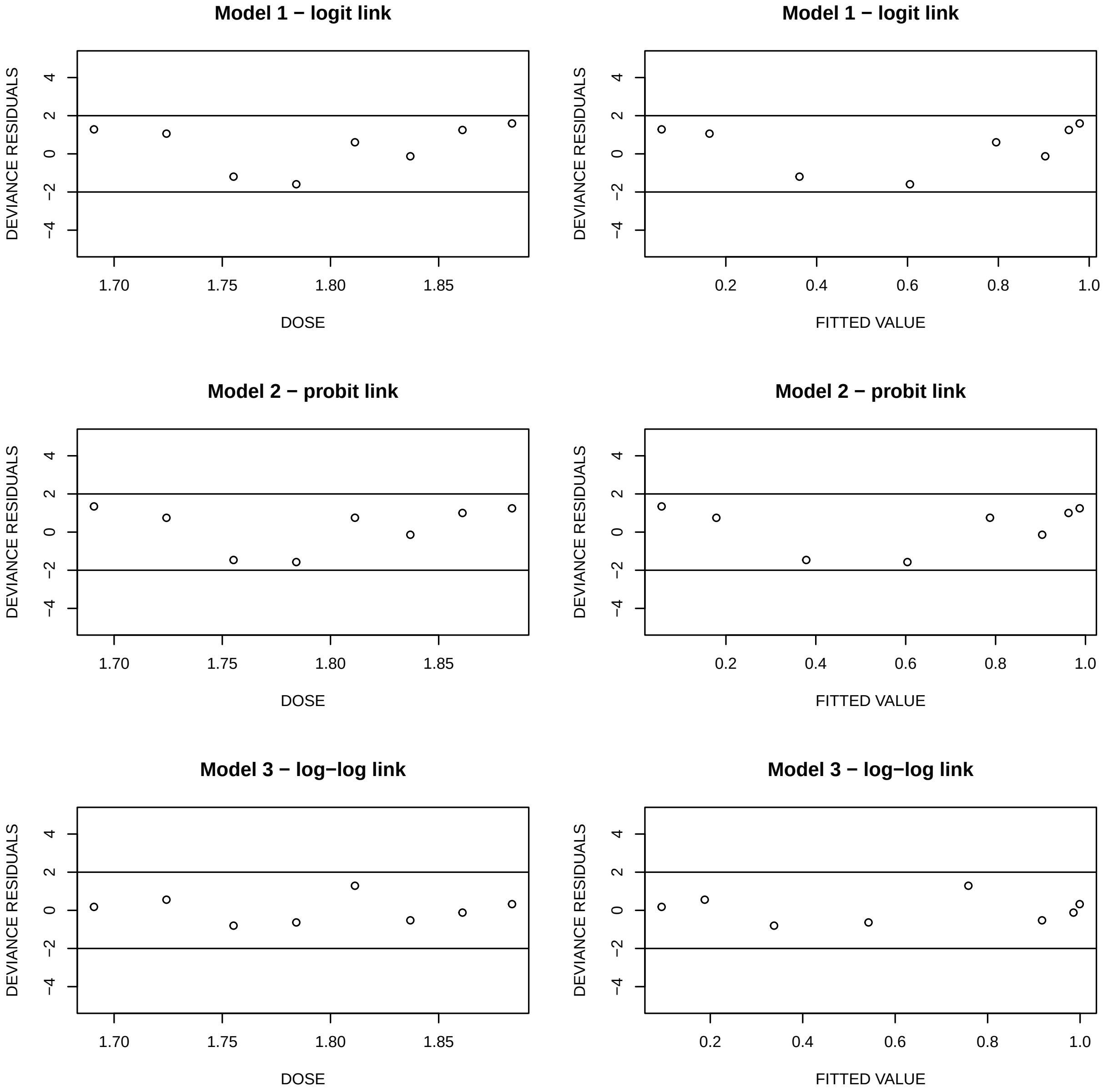
Fit of the Probit Model

```
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:

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

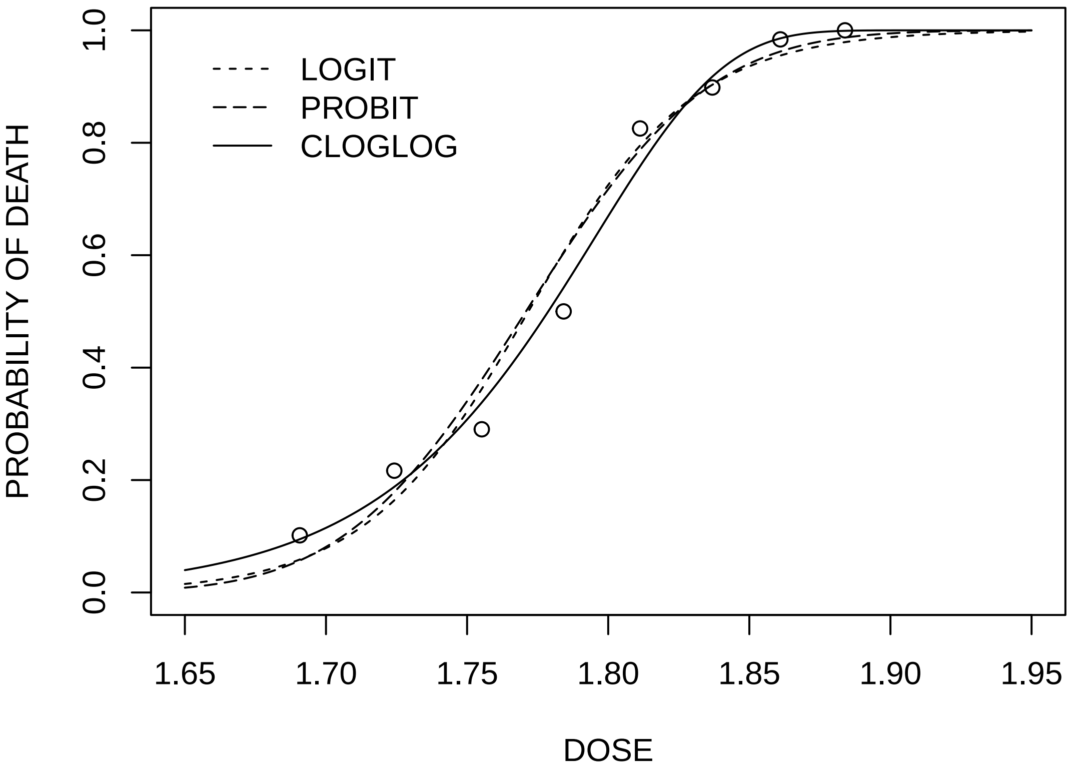
- Fitted probability of response:

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

- 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 expect 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 $\log\big(-\log(1-\pi)\big) = \beta_0 + \beta_1 x$, 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?

Reading week.

WEEK 7
18th to 22nd October

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
$$g(\pi_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$
where
  - $\boldsymbol{x}_i$ 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 $\boldsymbol{x}_i$ to probability π_i, and
  - $\boldsymbol{\beta}$ is a vector of regression parameters.
- When using the canonical link of Binomial distribution, i.e., $g(\,\cdot\,) = \mathsf{logit}(\,\cdot\,)$, we have
$$\mathsf{logit}(\pi_i) = \mathsf{log}\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i^\top \boldsymbol{\beta},$$
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 $\hat{\boldsymbol{\beta}} = \widehat{\mathsf{Var}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})$, where $\boldsymbol{I}^{-1}$ is the inverse of the information matrix evaluated at MLE $\hat{\boldsymbol{\beta}}$.
- Standard error: $\mathsf{se}(\hat{\beta}_k) = \sqrt{\left[\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})\right]_{kk}} = \sqrt{I^{kk}(\hat{\boldsymbol{\beta}})}$.
- Wald-test of a single parameter: H_0: $\beta_k = \beta^\star$ vs H_A: $\beta_k \neq \beta^\star$:
$$\frac{(\hat{\beta}_k - \beta^\star)^2}{I^{kk}(\hat{\beta}_k)} \overset{H_0}{\sim} \chi^2_1,$$
or
$$\frac{\hat{\beta}_k - \beta^\star}{\mathsf{se}(\hat{\beta}_k)} \sim \mathcal{N}(0,1) \text{ under } H_0.$$
For testing H_0: $\beta_k = 0$, we have $\frac{\hat{\beta}_k}{\mathsf{se}(\hat{\beta}_k)}$, reported as “`z-value`” in `glm()` summary.
- Confidence interval for a single β_k:
$$\hat{\beta}_k \pm Z_{1-\alpha/2}\,\mathsf{se}(\hat{\beta}_k).$$
- Confidence interval for $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$:
$$\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} \pm Z_{1-\alpha/2}\sqrt{\boldsymbol{x}_i^\top \boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i},$$
or equivalently
$$\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} \pm Z_{1-\alpha/2}\sqrt{\boldsymbol{x}_i^\top \widehat{\mathsf{Var}}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})\boldsymbol{x}_i}.$$
How about CI for $\pi_i = \mathsf{expit}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$?

Model Checking and Selection

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

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

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

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

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

$$\Delta D = D_0 - D_{\text{A}} \sim \chi^2_{p-q} \text{ under } H_0.$$

- Deviance Residuals:

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

  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 $\boldsymbol{x}_i$).
  - 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:

$$\pi(x) = \mathbb{P}(X \le x) = \int_{-\infty}^{x} f(x)\,\mathrm{d}x = F(x).$$

- 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)

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

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

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

  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:

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

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

- The likelihood and log-likelihood are then

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

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

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

  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:

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

Estimation of β from log linear regression

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

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

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

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

- These can also be found using general exponential family results.
- Use the above to estimate $\hat{\beta}$ 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., $\hat{\mu}_i \exp\{\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}\}$).
- Recall the Likelihood Ratio or Deviance Statistic has the form:

$$D = -2 \log\left(\frac{L(\hat{\boldsymbol{\mu}})}{L(\tilde{\boldsymbol{\mu}})}\right) = 2\big(\ell(\tilde{\boldsymbol{\mu}}) - \ell(\hat{\boldsymbol{\mu}})\big).$$

- Under H_0: constrained model is as adequate as saturated model, we have the following asymptotic distribution result:

$$D \sim \chi^2_{n-p}.$$

- For the Poisson we have:

$$\begin{aligned} D &= 2 \sum_{i=1}^{n} \Big(\big(y_i \log(\tilde{\mu}_i) - \tilde{\mu}_i\big) - \big(y_i \log(\hat{\mu}_i) - \hat{\mu}_i\big)\Big) \\ &= 2 \sum_{i=1}^{n} \left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right) \\ &= 2 \sum_{i=1}^{n} \left(O_i \log\left(\frac{O_i}{E_i}\right) - (O_i - E_i)\right). \end{aligned}$$

- 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:

$$\frac{\partial \ell}{\partial \beta_0} = \frac{\partial}{\partial \beta_0}\left[\sum_{i=1}^{n} y_i \boldsymbol{x}_i^\top \boldsymbol{\beta} - \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\}\right] = \sum_{i=1}^{n}(y_i - \mu_i) \implies \sum_{i=1}^{n}(y_i - \hat{\mu}_i) = 0,$$

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:

$$\Delta D = D_0 - D_{\mathrm{A}} \sim \chi^2_{q-p} \text{ under } H_0.$$

- The p-value for this test is given by:

$$p\text{-value} = \mathbb{P}(\chi^2_{q-p} > \Delta D).$$

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:

$$r_i^D = \mathrm{sign}(y_i - \hat{\mu}_i)\sqrt{|d_i|},$$

and are approximately $\mathcal{N}(0, 1)$ if H_0 holds.

- We can use the residual plots to evaluate the fit of a model.

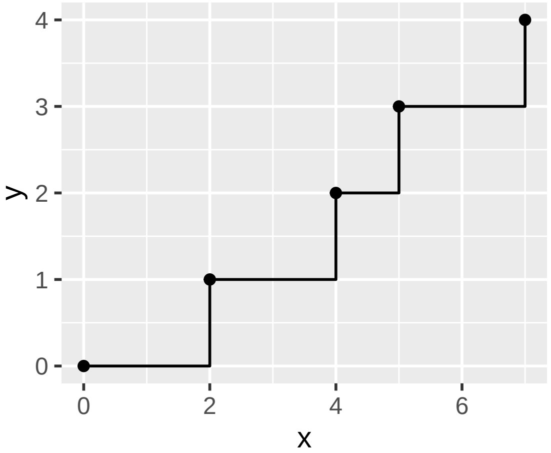
Regression for Poisson Processes

- 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.,

$$\mathbb{P}\big(N(t) = n; \lambda\big) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}, \ \lambda > 0, \ n = 0, 1, 2, \ldots.$$

- Expected number of events in $(0, t]$ is

$$\mathbb{E}\big[N(t)\big] = \mu(t) = \lambda t.$$

  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:

$$\log\big(\mu(t)\big) = \log(\lambda t) = \log(\lambda) + \log(t).$$

For each subject $i = 1, \ldots, n$ we observe:

- $N_i(t_i)$ = the number of events observed over $(0, t_i]$.
- Explanatory variables: $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip-1})^\top$.

Log Linear Regression Model for a Time Homogeneous Poisson Process

$$\begin{aligned} \log\big(\mu_i(t_i)\big) &= \log(\lambda_i) + \log(t_i) \\ &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \log(t_i) \qquad\qquad \text{e.g., } \log(\lambda_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}. \end{aligned}$$

- 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.

WEEK 8
25th to 29th October

Topic 4b: Ship Damage Example

Recall: Regression for Poisson Processes

1. Fitting the main effects log linear model:
   - Introduction of the data set.
   - Model 1: main effects + offset(log(months)).
2. Model selection:
   - Use Deviance tests of nested non-saturated models.
3. Model interpretation:
   - Show that β_k has log relative rate interpretation.
   - Wald based confidence intervals and hypothesis tests.

Example: Ship Damage Incidents

Example: Ship Damage Incidents

- McCullagh and Nelder (1989) discuss the analysis of a data set which records the number of times a certain type of damage incident occurs in cargo ships.
- Damage is caused by waves and occurs in the forward section of various cargo carrying vessels
- In order to prevent this type of damage from occurring in the future, the investigators want to identify risk factors including:
  - **Ship type** (A-E),
  - **Year of construction** (1960-1964; 1965-1969; 1970-1974; 1975-1979),
  - **Period of operation** (1960-1974; 1975-1979).

Ship Damage Data Set

In the dataset we have adopted the following coding conventions:

- `type`: The ship type variable is (1, 2, 3, 4, 5) for ship types A, B, C, D, and E, respectively
- `cyr`: The year of construction variable is (1, 2, 3, 4) for eras 1960-1964, 1965-1969, 1970-1974, and 1975-1979, respectively
- `oyr`: The year of operation variable is 1 for 1960-74 and 2 for 1975-1979
- `months`: The total number of months of operation for ships of that type and construction year during the period of operation
- `y`: The number of damage incidents for ships of that type and construction year during the period of operation

Ship Damage Data Set (`ship.dat`)

First ten rows of `ship.dat`:

```
   type cyr oyr months  y
1     1   1   1    127  0
2     1   1   2     63  0
3     1   2   1   1095  3
4     1   2   2   1095  4
5     1   3   1   1512  6
6     1   3   2   3353 18
7     1   4   2   2244 11
8     2   1   1  44882 39
9     2   1   2  17176 29
10    2   2   1  28609 58
```

R Code & Output (Models 1 and 2)

```r
# input dataset and create factor variables
ship.dat <- read.table("ship.dat", header = T)
ship.dat$typef <- factor(ship.dat$type)
ship.dat$cyrf <- factor(ship.dat$cyr)
ship.dat$oyrf <- factor(ship.dat$oyr)
ship.dat
# fitting the main effects with the offset term
model1 <- glm(y ~ typef + cyrf + oyrf + offset(log(months)), family = poisson,
  data = ship.dat)
summary(model1)
# fitting all main effects (treating offset as a covariate for
# diagnostics)
model2 <- glm(y ~ typef + cyrf + oyrf + log(months), family = poisson,
  data = ship.dat)
summary(model2)
```

Model 1: Main effects + offset(log(months))

- Time homogenous Poisson process: $\mathbb{E}\big[N_i(t_i)\big] = \mu_t(t_i) = \lambda t_i$.
- Log linear regression model:

$$\log\big(\mu_i(t_i)\big) = \log(\lambda_i) + \log(t_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \log(t_i).$$

- Ship Damage main effects model:

$$\begin{aligned}\log\big(\mu_i(t_i)\big) = \beta_0 + {} & \overbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}}^{\text{ship type}} + \\ & \underbrace{\beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7}}_{\text{year of construction}} + \underbrace{\beta_8 x_{i8}}_{\text{operation year}} + \underbrace{\log(t_i)}_{\text{offset}},\end{aligned}$$

where

$$\begin{aligned} x_{i1} &= \mathbb{I}\{\text{type B}\}, & x_{i5} &= \mathbb{I}\{\text{1965-1969}\}, \\ x_{i2} &= \mathbb{I}\{\text{type C}\}, & x_{i6} &= \mathbb{I}\{\text{1970-1974}\}, \\ x_{i3} &= \mathbb{I}\{\text{type D}\}, & x_{i7} &= \mathbb{I}\{\text{1975-1979}\}, \\ x_{i4} &= \mathbb{I}\{\text{type E}\}, & x_{i8} &= \mathbb{I}\{\text{1975-1979}\}. \end{aligned}$$

Model 1: Main effects + offset(log(months))

```
summary(model1)


Call:
glm(formula = y ~ typef + cyrf + oyrf + offset(log(months)),
    family = poisson, data = ship.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6768  -0.8293  -0.4370   0.5058   2.7912

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.40590    0.21744 -29.460  < 2e-16 ***
typef2      -0.54334    0.17759  -3.060  0.00222 **
typef3      -0.68740    0.32904  -2.089  0.03670 *
typef4      -0.07596    0.29058  -0.261  0.79377
typef5       0.32558    0.23588   1.380  0.16750
cyrf2        0.69714    0.14964   4.659 3.18e-06 ***
cyrf3        0.81843    0.16977   4.821 1.43e-06 ***
cyrf4        0.45343    0.23317   1.945  0.05182 .
oyrf2        0.38447    0.11827   3.251  0.00115 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 146.328  on 33  degrees of freedom
Residual deviance:  38.695  on 25  degrees of freedom
AIC: 154.56

Number of Fisher Scoring iterations: 5
```

Model 2: Main effects + log(months)

```
summary(model2)


Call:
glm(formula = y ~ typef + cyrf + oyrf + log(months), family = poisson,
    data = ship.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6580  -0.8939  -0.4900   0.4676   2.7435

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.5940     0.8724  -6.412 1.43e-10 ***
typef2       -0.3499     0.2702  -1.295  0.19539
typef3       -0.7631     0.3382  -2.257  0.02404 *
typef4       -0.1355     0.2971  -0.456  0.64842
typef5        0.2739     0.2418   1.133  0.25719
cyrf2         0.6625     0.1536   4.312 1.61e-05 ***
cyrf3         0.7597     0.1777   4.276 1.90e-05 ***
cyrf4         0.3697     0.2458   1.504  0.13259
oyrf2         0.3703     0.1181   3.134  0.00172 **
log(months)   0.9027     0.1018   8.867  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 614.539  on 33  degrees of freedom
Residual deviance:  37.804  on 24  degrees of freedom
AIC: 155.67

Number of Fisher Scoring iterations: 5
```

Summary of Model 1 versus Model 2

- $\log(\cdot)$ is the canonical link for the Poisson, so it is the default when `family=poisson`.
- Model 1: main effects + `offset(log(months))`: $\boldsymbol{x}^\top\boldsymbol{\beta} = \log(t_i)$.
  - The offset explains some variation in the number of damage incidents due to different amounts of time at risk.
- Model 2: main effects + `log(months)`: $\boldsymbol{x}^\top\boldsymbol{\beta}\log(t_i)$.
  - Examine $\hat{\beta}_9$ the coefficient for `log(months)`.
  - Conduct a Wald-based test of H_0: $\beta_9 = 1$ versus H_A: $\beta_9 \neq 1$:

$$p = 2\,\mathbb{P}\left(Z > \frac{|\hat{\beta}_9 - 1|}{\mathsf{se}(\hat{\beta}_9)}\right) = 2\,\mathbb{P}(Z > \frac{|0.9027 - 1|}{0.1018}) = 2\,\mathbb{P}(Z > |-0.9558|) = 0.34.$$

  Therefore, do not reject H_0: $\beta_9 = 1$.

- We will not typically do this check and just use `offset(log(ti))` since it's implied through the assumption of a time homogenous Poisson Process.

R Code (Models 3a, 3b, 3c)

Now, consider various models nested within model 1 to see if any of the main effects are not significant.

```r
# testing for the association between ship type and frequency of
# events
model3a <- glm(y ~ cyrf + oyrf + offset(log(months)), family = poisson,
  data = ship.dat)
model3a$deviance
model3a$df.residual
1 - pchisq(model3a$deviance - model1$deviance, model3a$df.residual - model1$df.residual)
# testing for association between year of construction and event
# frequency
model3b <- glm(y ~ typef + oyrf + offset(log(months)), family = poisson,
  data = ship.dat)
model3b$deviance
model3b$df.residual
1 - pchisq(model3b$deviance - model1$deviance, model3b$df.residual - model1$df.residual)
# testing for the association between year of operation and event
# frequency
model3c <- glm(y ~ typef + cyrf + offset(log(months)), family = poisson,
  data = ship.dat)
model3c$deviance
model3c$df.residual
1 - pchisq(model3c$deviance - model1$deviance, model3c$df.residual - model1$df.residual)
```

Model 3a: `cyrf + oyrf + offset(log(months))`

- Use this model to test:
  - H_0: Type of Ship is unimportant (i.e., $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$).
  - H_A: $\beta_1 \neq 0$ or $\cdots$ or $\beta_4 \neq 0$.

```r
model3a <- glm(y ~ cyrf + oyrf + offset(log(months)), family = poisson,
  data = ship.dat)
model3a$deviance

[1] 62.36534

model3a$df.residual

[1] 29

1 - pchisq(model3a$deviance - model1$deviance, model3a$df.residual - model1$df.residual)

[1] 9.299568e-05
```

$$\Delta D = D_0 - D_A \sim \chi^2_{(4)} \text{ under } H_0.$$

$$p = \mathbb{P}\Big(\chi^2_{(4)} > (62.365 - 38.695)\Big) < 0.001.$$

- Reject the null hypothesis of no variation in the accident rate across ships of different types.
- This is strong evidence of a need to adjust for the difference in the accident rates between ship types.

Model 3b: `typef + oyrf + offset(log(months))`

- Use this model to test:
  - H_0: Construction year is unimportant (i.e., $\beta_5 = \beta_6 = \beta_7 = 0$).
  - H_A: $\beta_5 \neq 0$ or $\cdots$ or $\beta_7 \neq 0$.

```
model3b <- glm(y ~ typef + oyrf + offset(log(months)), family = poisson,
  data = ship.dat)
model3b$deviance

[1] 70.10294

model3b$df.residual

[1] 28

1 - pchisq(model3b$deviance - model1$deviance, model3b$df.residual - model1$df.residual)

[1] 6.974977e-07
```

- Reject the null hypothesis of no variation in the accident rate across ships of different construction years.

Model 3c: `typef + cyrf + offset(log(months))`

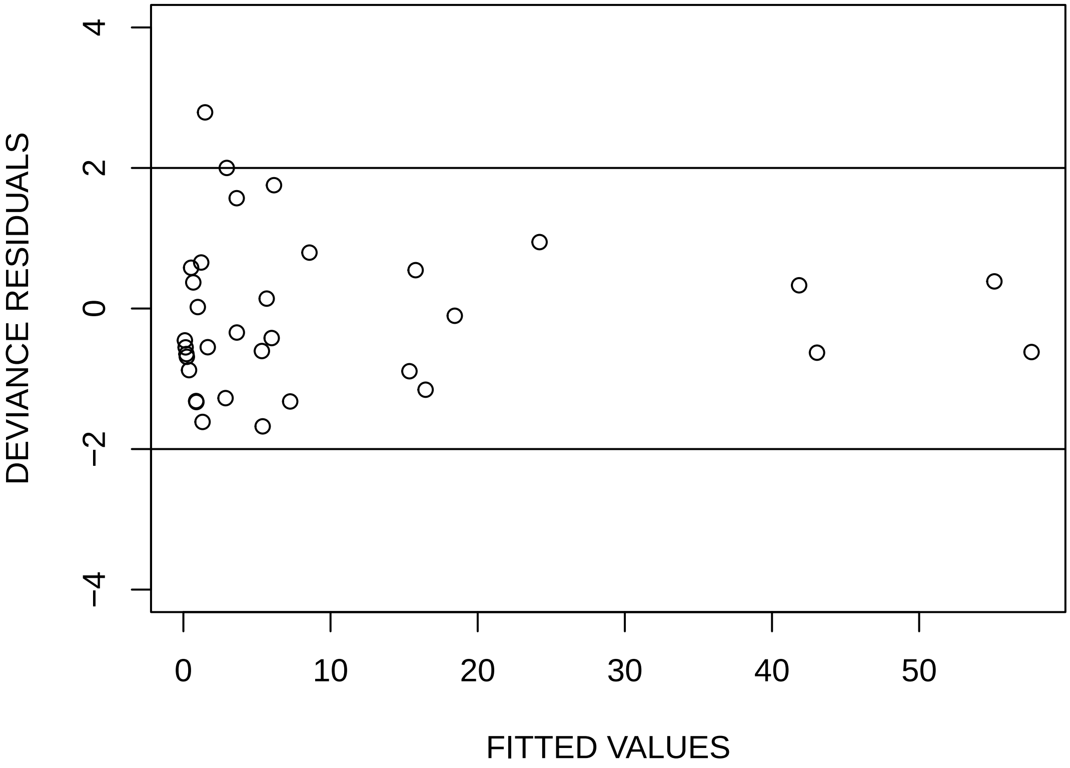
- Use this model to test:
  - H_0: Operation year is unimportant (i.e., $\beta_8 = 0$).
  - H_A: $\beta_8 \neq 0$.

```
model3c <- glm(y ~ typef + cyrf + offset(log(months)), family = poisson,
  data = ship.dat)
model3c$deviance

[1] 49.35519

model3c$df.residual

[1] 26

1 - pchisq(model3c$deviance - model1$deviance, model3c$df.residual - model1$df.residual)

[1] 0.001094692
```

- Reject the null hypothesis of no variation in the accident rate across ships of different periods of operation.
- We are unable to remove any of the main effects from the model (all are statistically significant).
- Next, consider adding interaction effects.

R Code (Models 4, 5, 6)

```
# testing for the interaction between type of ship and year of
# construction
model4 <- glm(y ~ typef + cyrf + oyrf + typef * cyrf + offset(log(months)),
  family = poisson, data = ship.dat)
model4$deviance
model4$df.residual
1 - pchisq(model1$deviance - model4$deviance, model1$df.residual - model4$df.residual)
summary(model4)
mrho <- summary(model4, corr = T)$correlation
mrho
# testing for the interaction between type of ship and year of
# operation
model5 <- glm(y ~ typef + cyrf + oyrf + typef * oyrf + offset(log(months)),
  family = poisson, data = ship.dat)
1 - pchisq(model1$deviance - model5$deviance, model1$df.residual - model5$df.residual)
# testing for the interaction between year of construction and
# operation
model6 <- glm(y ~ typef + cyrf + oyrf + cyrf * oyrf + offset(log(months)),
  family = poisson, data = ship.dat)
1 - pchisq(model1$deviance - model6$deviance, model1$df.residual - model6$df.residual)
# plot the residuals
ship.dat$rdeviance <- residuals.glm(model1, type = "deviance")
plot(model1$fitted.values, ship.dat$rdeviance, ylim = c(-4, 4), xlab = "FITTED VALUES",
  ylab = "DEVIANCE RESIDUALS")
abline(h = -2)
abline(h = 2)
```



Model 4: typef + cyrf + oyrf + typef*cyrf + offset(log(months))

- Use this model to test:
  - H_0: the typef*cyrf interaction is unimportant (Model 1).
  - H_A: (model 4).

```
model4 <- glm(y ~ typef + cyrf + oyrf + typef * cyrf + offset(log(months)),
  family = poisson, data = ship.dat)
model4$deviance

[1] 14.58688

model4$df.residual

[1] 13

1 - pchisq(model1$deviance - model4$deviance, model1$df.residual - model4$df.residual)

[1] 0.01966268
```

$$\Delta D = D_0 - D_A \sim \chi^2_{(12)} \text{ under } H_0.$$

$$p = \mathbb{P}\Big(\chi^2_{(12)} > (38.695 - 14.587)\Big) < 0.0197.$$

  - Reject the null hypothesis that the main effects model is adequate.
  - That is, we would choose model 4 over model 1.

```
summary(model4)


Call:
glm(formula = y ~ typef + cyrf + oyrf + typef * cyrf + offset(log(months)),
    family = poisson, data = ship.dat)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.99643  -0.09176  -0.00008   0.13849   2.53827

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -23.9891  6625.5245  -0.004  0.99711
typef2          17.0506  6625.5245   0.003  0.99795
typef3          17.0863  6625.5245   0.003  0.99794
typef4          -0.5962  9331.1044   0.000  0.99995
typef5           0.8799 11522.0954   0.000  0.99994
cyrf2           18.0324  6625.5245   0.003  0.99783
cyrf3           18.3969  6625.5245   0.003  0.99778
cyrf4           18.2860  6625.5245   0.003  0.99780
oyrf2            0.3850     0.1186   3.246  0.00117 **
typef2:cyrf2   -17.3620  6625.5245  -0.003  0.99791
typef3:cyrf2   -18.6108  6625.5246  -0.003  0.99776
typef4:cyrf2   -18.4024 11467.2826  -0.002  0.99872
typef5:cyrf2     0.4496 11522.0955   0.000  0.99997
typef2:cyrf3   -17.6110  6625.5245  -0.003  0.99788
typef3:cyrf3   -17.6160  6625.5246  -0.003  0.99788
typef4:cyrf3     1.0922  9331.1044   0.000  0.99991
typef5:cyrf3    -0.8285 11522.0954   0.000  0.99994
typef2:cyrf4   -17.7124  6625.5245  -0.003  0.99787
typef3:cyrf4   -17.3813  6625.5246  -0.003  0.99791
typef4:cyrf4    -0.3254  9331.1044   0.000  0.99997
typef5:cyrf4    -1.8570 11522.0955   0.000  0.99987
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 146.328  on 33  degrees of freedom
Residual deviance:  14.587  on 13  degrees of freedom
AIC: 154.45

Number of Fisher Scoring iterations: 17
```

- Huge standard errors!
- This model is overparameterized!

- Twelve interaction terms.
- Type 4: no events for `cyr` 1 or 2.

Model 5: `typef + cyrf + oyrf + typef*oyrf + offset(log(months))`

- Use this model to test:
  - H_0: the `typef*oyrf` interaction is unimportant (Model 1).
  - H_A: (model 5).

```
model5 <- glm(y ~ typef + cyrf + oyrf + typef * oyrf + offset(log(months)),
  family = poisson, data = ship.dat)
1 - pchisq(model1$deviance - model5$deviance, model1$df.residual - model5$df.residual)

[1] 0.2936317
```

  - Do not reject the null hypothesis that the main effects model is adequate.
  - The interaction between ship type and year of operation is not significant.

Model 6: `typef + cyrf + oyrf + cyrf*oyrf + offset(log(months))`

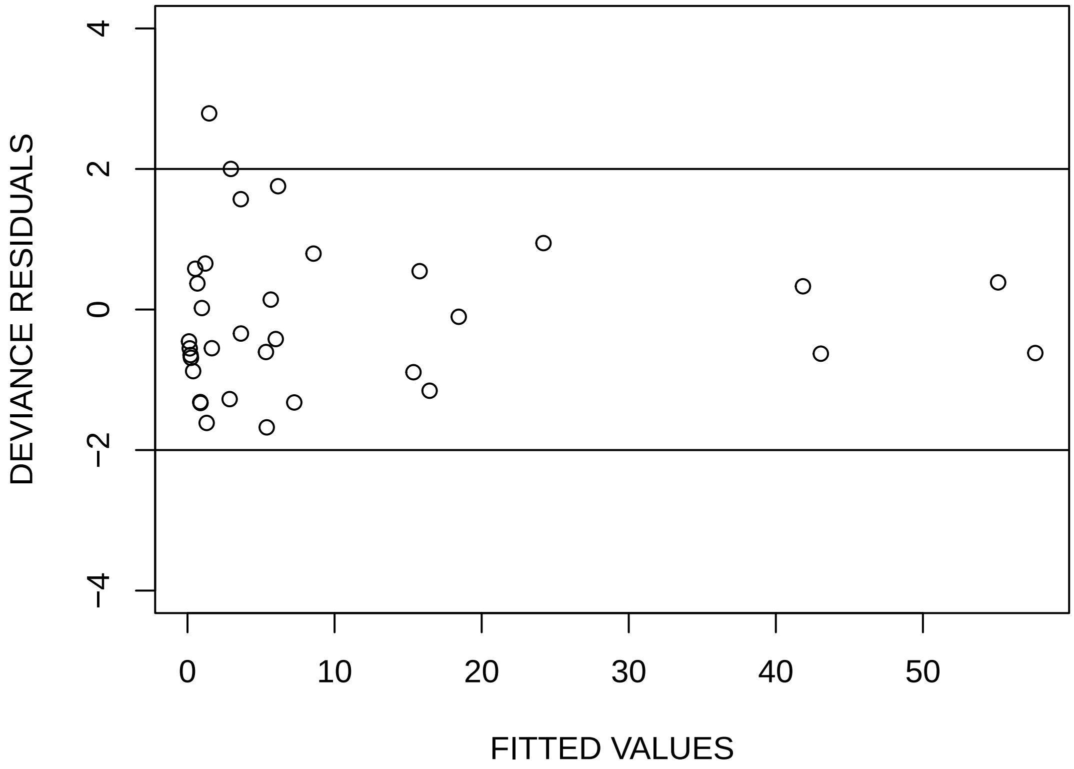
- Use this model to test:
  - H_0: the `cyrf*oyrf` interaction is unimportant (Model 1).
  - H_A: (model 6).

```
model6 <- glm(y ~ typef + cyrf + oyrf + cyrf * oyrf + offset(log(months)),
  family = poisson, data = ship.dat)
1 - pchisq(model1$deviance - model6$deviance, model1$df.residual - model6$df.residual)

[1] 0.4091268
```

  - Do not reject the null hypothesis that the main effects model is adequate.
  - The interaction between year of construction and year of operation is not significant.

Model 1: `typef + cyrf + oyrf +offset(log(months))`

- Conclude that the best fitting model is the main effects model.
- Check the residual plot:



- $\hat{\mu}_i = \exp\left\{\boldsymbol{x}_i^\top \hat{\beta} + \log(t_i)\right\}$.
- $D = \sum_i 2 \log\left(\frac{y_i}{\hat{\mu}_i}\right) = \sum_i d_i$.
- $r_i^d = \text{sign}(y_i - \hat{\mu}_i)\sqrt{|d_i|}$.

Interpretation of Model 1: Main effects + `offset(log(months))`

```
model1 <- glm(y ~ typef + cyrf + oyrf + offset(log(months)), family = poisson,
  data = ship.dat)
summary(model1)


Call:
glm(formula = y ~ typef + cyrf + oyrf + offset(log(months)),
    family = poisson, data = ship.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6768  -0.8293  -0.4370   0.5058   2.7912

Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -6.40590    0.21744 -29.460  < 2e-16 ***
typef2      -0.54334    0.17759  -3.060  0.00222 ** 
typef3      -0.68740    0.32904  -2.089  0.03670 *  
typef4      -0.07596    0.29058  -0.261  0.79377    
typef5       0.32558    0.23588   1.380  0.16750    
cyrf2        0.69714    0.14964   4.659 3.18e-06 ***
cyrf3        0.81843    0.16977   4.821 1.43e-06 ***
cyrf4        0.45343    0.23317   1.945  0.05182 .  
oyrf2        0.38447    0.11827   3.251  0.00115 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 146.328  on 33  degrees of freedom
Residual deviance:  38.695  on 25  degrees of freedom
AIC: 154.56

Number of Fisher Scoring iterations: 5
```

```
summary(model1)$cov.unscaled
```

```
             (Intercept)        typef2        typef3       typef4        typef5
(Intercept)  0.047281921 -0.0313338453 -0.0270722494 -0.023415086 -0.0241001095
typef2      -0.031333845  0.0315381717  0.0253121856  0.023057691  0.0239048348
typef3      -0.027072249  0.0253121856  0.1082700615  0.022710437  0.0243415185
typef4      -0.023415086  0.0230576907  0.0227104372  0.084435953  0.0228773141
typef5      -0.024100109  0.0239048348  0.0243415185  0.022877314  0.0556390922
cyrf2       -0.015756834  0.0022749529  0.0017647174  0.001203482 -0.0001442043
cyrf3       -0.020308913  0.0081833789  0.0025425848  0.001410245 -0.0014853352
cyrf4       -0.020358789  0.0094600451  0.0074478910 -0.006543921  0.0029036746
oyrf2       -0.005558091  0.0005331834 -0.0001195467 -0.000162536  0.0007514856
                   cyrf2        cyrf3        cyrf4         oyrf2
(Intercept) -0.0157568339 -0.020308913 -0.020358789 -0.0055580913
typef2       0.0022749529  0.008183379  0.009460045  0.0005331834
typef3       0.0017647174  0.002542585  0.007447891 -0.0001195467
typef4       0.0012034818  0.001410245 -0.006543921 -0.0001625360
typef5      -0.0001442043 -0.001485335  0.002903675  0.0007514856
cyrf2        0.0223925335  0.016093453  0.016591557 -0.0021248406
cyrf3        0.0160934529  0.028823065  0.021702485 -0.0052926926
cyrf4        0.0165915573  0.021702485  0.054368442 -0.0086966553
oyrf2       -0.0021248406 -0.005292693 -0.008696655  0.0139882936
```

Interpretation of Log Linear Models for Poisson Processes

- Focus on interpretation of Model 1, the main effects model.
- Recall the form of the model

$$\log\big(\mu_i(t_i)\big) = \log(\lambda_i) + \log(t_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \log(t_i).$$

- This is based on the Poisson distribution with the expected number of events occurring over $(0, t]$ given by

$$\mathbb{E}\big[N_i(t_i)\big] = \mu_i(t_i) = \lambda_i t_i.$$

- $\lambda =$ rate parameter (expected number of events per unit time).
- The regression parameters of this log linear model will have a log Relative Rate (RR) interpretation:

$$\text{RR} = \frac{\lambda_1}{\lambda_2} = \frac{\text{Number of events in group 1 per unit time}}{\text{Number of events in group 2 per unit time}}.$$

Interpretation of Model 1: RR for A vs C

Task 1: Controlling for periods of construction and operation estimate the relative rate of accidents for ships of type A versus type C.

type	cyr	oyr	$\boldsymbol{x}_i$	$\log(\lambda_i)$
A	—	—	$(1,0,0,0,0,x_5,x_6,x_7,x_8)$	$\beta_0+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8$
C	—	—	$(1,0,1,0,0,x_5,x_6,x_7,x_8)$	$\beta_0+\beta_2+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8$
			$\log(\lambda_A/\lambda_C) =$	$-\beta_2$

	β_2	$\exp\{-\beta_2\}$
MLE	-0.6874	1.990
95 % CI	$-0.6874 \pm 1.96(0.329) = (-1.332, -0.0426)$	$(e^{0.0426}, e^{1.332}) = (1.04, 3.79).$

For ships constructed in the same period and operated in the same period the rate of accidents for ships of type A is 1.99 times higher than the rate of accidents for ships of type C. A 95% confidence interval for this relative rate is $(1.04, 3.79)$.

- Note that the null hypothesis of no effect is equivalent to RR $= 1$ or $\log(\text{RR}) = 0$:

$$H_0\colon \beta_2 = 0 \text{ versus } H_A\colon \beta_2 \neq 0.$$

- The R output includes the p-value for this test:

$$2\,\mathbb{P}\left(Z > \frac{|\hat{\beta}_2 - 0|}{\text{se}(\hat{\beta}_2)}\right) = 2\,\mathbb{P}(Z > 2.089) = 0.0367.$$

- Therefore, we reject the null hypothesis that the rate of accidents is the same for ships of types A and C (controlling for periods of construction and operation).

Interpretation of Model 1: RR for E vs B

Task 2: Controlling for periods of construction and operation estimate the relative rate of accidents for ships of type E versus type B.

type	cyr	oyr	$\boldsymbol{x}_i$	$\log(\lambda_i)$
E	—	—	$(1,0,0,0,1,x_5,x_6,x_7,x_8)$	$\beta_0+\beta_4+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8$
B	—	—	$(1,1,0,0,0,x_5,x_6,x_7,x_8)$	$\beta_0+\beta_1+\beta_5x_5+\beta_6x_6+\beta_7x_7+\beta_8x_8$
			$\log(\lambda_E/\lambda_B) =$	$\beta_4 - \beta_1$

- Note that the log relative risk is a linear combination of two regression parameters.
- Recall that since $\hat{\boldsymbol{\beta}}$ is an MLE, $\hat{\boldsymbol{\beta}} \sim \text{MVN}\Big(\boldsymbol{\beta}, I^{-1}(\hat{\boldsymbol{\beta}})\Big)$

$$\boldsymbol{x}^\top \hat{\boldsymbol{\beta}} \sim \mathcal{N}\Big(\boldsymbol{x}^\top \boldsymbol{\beta}, \boldsymbol{x}^\top I^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{x}\Big).$$

- In order to estimate $\text{se}(\beta_4 - \beta_1)$:
  (i) If working in R, we can define the contrast $\boldsymbol{c} = (0, -1, 0, 0, 1, 0, 0, 0, 0)^\top$ and

$$\text{se}(\hat{\beta}_4 - \hat{\beta}_1) = \sqrt{\boldsymbol{c}^\top \boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{c}}.$$

```
x = as.matrix(c(0, -1, 0, 0, 1, 0, 0, 0, 0), ncol = 1)
v = summary(model1)$cov.unscaled
sqrt(t(x) %*% v %*% x)

          [,1]
[1,] 0.1984127
```

  (ii) If working by hand with the R covariance or correlation matrix:

$$\begin{aligned}\text{se}(\hat{\beta}_4 - \hat{\beta}_1) &= \sqrt{\text{Var}(\hat{\beta}_4) + \text{Var}(\hat{\beta}_1) - 2\,\text{Cov}(\hat{\beta}_4, \hat{\beta}_1)} \\ &= \sqrt{(0.05564) + (0.03154) - 2(0.02390)} \\ &= 0.198.\end{aligned}$$

- Now to estimate the relative rate $\text{exp}\{\beta_4 - \beta_1\}$:

	$\beta_4 - \beta_1$	$\text{exp}\{\beta_4 - \beta_1\}$
MLE	$-0.3256 - (-0.5433) = 0.8689$	$\text{exp}\{0.08689\} = 2.38$
95 % CI	$0.8669 \pm 1.96(0.198) = (0.481, 1.257)$	$(e^{0.481}, e^{1.257}) = (1.62, 3.51)$

- *For ships constructed and operated in the same periods those of type E had an estimated* 2.38*, 95 % CI* $(1.62, 3.51)$*, times higher accident rate than those of type B.*
- Here the null hypothesis of no effect is that ships of types E and B have the same accident rate. That is,

$$H_0\colon \beta_4 - \beta_1 = 0 \text{ vs } H_A\colon \beta_4 - \beta_1 \neq 0.$$

- We test this using a Wald test. Since $\boldsymbol{x}^\top \hat{\boldsymbol{\beta}} \sim \mathcal{N}\Big(\boldsymbol{x}^\top \boldsymbol{\beta}, \boldsymbol{x}^\top I^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{x}\Big)$. Then

$$\frac{\boldsymbol{x}^\top \hat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{x}^\top I^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{x}}} \sim \mathcal{N}(0, 1).$$

- The p-value for this test is:

$$2\,\mathbb{P}\left(Z > \frac{\boldsymbol{x}^\top \hat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{x}^\top I^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{x}}}\right) = 2\,\mathbb{P}\left(Z > \frac{0.8689}{0.198}\right) < 0.001.$$

- Therefore, we reject the null hypothesis that the accident rate is the same for ships of types E and B (controlling for periods of contraction and operation).

Interpretation of Model 1: Expected Number Events

Task 3: Estimate the expected number of accidents for a group of 10 type B ships built in 1970 and operated during the entire period 1975-1979.

$$\log\big(\mu_i(t_i)\big) = \log(\lambda_i) + \log(t_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \log(t_i).$$

- Estimate $\log(\lambda_i)$ the log of the event rate and its CI:

type	cyr	oyr	$\boldsymbol{x}_i$	$\log(\lambda_i)$
B	70-74	75-79	$(1, 1, 0, 0, 0, 0, 1, 0, 1)$	$\beta_0 + \beta_1 + \beta_6 + \beta_8$

```
x = as.matrix(c(1, 1, 0, 0, 0, 0, 1, 0, 1), ncol = 1)
v = summary(model1)$cov.unscaled
t(x) %*% model1$coeff

          [,1]
[1,] -5.746352

sqrt(t(x) %*% v %*% x)

          [,1]
[1,] 0.1186486

t(x) %*% model1$coef + c(-1, 1) * qnorm(0.975) * sqrt(t(x) %*% v %*% x)

[1] -5.978899 -5.513805
```

Interpretation of Model 1: Expected Number Events

- Determine the offset t_i = months:

$$\begin{aligned} t_i &= \text{total amount of time at risk of an accident} \\ &= (\text{\# ships})(\text{length of operation}) \\ &= (10)(5 \times 12) \\ &= 600. \end{aligned}$$

- Calculate the expected number of accidents $\hat{\mu}_i$:

$$\begin{aligned} \log(\hat{\mu}_i) &= \log(\hat{\lambda}_i) + \log(t_i) \\ \hat{\mu}_i &= \hat{\lambda}_i t_i \\ &= \exp\{-5.7463\} \times 600 \\ &= 1.92. \end{aligned}$$

- With 95 % CI: $(600e^{-5.5138}, 600e^{-5.9789}) = (1.52, 2.42)$.
- *The estimated number of accidents for a group of 10 type B ships built in 1970 and operated during the entire period 1975-1979 is* 1.92 *with a 95 % CI of* $(1.52, 2.42)$.

Topic 4c: Poisson Approximation to the Binomial Distribution

Log Linear Models

Previously we used a Poisson GLM to model count data arising from a time homogeneous Poisson process:

- $N_i(t_i)$ = the number of events observed over $(0, t_i]$:
$$\mathbb{E}\big[N_i(t_i)\big] = \mu_i(t_i) = \lambda_i t_i$$
- Explanatory variables: $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip-1})^\top$.
$$\log\big(\mu_i(t_i)\big) = \log(\lambda_i) + \log(t_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \log(t_i).$$

We will consider three other types of data we can analyse with a Poisson GLM:

1. Approximating binomial data (topic 4c).
2. Time non-homogeneous Poisson processes (topic 4d).
3. Contingency tables/Multinomial data (topic 4e).

Poisson Approximation to the Binomial

- Suppose: $Y \sim \text{BIN}(m, \pi)$ so that $\mathbb{E}[Y] = m\pi$.
- Set $\mu = m\pi$ and examine pmf of Y in terms of μ:
$$f(y) = \binom{m}{y} \pi^y (1-\pi)^{m-y}$$
$$f(y) = \quad = \frac{(m)(m-1)\cdots(m-y)(m-y-1)\cdots(1)}{(m-y)!y!} \left(\frac{\mu}{m}\right)^y \left(1 - \frac{\mu}{m}\right)^{m-y}$$
$$= \underbrace{\frac{(m)(m-1)\cdots(m-(y-1))}{(m)(m)\cdots(m)}}_{\to 1 \text{ as } m \to \infty} \frac{\mu^y}{y!} \underbrace{\left(1 - \frac{\mu}{m}\right)^m}_{\to e^{-\mu}} \underbrace{\left(1 - \frac{\mu}{m}\right)^{-y}}_{\to 1}.$$
- Recall:
$$\lim_{n \to \infty} \left(1 + \frac{a}{n}\right)^n = e^a.$$
- Therefore, as $m \to \infty$ with $\mu = m\pi$ fixed:
$$f(y) \to \frac{\mu^y e^{-\mu}}{y!} \text{ the pmf of the Poisson.}$$
- So for $Y \sim \text{BIN}(m, \pi)$, as $m \to \infty$, $\pi \to 0$ with $\mathbb{E}[Y] = \mu = m\pi$ fixed we have:
$$Y \sim \text{POI}(\mu = m\pi).$$
- Using a Poisson GLM (with log link):
$$\log(\mu) = \log(\pi) + \log(m) = \boldsymbol{x}^\top \boldsymbol{\beta} + \underbrace{\log(m)}_{\text{offset}}.$$
- Use the Poisson distribution to model Binomial data.
- Use with large population (m large) and low event rate (π).
- Example: Today and Problem 3.1 in course notes.

Non Melanoma Skin Cancer

Schwarz (2015) gives the incidence of non melanoma skin cancer among women in the early 1970s in Minneapolis-St Paul and Dallas-Fort Worth.

City	Age	Count	Pop. Size
msp	15-25	1	172675
msp	25-34	16	123065
msp	35-44	30	96216
msp	45-54	71	92051
msp	55-64	102	72159
msp	65-74	130	54722
msp	75-84	133	32185
msp	85+	40	8328
dfw	15-25	4	181343
dfw	25-34	38	146207
dfw	35-44	119	121374
dfw	45-54	221	111353
dfw	55-64	259	83004
dfw	65-74	310	55932
dfw	75-84	226	29007
dfw	85+	65	7538

Binomial and Poisson Models

- Binomial model:

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

  where $x_{i1} = \mathbb{I}\{\texttt{city=msp}\}$, $x_{i2} = \mathbb{I}\{\text{agegroup } j\}$, $j = 2, 3, \ldots, 8$. β_1 and β_j have $\log(\text{OR})$ interpretations.

- Poisson model:

$$\log(\mu_i) = \alpha_0 + \alpha_1 x_{i1} + \alpha_j x_{ij} + \log(m_i).$$

  α_1 and α_j have $\log(\text{RR})$ interpretations.

Binomial Model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_j x_{ij}, \; j = 2, 3, \ldots, 8.$$

```
melanoma <- read.table("melanoma.txt", header = T)
melanoma$resp = cbind(melanoma$Count, melanoma$Population - melanoma$Count)
fit.binomial = glm(resp ~ factor(City) + factor(Age), family = binomial,
  data = melanoma)
summary(fit.binomial)


Call:
glm(formula = resp ~ factor(City) + factor(Age), family = binomial,
    data = melanoma)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
```

```
-1.49511  -0.47903   0.01814   0.37356   1.23840

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -10.85279    0.44749 -24.253  < 2e-16 ***
factor(City)msp   -0.80692    0.05228 -15.433  < 2e-16 ***
factor(Age)25-34   2.63034    0.46747   5.627 1.84e-08 ***
factor(Age)35-44   3.84801    0.45467   8.463  < 2e-16 ***
factor(Age)45-54   4.59672    0.45104  10.191  < 2e-16 ***
factor(Age)55-64   5.08987    0.45031  11.303  < 2e-16 ***
factor(Age)65-74   5.64998    0.44976  12.562  < 2e-16 ***
factor(Age)75-84   6.06540    0.45035  13.468  < 2e-16 ***
factor(Age)85+     6.18590    0.45782  13.512  < 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: 2794.7794  on 15  degrees of freedom
Residual deviance:    8.0828  on  7  degrees of freedom
AIC: 120.29

Number of Fisher Scoring iterations: 4
```

Poisson Model

$$\log(\mu_i) = \alpha_0 + \alpha_1 x_{i1} + \alpha_j x_{ij} + \log(m_i),\ j = 2, 3, \ldots, 8.$$

```
fit.poisson = glm(Count ~ factor(City) + factor(Age) + offset(log(Population)),
  family = poisson, data = melanoma)
summary(fit.poisson)
```

```

Call:
glm(formula = Count ~ factor(City) + factor(Age) + offset(log(Population)),
    family = poisson, data = melanoma)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5043  -0.4816   0.0169   0.3697   1.2504

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -10.85360    0.44749 -24.255  < 2e-16 ***
factor(City)msp   -0.80428    0.05221 -15.406  < 2e-16 ***
factor(Age)25-34   2.63019    0.46746   5.627 1.84e-08 ***
factor(Age)35-44   3.84735    0.45466   8.462  < 2e-16 ***
factor(Age)45-54   4.59519    0.45103  10.188  < 2e-16 ***
factor(Age)55-64   5.08728    0.45030  11.298  < 2e-16 ***
factor(Age)65-74   5.64541    0.44975  12.552  < 2e-16 ***
factor(Age)75-84   6.05855    0.45032  13.454  < 2e-16 ***
factor(Age)85+     6.17819    0.45774  13.497  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2790.340  on 15  degrees of freedom
Residual deviance:    8.195  on  7  degrees of freedom
AIC: 120.44

Number of Fisher Scoring iterations: 4
```

Example: Non Melanoma Skin Cancer

1. What is the OR and RR for developing non melanoma skin cancer for women in Dallas-Forth Worth versus those in Minneapolis-St Paul, controlling for age?

$$\widehat{\text{OR}} = \mathsf{exp}\{-\hat{\beta}_1\} = \mathsf{exp}\{0.80692\} = 2.2410.$$

$$\widehat{\text{RR}} = \mathsf{exp}\{-\hat{\alpha}_1\} = \mathsf{exp}\{0.80428\} = 2.2351.$$

2. What is the predicted number of skin cancer cases in Dallas-Fort Worth among women age 25-34?

$$\hat{Y}_i = m_i\hat{\pi}_i = (146207)\,\mathsf{expit}(\hat{\beta}_0 + \hat{\beta}_2) = 39.25427.$$

$$\hat{\mu}_i = m_i\hat{\pi}_i = (146207)\,\mathsf{exp}\{\hat{\alpha}_0 + \hat{\alpha}_2\} = 39.22713.$$

- m (population size) is very large and π (probability of getting non melanoma skin cancer) is very small so the Poisson approximation holds.
- Inference from the two models is nearly identical.
- We might prefer the RR interpretation over the OR interpretation.

Topic 4d: Time Non-homogeneous Poisson Processes

Time Non-Homogeneous Poisson Processes

- Now consider

$$\mathbb{E}\big[N(t)\big] = \mu(t) = \lambda(t) \qquad (\text{not} = \lambda t).$$

- The rate is now a *function* of time.
- Lots of possible ways to model the rate $\lambda(t)$.
  - Piecewise constant:
  $$\lambda(t) = b_1\,\mathbb{I}\{0 < t < t_1\} + b_2\,\mathbb{I}\{t_1 \leq t < t_2\} + \cdots .$$
  - Piecewise linear:
  $$\lambda(t) = (m_1 t + b_1)\,\mathbb{I}\{0 < t < t_1\} + (m_2 t + b_2)\,\mathbb{I}\{t_1 \leq t < t_2\} + \cdots .$$
  - Quadratic:
  $$\lambda(t) = at^2 + bt + c.$$
  - Splines, etc.

Example: Rat Tumour Data

Rat Tumour Data

- Here we consider data from a study of the development of mammary tumours in rats reported in Gail et al. (1980).
- This study was a carcinogenicity experiment in which 48 rats were exposed to a carcinogen,
  - 23 were then assigned to a treatment group where the treatment was designed to reduce the development of tumours,
  - 25 were assigned to the control group.
- The rats were carefully examined over 122 days for the development of new tumours (multiple tumours could develop).
- The day (time) of each tumour was recorded.
- Our aim here is to estimate the expected number of tumours in the two groups and make treatment comparisons.

We show the first 5 IDs for each group.

Times to tumours in days$^{\text{(number of tumours detected)}}$

Treatment Group		Control Group	
ID	Days of Tumour Detection	ID	Days of Tumour Detection
1	122	1	$3, 42, 59, 61^{(2)}, 112, 119$
2	—	2	$28, 31, 35, 45, 52, 59^{(2)}, 77, 85, 107, 112$
3	$3, 88$	3	$31, 38, 48, 52, 74, 77, 101^{(2)}, 119$
4	92	4	$11, 114$
5	$70, 74, 85, 92$	5	$35, 45, 74^{(2)}, 77, 80, 85, 90^{(2)}$

Timeline plots for data from Gail et al. (1980)

R Code & Rat Tumour Data Structure

```r
rats <- read.table("rats.dat", header = F)
dimnames(rats)[[2]] <- c("id", "start", "stop", "status", "enum", "trt")
# function to covert data to the structure of one line per interval
# per subject
gd.pw.f <- function(indata) {
  pid <- sort(unique(indata$id))
  data <- matrix(0, nrow = (length(pid) * 4), ncol = 5)
  for (i in 1:length(pid)) {
    tmp <- indata[indata$id == pid[i], ]
    etime <- floor(tmp$stop[tmp$status == 1])
    startpos <- 4 * (i - 1) + 1
    stoppos <- 4 * i
    data[startpos:stoppos, 1] <- rep(pid[i], 4)
    data[startpos:stoppos, 2] <- c(1, 2, 3, 4)
    data[startpos:stoppos, 3] <- c(sum((etime > 0) & (etime <= 30)),
      sum((etime > 30) & (etime <= 60)), sum((etime > 60) & (etime <=
        90)), sum((etime > 90) & (etime <= 122)))
    data[startpos:stoppos, 4] <- c(30, 30, 30, 32)
    data[startpos:stoppos, 5] <- rep(unique(tmp$trt), 4)
  }
  data <- data.frame(data)
  dimnames(data)[[2]] <- c("id", "interval", "count", "len", "trt")
  return(data)
}
rats.pw <- gd.pw.f(rats)
rats.pw[1:20, ]
```

```
rats.pw[1:20, ]

   id interval count len trt
1   1        1     0  30   1
2   1        2     0  30   1
3   1        3     0  30   1
4   1        4     1  32   1
5   2        1     0  30   1
6   2        2     0  30   1
7   2        3     0  30   1
8   2        4     0  32   1
9   3        1     1  30   1
10  3        2     0  30   1
11  3        3     1  30   1
12  3        4     0  32   1
13  4        1     0  30   1
14  4        2     0  30   1
15  4        3     0  30   1
16  4        4     1  32   1
17  5        1     0  30   1
18  5        2     0  30   1
19  5        3     3  30   1
20  5        4     1  32   1
```

- Consider four time intervals.
- One line of data per interval.
- `count` = number events in interval.
- `len` = days spent in interval.
- `trt` = treatment group.

1. Model Control Group Only (`pfitC`)

$$\log(\mu_{ik}) = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}_{\text{interval}} + \texttt{offset}(\log(\texttt{len}_{ik})).$$

- To start, we fit a piecewise constant model for control rats:

$$\log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \log(t_i).$$

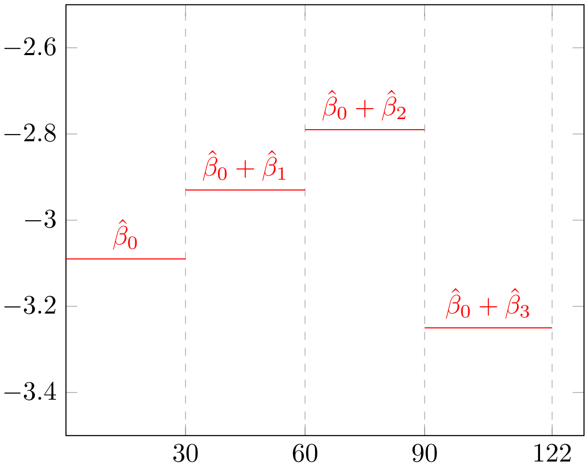
- `interval` is a categorical variable at 4 levels:

$$x_{i1} = \mathbb{I}\{\text{interval } 2\}, \quad x_{i2} = \mathbb{I}\{\text{interval } 3\}, \quad x_{i3} = \mathbb{I}\{\text{interval } 4\}.$$

- Include `offset(log(len))` to account for the fact that different intervals are of different durations.

```
pfitC <- glm(count ~ factor(interval) + offset(log(len)), family = poisson(link = log),
  data = rats.pw, subset = (trt == 0))
summary(pfitC)


Call:
glm(formula = count ~ factor(interval) + offset(log(len)), family = poisson(link = log),
    data = rats.pw, subset = (trt == 0))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9183  -1.5748  -0.2736   0.6262   2.8959

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.0937     0.1715 -18.039   <2e-16 ***
factor(interval)2   0.1625     0.2333   0.697    0.486
factor(interval)3   0.3023     0.2262   1.337    0.181
factor(interval)4  -0.1569     0.2483  -0.632    0.527
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 167.79  on 99  degrees of freedom
Residual deviance: 163.31  on 96  degrees of freedom
AIC: 345.25

Number of Fisher Scoring iterations: 5
```

Plot of log$\big(\lambda(t)\big)$ for `pfitC`



- $\log(\hat\lambda_1) = \hat\beta_0 = -3.09$.
- $\log(\hat\lambda_2) = \hat\beta_0 + \hat\beta_1 = -3.09 + 0.16 = -2.93$.

- $\log(\hat{\lambda}_3) = \hat{\beta}_0 + \hat{\beta}_2 = -3.09 + 0.3 = -2.79.$
- $\log(\hat{\lambda}_4) = \hat{\beta}_0 + \hat{\beta}_3 = -3.09 - 0.16 = -3.25.$

Interpretation of `pfitC`

$$\log(\mu_i) = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}_{\text{interval}} + \log(t_i)$$

- Relative Rate of events in interval 2 versus interval 1:

$$\exp\{\beta_1\} = \frac{\lambda(\text{interval 2})}{\lambda(\text{interval 1})} = \exp\{0.16254\} = 1.176.$$

- Notice none of $\beta_1, \beta_2, \beta_3$ are statistically significant.
- There is a trend of a slightly higher rate in intervals 2 and 3 (versus interval 1) but the event rate does not differ significantly across follow-up time in the control rats.

2. Model Control and Treatment Groups (`pfit`)

- Now, fit a model to both the treatment and control groups.
- $x_{i4} = \mathbb{I}\{\text{treatment group}\}$.
- Assume a piecewise constant baseline rate function.
- Model is now:

$$\log(\mu_i) = \beta_0 + \underbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}_{\text{interval}} + \beta_4 x_{i4} + \texttt{offset}(\log(t_i)).$$

- $\exp\{\beta_1\}$ is now RR of events for interval 2 versus interval 1, for two rats of the same treatment group.

```
pfit <- glm(count ~ factor(interval) + trt + offset(log(len)), family = poisson(link = log),
  data = rats.pw)
summary(pfit)


Call:
glm(formula = count ~ factor(interval) + trt + offset(log(len)),
    family = poisson(link = log), data = rats.pw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8335  -1.1994  -0.3302   0.4701   3.0551

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.08818    0.15079 -20.480  < 2e-16 ***
factor(interval)2  0.17185    0.19590   0.877    0.380
factor(interval)3  0.20634    0.19438   1.062    0.288
factor(interval)4 -0.06454    0.20412  -0.316    0.752
trt               -0.82302    0.15171  -5.425 5.79e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

    Null deviance: 301.37  on 191  degrees of freedom
Residual deviance: 266.32  on 187  degrees of freedom
AIC: 547.75

Number of Fisher Scoring iterations: 5
```

Interpretation of `pfit`

- Relative Rate of events for treatment versus control rats:

$$\mathsf{exp}\{\beta_4\} = \frac{\lambda(\text{treatment})}{\lambda(\text{control})} = \mathsf{exp}\{-0.8230\} = 0.44.$$

- Controlling for interval of follow-up, the rate of tumour development in treated rats in 0.44 times that of control rats.
- That is, treatment looks beneficial.
- Notice that β_4 is statistically significant.
- $\beta_1, \beta_2, \beta_3$ are still not statistically significant.
- Consider do we really need to use a time non-homogeneous model for this data?

3. Time Homogeneous Model (`fit`)

$$\mathsf{log}(\mu_i) = \beta_0 + \beta_4 x_{i4} + \mathsf{log}(t_i).$$

- $\beta_0 =$ log rate of tumour development, per day, control group.
- $\beta_4 =$ log Relative Rate (RR) of tumour development in treated vs control rats.
- This model is nested within the time non-homogeneous model.
- Consider `pfit` model with $\beta_1 = \beta_2 = \beta_3 = 0$.
- We can carry out a likelihood ratio test

```
fit <- glm(count ~ trt + offset(log(len)), family = poisson(link = log),
  data = rats.pw)
summary(fit)


Call:
glm(formula = count ~ trt + offset(log(len)), family = poisson(link = log),
    data = rats.pw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7800  -1.1421  -0.4235   0.4009   3.2673

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -3.00562    0.08138 -36.934  < 2e-16 ***
trt         -0.82302    0.15171  -5.425 5.79e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 301.37  on 191  degrees of freedom
Residual deviance: 269.06  on 190  degrees of freedom
AIC: 544.49

Number of Fisher Scoring iterations: 5
```

Interpretation of `fit`

- Note $\hat{\beta}_4 = -0.8230$ is almost unchanged versus model `fit`.
- Likelihood Ratio/Deviance test of H_0: $\beta_1 = \beta_2 = \beta_3 = 0$:

$$\Delta D = D_0 - D_A = 269.060 - 266.323 \sim \chi^2_{(3)} \text{ under } H_0.$$

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

```
[1] 0.4340077
```

- Do not reject H_0.
- Conclude that the time homogeneous model (model 3) is probably OK in this case.
- However, we retain it for generality and for the following analysis.

4. Time Non-Homogeneous Model with Treatment Interaction (`ifit`)

- Q: Is the treatment effect constant over time?
- Model with interaction:

$$\begin{aligned}\log(\mu_i) = \beta_0 + &\overbrace{\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}}^{\text{interval}} + \overbrace{\beta_4 x_{i4}}^{\text{treatment}} + \\ &+ \underbrace{\beta_5 x_{i1} x_{i4} + \beta_6 x_{i2} x_{i4} + \beta_7 x_{i3} x_{i4}}_{\text{interval}*\text{treatment}} + \log(t_i)\end{aligned}$$

- Model `pfit` (time non-homogeneous, without interaction) is nested within this model (consider `ifit` with $\beta_5 = \beta_6 = \beta_7 = 0$).

```
ifit <- glm(count ~ offset(log(len)) + factor(interval) * trt, family = poisson(link = log),
  data = rats.pw)
summary(ifit)
```

```
Call:
glm(formula = count ~ offset(log(len)) + factor(interval) * trt,
    family = poisson(link = log), data = rats.pw)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9183  -1.2158  -0.3241   0.5125   2.8959

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -3.09371    0.17150 -18.039   <2e-16 ***
factor(interval)2      0.16252    0.23326   0.697   0.4860
factor(interval)3      0.30228    0.22617   1.337   0.1814
factor(interval)4     -0.15691    0.24833  -0.632   0.5275
trt                   -0.80392    0.31755  -2.532   0.0114 *
factor(interval)2:trt  0.03164    0.42972   0.074   0.9413
factor(interval)3:trt -0.37639    0.44663  -0.843   0.3994
factor(interval)4:trt  0.28653    0.43808   0.654   0.5131
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 301.37  on 191  degrees of freedom
Residual deviance: 263.92  on 184  degrees of freedom
AIC: 551.35

Number of Fisher Scoring iterations: 5
```

Interpretation of `ifit`

- Note $\hat{\beta}_4 = -0.8038$ is very similar to `pfit`.
- Likelihood Ratio/Deviance test of H_0: $\beta_5 = \beta_6 = \beta_7 = 0$:

$$\Delta D = D_0 - D_A = 266.323 - 263.917 \sim \chi^2_{(3)} \text{ under } H_0.$$

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

[1] 0.4926145
```

- Do not reject H_0.
- We do not have evidence that the treatment effect varies across the time intervals.

Summary of Rat Tumour Data Analysis

- Looks like a piecewise constant rate function is not necessary.
- The best model (of the ones we examined) is `fit`:

$$\log(\mu_i) = \beta_0 + \beta_4 x_{i4} + \log(t_i).$$

- Interpretation: The relative rate for tumour development in treated versus control rats is:

$$\exp\{\hat{\beta}_4\} = \exp\{-0.822995\} = 0.439.$$

- That is, treatment is beneficial (treated rates get fewer tumours).
- Prediction: Expected number of tumours for a treated rat observed for 70 days?

$$\log(\hat{\mu}) = \hat{\beta}_0 + \hat{\beta}_4 + \log(70) = -3.00562 - 0.82302 + \log(70) = 0.41986.$$

$$\hat{\mu} = \exp\{0.41986\} = 1.5217.$$

Topic 4e: Introduction of Contingency Tables

Analysis of Contingency Tables

- Contingency tables can be formed to display data when all variables are categorical.
- Below is a two-dimensional $I \times J$ contingency table.

$$\begin{array}{cc|ccccccc|c}
 & & \multicolumn{7}{c}{\text{Factor } W} & \\
 & & 1 & 2 & 3 & \cdots & j & \cdots & J & \\ \hline
 & 1 & y_{11} & y_{12} & y_{13} & \cdots & y_{1j} & \cdots & y_{1J} & y_{1\bullet} \\
 & 2 & y_{21} & y_{22} & y_{23} & \cdots & y_{2j} & \cdots & y_{2J} & y_{2\bullet} \\
 & 3 & y_{31} & y_{32} & y_{33} & \cdots & y_{3j} & \cdots & y_{3J} & y_{3\bullet} \\
\text{Factor } V & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\
 & i & y_{i1} & y_{i2} & y_{i3} & \cdots & y_{ij} & \cdots & y_{iJ} & y_{i\bullet} \\
 & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\
 & I & y_{I1} & y_{I2} & y_{I3} & \cdots & y_{Ij} & \cdots & y_{IJ} & y_{I\bullet} \\ \hline
 & & y_{\bullet 1} & y_{\bullet 2} & y_{\bullet 3} & \cdots & y_{\bullet j} & \cdots & y_{\bullet J} & y_{\bullet\bullet}
\end{array}$$

- I = Number of rows; J = Number of columns.
- Row Totals: $y_{i\bullet} = \sum_{j=1}^{J} y_{ij}$.
- Column Totals: $y_{\bullet j} = \sum_{i=1}^{I} y_{ij}$.
- Grand Total: $y_{\bullet\bullet} = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij}$.
- Want to assess the nature/significance of ANY associations between the variables.
- No special response variable — all factors are of equal importance.
- Contingency tables are a cross-classification of units with respect to the factors of interest.
- The observations y_{ij} consist of all the cell counts of the contingency table — these will be our "responses."

Example: 2-way Contingency Table

Breast Self-Examination Contingency Table

- Senie *et al.* (1981) investigated the relationship between age and frequency of breast self-examination in a sample of women.
- Two factors: Age (at 3 levels) and Frequency (at 3 levels).

		Frequency of breast self-examination			
		Monthly	Occasionally	Never	Total
Age	<45	91	90	51	232
	45–59	150	200	155	505
	≥60	109	198	172	479
	Total	350	488	378	1216

- Is than an association between age and exam frequency?

Basic Assumption in Contingency Tables

- Basic Assumption: Each cell count has an independent Poisson distribution with mean μ_{ij} for the (i, j) cell

$$\mathbb{P}(Y_{ij} = y_{ij}) = \frac{\mu_{ij}^{y_{ij}} e^{-\mu_{ij}}}{y_{ij}!}, \; y_{ij} = 0, 1, 2, \ldots .$$

- The joint distribution is

$$\mathbb{P}(Y_{ij} = y_{ij}, i = 1, \ldots, I, j = 1, \ldots, J) = \prod_{i=1}^{I} \prod_{j=1}^{J} \left(\frac{\mu_{ij}^{y_{ij}} e^{-\mu_{ij}}}{y_{ij}!} \right)$$

- We will condition on the relevant fixed totals (row, column, or grand) (possibly fixed by design) to get a multinomial or product multinomial distribution.
- Will show that these can all by analysed using Poisson GLMs.

The Multinomial Distribution

- Assume the total number of units is fixed $Y_{\bullet\bullet} = y_{\bullet\bullet}$ $(= n)$.
- Units are then cross-classified by 2 factors V and W.
- Our assumption of $Y_{ij} \sim \text{POI}(\mu_{ij})$ independently implies

$$Y_{\bullet\bullet} \sim \text{POI}(\mu_{\bullet\bullet}), \text{ where } \mu_{\bullet\bullet} = \sum \sum \mu_{ij}.$$

- To get the joint distribution of the Y_{ij}'s, we must condition on the grand total $Y_{\bullet\bullet} = y_{\bullet\bullet}$ since this is a fixed design:

$$\begin{aligned}
\mathbb{P}(Y_{ij} = y_{ij} \forall i, j \mid Y_{\bullet\bullet} = y_{\bullet\bullet}) &= \frac{\mathbb{P}(Y_{ij} = y_{ij} \forall i, j, Y_{\bullet\bullet} = y_{\bullet\bullet})}{\mathbb{P}(Y_{\bullet\bullet} = y_{\bullet\bullet})} \\
&= \frac{\prod_{i=1}^{I} \prod_{j=1}^{J} \left(\frac{\mu_{ij}^{y_{ij}} \exp\{-\mu_{ij}\}}{y_{ij}!} \right)}{\mu_{\bullet\bullet}^{y_{\bullet\bullet}} \exp\{-\mu_{\bullet\bullet}\} / y_{\bullet\bullet}!} \\
&= \left(\frac{y_{\bullet\bullet}!}{\prod \prod y_{ij}!} \right) \left(\frac{\prod \prod \mu_{ij}^{y_{ij}}}{\mu_{\bullet\bullet}^{y_{\bullet\bullet}}} \right) \underbrace{\left(\frac{\exp\{-\sum \sum \mu_{ij}\}}{\exp\{-\mu_{\bullet\bullet}\}} \right)}_{= 1 \text{ since } \mu_{\bullet\bullet} = \sum \sum \mu_{ij}} \\
&= \left(\frac{y_{\bullet\bullet}!}{\prod \prod y_{ij}!} \right) \qquad \underbrace{\prod_{i=1}^{I} \prod_{j=1}^{J} \left(\frac{\mu_{ij}}{\mu_{\bullet\bullet}} \right)^{y_{ij}}}_{\text{since } \mu_{\bullet\bullet}^{y_{\bullet\bullet}} = \mu_{\bullet\bullet}^{\sum \sum y_{ij}} = \prod \prod \mu_{\bullet\bullet}^{y_{ij}}}
\end{aligned}$$

- Recall the standard Multinomial distribution:

$$f(x_1, \dots, x_k; n, \pi_1, \dots, \pi_k) = \mathbb{P}(X_1 = x_1, \dots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} \pi_1^{x_1} \cdots \pi_k^{x_k},$$

  where $\sum \pi_i = 1$ and $\sum x_i = n$.

- The pmf on the previous slide is a multinomial distribution with

$$\pi_{ij} = \mu_{ij}/\mu_{\bullet\bullet} = \mathbb{P}(\text{level } i \text{ of factor } V \text{ and level } j \text{ of factor } W).$$

- Note that $\sum \sum \pi_{ij} = 1$

$$\mathbb{P}(Y_{ij} = y_{ij} \forall i, j \mid Y_{\bullet\bullet} = y_{\bullet\bullet}) = \left(\frac{y_{\bullet\bullet}!}{\prod \prod y_{ij}!} \right) \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{y_{ij}}.$$

Multinomial Likelihood

$$\mathbb{P}(Y_{ij} = y_{ij} \forall i, j \mid Y_{\bullet\bullet} = y_{\bullet\bullet}) = \left(\frac{y_{\bullet\bullet}!}{\prod \prod y_{ij}!} \right) \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{y_{ij}}.$$

- $\boldsymbol{\pi} = (\pi_{11}, \dots, \pi_{IJ})^\top$ be the parameter vector.
- The likelihood and log-likelihood are given by:

$$L(\boldsymbol{\pi}) = \prod_i \prod_j \pi_{ij}^{y_{ij}}, \text{ where } \sum \sum \pi_{ij} = 1$$

$$\ell(\boldsymbol{\pi}) = \sum_i \sum_j y_{ij} \log(\pi_{ij}).$$

Testing for Independence in a 2-way Table

- Thinking back to the contingency table, we might be interested in testing the hypothesis that the two methods of classification are independent:

$$H_0\colon \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j} \ \forall i, j$$

$$H_A\colon \pi_{ij} \neq \pi_{i\bullet}\pi_{\bullet j} \text{ for some } i, j,$$

  where $\pi_{i\bullet} = \sum_{j=1}^{J} \pi_{ij}$ and $\pi_{\bullet j} = \sum_{i=1}^{I} \pi_{ij}$.

- Consider the log-likelihood under H_0 (independence):

$$\begin{aligned} \ell(\boldsymbol{\pi}) &= \sum_i \sum_j y_{ij} \log(\pi_{i\bullet}\pi_{\bullet j}) \\ &= \sum_i \sum_j y_{ij} \big(\log(\pi_{i\bullet}) + \log(\pi_{\bullet j}) \big) \\ &= \sum_i y_{i\bullet} \log(\pi_{i\bullet}) + \sum_j y_{\bullet j} \log(\pi_{\bullet j}). \end{aligned}$$

- The parameters are constrained by $\sum \pi_{i\bullet} = 1$ and $\sum \pi_{\bullet j} = 1$.
- The MLEs of $\pi_{i\bullet}$ and $\pi_{\bullet j}$ under H_0 are:

$$\hat{\pi}_{i\bullet} = \frac{y_{i\bullet}}{y_{\bullet\bullet}}, \qquad \hat{\pi}_{\bullet j} = \frac{y_{\bullet j}}{y_{\bullet\bullet}}.$$

- And the log-likelihood evaluated at the MLE is:

$$\ell(\hat{\boldsymbol{\pi}}) = \sum_i \sum_j y_{ij} \log\left(\frac{y_{i\bullet} y_{\bullet j}}{y_{\bullet\bullet}^2}\right).$$

- Next consider working under H_A (unconstrained).
- The unconstrained MLEs are: $\tilde{\pi}_{ij} = \frac{y_{ij}}{y_{\bullet\bullet}}$.
- And the log-likelihood evaluated at the unconstrained MLE is:

$$\ell(\tilde{\boldsymbol{\pi}}) = \sum_i \sum_j y_{ij} \log\left(\frac{y_{ij}}{y_{\bullet\bullet}}\right).$$

- To test for independence we could use a Likelihood Ratio/Deviance test for the multinomial:

$$\begin{aligned} D &= 2\big(\ell(\tilde{\boldsymbol{\pi}}) - \ell(\hat{\boldsymbol{\pi}})\big) \\ &= 2\sum_i \sum_j y_{ij} \log\left(\frac{y_{ij}}{y_{\bullet\bullet}} \Big/ \frac{y_{i\bullet} y_{\bullet j}}{y_{\bullet\bullet}^2}\right) \\ &= 2\sum_i \sum_j y_{ij} \log\left(\frac{y_{ij}}{y_{i\bullet} y_{\bullet j}/y_{\bullet\bullet}}\right) \\ &= 2\sum_i \sum_j O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right). \end{aligned}$$

- Note this has the usual form of a Deviance Statistic with

$$O_{ij} = y_{ij} \quad \text{and} \quad E_{ij} = y_{\bullet\bullet}\hat{\pi}_{ij} \text{ under } H_0.$$

- We know $D \sim \chi^2_{(n-p)}$, but what are the degrees of freedom here?

$$\begin{aligned} n - p &= (\#\text{ parameters saturated}) - (\#\text{ parameters unsaturated}) \\ &= (IJ - 1) - \big((I-1) + (J-1)\big) \\ &= IJ - I - J + 1 \\ &= (I-1)(J-1). \end{aligned}$$

Example: Breast Self-Examination Data ($\tilde{\mu}_{ij}$ vs $\hat{\mu}_{ij}$)

- Observed Data: $y_{ij} = \tilde{\mu}_{ij} = \tilde{\pi}_{ij} y_{\bullet\bullet}$:

		Frequency of breast self-examination			
		Monthly	Occasionally	Never	Total
Age	<45	91	90	51	232
	45–59	150	200	155	505
	≥60	109	198	172	479
	Total	350	488	378	1216

- Expected Data under H_0: $\hat{\mu}_{ij} = \hat{\pi}_{ij} y_{\bullet\bullet} = y_{i\bullet} y_{\bullet j}/y_{\bullet\bullet}$:

		Frequency of breast self-examination			
		Monthly	Occasionally	Never	Total
Age	<45	66.78	93.11	72.12	232
	45–59	145.35	202.66	156.98	505
	≥60	137.87	192.23	148.90	479
	Total	350	488	378	1216

Example: Breast Self-Examination Data: ($\tilde{\pi}_{ij}$ vs $\hat{\pi}_{ij}$)

- Unconstrained MLEs: $\tilde{\pi}_{ij} = y_{ij}/y_{\bullet\bullet}$ (as percentages):

		Frequency of breast self-examination			
		Monthly	Occasionally	Never	Row %
Age	<45	7.48	7.40	4.19	19.07
	45–59	12.34	16.45	12.75	41.54
	≥60	8.96	16.28	14.14	39.38
	Column %	28.78	40.13	31.08	100

- Constrained MLEs under H_0: $\hat{\pi}_{ij} = \hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j} = y_{i\bullet}y_{\bullet j}/y_{\bullet\bullet}^2$:

		Frequency of breast self-examination			
		Monthly	Occasionally	Never	Row %
Age	<45	5.49	7.66	5.93	19.08
	45–59	11.95	16.67	12.91	41.53
	≥60	11.34	15.81	12.25	39.40
	Column %	28.78	40.14	31.09	100

Example: Breast Self-Examination Data (Testing Independence)

- Use the Likelihood Ratio/Deviance test derived for the Multinomial Distribution

$$D = 2\sum_i \sum_j y_{ij} \log\left(\frac{y_{ij}}{y_{i\bullet}y_{\bullet j}/y_{\bullet\bullet}}\right) = 25.19226.$$

- Compare to a $\chi^2_{(4)}$ distribution:

$$p = \mathbb{P}\left(\chi^2_{(4)} > 25.19226\right) < 0.001.$$

- So we reject the null hypothesis that age and frequency of breast self-examination are independent.

The Product Multinomial Distribution

- Previously, we assumed the grand total $Y_{\bullet\bullet} = y_{\bullet\bullet}$ was fixed.
- Now assume that the row totals $Y_{i\bullet} = y_{i\bullet}$ are fixed.
  - Choose a sample of fixed size from populations $i = 1, \ldots, I$ and then classify the units with response to Factor W.

- Our assumption of $Y_{ij} \sim \text{POI}(\mu_{ij})$ independently implies

$$Y_{i\bullet} \sim \text{POI}(\mu_{i\bullet}),\ \text{where } \mu_{i\bullet} = \sum_j \mu_{ij}.$$

- To get the joint distribution of the Y_{ij}'s we now condition on the row totals $Y_{i\bullet} = y_{i\bullet}$, $i = 1, \dots, I$

$$\mathbb{P}(Y_{ij} = y_{ij} \forall i, j \mid Y_{i\bullet} = y_{i\bullet} \forall i) = \frac{\mathbb{P}(Y_{ij} = y_{ij} \forall i, j, Y_{i\bullet} = y_{i\bullet} \forall i)}{\mathbb{P}(Y_{i\bullet} = y_{i\bullet} \forall i)}.$$

Example: Another Breast Self-Examination Study

- Imagine this time the investigators decided study a fixed number of women of each age group.
- The (hypothetical) 2-way contingency table is now:

Breast Self-Examination Contingency Table (Hypothetical)

		Frequency of breast self-examination			
		Monthly	Occasionally	Never	Total
	<45	78	78	44	200
Age	45–59	178	238	184	600
	≥60	91	165	144	400
	Total	347	481	372	1200

- We need to take this method of sampling into account in the analysis.

$$\begin{aligned}\mathbb{P}(Y_{ij} = y_{ij} \forall i, j \mid Y_{i\bullet} = y_{i\bullet} \forall i) &= \left(\prod_i \prod_j \left(\frac{\mu_{ij}^{y_{ij}} \exp\{-\mu_{ij}\}}{y_{ij}!}\right)\right) \Big/ \left(\prod_i \frac{\mu_{i\bullet}^{y_{i\bullet}} \exp\{-\mu_{i\bullet}\}}{y_{i\bullet}!}\right) \\ &= \left(\frac{\prod_i y_{i\bullet}!}{\prod\prod y_{ij}!}\right)\left(\frac{\prod\prod \mu_{ij}^{y_{ij}}}{\prod_i \mu_{i\bullet}^{y_{i\bullet}}}\right) \underbrace{\left(\frac{\exp\{-\sum\sum \mu_{ij}\}}{\exp\{-\sum_i \mu_i\}}\right)}_{= 1 \text{ since } \mu_{ij} = \sum_i \mu_{i\bullet} = \mu_{\bullet\bullet}} \\ &= \prod_{i=1}^{I} \underbrace{\left(\frac{y_{i\bullet}!}{\prod_j y_{ij}!} \prod_{j=1}^{J} \left(\frac{\mu_{ij}}{\mu_{i\bullet}}\right)^{y_{ij}}\right)}_{\text{Multinomial pmf for row } i}.\end{aligned}$$

- This is the product multinomial distribution with $\pi_{ij} = \mu_{ij}/\mu_{i\bullet}$.
- Here, π_{ij} = probability of being level j given population level i.
- Note that $\sum_j \pi_{ij} = 1$ for all i.

Product Multinomial Likelihood

$$\mathbb{P}(Y_{ij} = y_{ij} \forall i, j \mid Y_{i\bullet} = y_{i\bullet} \forall i) = \prod_{i=1}^{I}\left(\frac{y_{i\bullet}!}{\prod_j y_{ij}!} \prod_{j=1}^{J} \pi_{ij}^{y_{ij}}\right).$$

- Again, let $\boldsymbol{\pi} = (\pi_{11}, \dots, \pi_{IJ})^\top$ be the parameter vector.
- Note the π_{ij} have different interpretations here versus the multinomial case.
- The log-likelihood is given by:

$$\ell(\boldsymbol{\pi}) = \sum_i \sum_j y_{ij} \log(\pi_{ij}),\ \text{where } \sum_j \pi_{ij} = 1\ \forall i.$$

Testing for Independence with the Product Multinomial

- In this case we might be interested in testing where the probability of being at factor level j is the same across all stratum/populations $i = 1, \dots, I$

$$H_0\colon \pi_{1j} = \pi_{2j} = \cdots = \pi_{Ij} = \pi_j,\ j = 1, 2, \dots, J,$$
$$H_A\colon \text{at least one } \pi_{ij} \neq \pi_{i'j}.$$

- The log likelihood under H_0 (independence) is

$$\ell(\boldsymbol{\pi}) = \sum_i \sum_j y_{ij} \log(\pi_j) = \sum_j y_{\bullet j} \log(\pi_j).$$

- The parameters are constrained by $\sum_j \pi_j = 1$.
- The MLEs under H_0 are

$$\hat{\pi}_{ij} = \hat{\pi}_j = \frac{y_{\bullet j}}{y_{\bullet\bullet}}.$$

- Under H_A (unconstrained) the MLEs are

$$\tilde{\pi}_{ij} = \frac{y_{ij}}{y_{i\bullet}}.$$

- The Likelihood Ratio/Deviance test statistic is:

$$\begin{aligned} D &= 2\big(\ell(\tilde{\boldsymbol{\pi}}) - \ell(\hat{\boldsymbol{\pi}})\big) \\ &= 2\sum_i \sum_j y_{ij} \log\left(\frac{y_{ij}}{y_{i\bullet}} \Big/ \frac{y_{\bullet j}}{y_{\bullet\bullet}}\right) \\ &= 2\sum_i \sum_j y_{ij} \log\left(\frac{y_{ij}}{y_{i\bullet} y_{\bullet j}/y_{\bullet\bullet}}\right). \end{aligned}$$

- Which is identical to the Deviance statistic for testing independence under a multinomial distribution.
- Here, $D \sim \chi^2_{(I-1)(J-1)}$ since

$$n - p = I(J-1) - (J-1) = IJ - I - J + 1 = (I-1)(J-1).$$

Example: Another Breast Self-Examination Study

- Observed Data: y_{ij}:

Frequency of breast self-examination

		Monthly	Occasionally	Never	Total
Age	<45	78	78	44	200
	45–59	178	238	184	600
	≥60	91	165	144	400
	Total	347	481	372	1200

- Expected Data under H_0: $\hat{\mu}_{ij} = y_{i\bullet}\hat{\pi}_j = y_{i\bullet} y_{\bullet j}/y_{\bullet\bullet}$

Frequency of breast self-examination

		Monthly	Occasionally	Never	Total
Age	<45	57.83	80.17	62.00	200
	45–59	173.50	240.50	186.00	600
	≥60	115.67	160.33	124.00	400
	Total	347	481	372	1200

- Unconstrained MLEs: $\tilde{\pi}_{ij} = y_{ij}/y_{i\bullet}$ (as percentages):

		Frequency of breast self-examination			
		Monthly	Occasionally	Never	Total
Age	<45	39.00	39.00	22.00	100
	45–59	29.67	39.67	30.67	100
	≥60	22.75	41.25	36.00	100

- Constrained MLEs: $\hat{\pi}_{ij} = \hat{\pi}_j = y_{\bullet j}/y_{\bullet\bullet}$ (as percentages):

		Frequency of breast self-examination			
		Monthly	Occasionally	Never	Total
Age	<45	28.92	40.08	31.00	100
	45–59	28.92	40.08	31.00	100
	≥60	28.92	40.08	31.00	100

Example: Another Breast Self-Examination Study (Testing Independence)

- Use the Likelihood Ratio/Deviance test derived for the Multinomial Distribution

$$D = 2 \sum_i \sum_j y_{ij} \log\left(\frac{y_{ij}}{y_{i\bullet} y_{\bullet j}/y_{\bullet\bullet}} \right) = 21.25615.$$

- Compare to a $\chi^2_{(4)}$ distribution:

$$p = \mathbb{P}\Big(\chi^2_{(4)} > 21.25615\Big) < 0.001.$$

- So we reject the null hypothesis that age and frequency of breast self-examination are independent.

Summary

- Today we considered simple 2-way contingency tables.
- With the basic Poisson assumption for the cell counts, depending on the type of sampling used, we can test for independence using:
  1. Multinomial distribution (condition on $y_{\bullet\bullet}$).
  2. Product multinomial (condition on $y_{i\bullet}$, $i = 1, 2, \ldots, I$).
- Both yield the same Likelihood Ratio/Deviance test statistic.
- Interestingly we can also use log-linear models to assess these independence hypotheses (next week).
- Easily generalizable to 3-way (and more) contingency tables.

Topic 4f: Log Linear Models for Two-way Tables

Likelihood Based Analysis of 2-way Contingency Tables

		Factor W							
		1	2	3	$\cdots$	j	$\cdots$	J	
	1	y_{11}	y_{12}	y_{13}	$\cdots$	y_{1j}	$\cdots$	y_{1J}	$y_{1\bullet}$
	2	y_{21}	y_{22}	y_{23}	$\cdots$	y_{2j}	$\cdots$	y_{2J}	$y_{2\bullet}$
	3	y_{31}	y_{32}	y_{33}	$\cdots$	y_{3j}	$\cdots$	y_{3J}	$y_{3\bullet}$
Factor V	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$.
	i	y_{i1}	y_{i2}	y_{i3}	$\cdots$	y_{ij}	$\cdots$	y_{iJ}	$y_{i\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	I	y_{I1}	y_{I2}	y_{I3}	$\cdots$	y_{Ij}	$\cdots$	y_{IJ}	$y_{I\bullet}$
		$y_{\bullet 1}$	$y_{\bullet 2}$	$y_{\bullet 3}$	$\cdots$	$y_{\bullet j}$	$\cdots$	$y_{\bullet J}$	$y_{\bullet\bullet}$

- Previously: Previously: Derived Likelihood Ratio/Deviance tests for testing for independence between Factor V and Factor W.

- Basic Assumption: $Y_{ij} \sim \text{POI}(\mu_{ij})$, $\forall i, j$.

- When we condition on the Grand Total the joint distribution becomes Multinomial, and we want to test:

$$H_0\colon \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j} \ \forall i, j$$

$$H_A\colon \pi_{ij} \neq \pi_{i\bullet}\pi_{\bullet j} \text{ for some } i, j.$$

- When we condition on the Row Totals the joint distribution becomes Product Multinomial, and we want to test:

$$H_0\colon \pi_{1j} = \pi_{2j} = \cdots = \pi_{Ij} = \pi_j, \ j = 1, 2, \ldots, J,$$

$$H_A\colon \text{at least one } \pi_{ij} \neq \pi_{i'j}.$$

- In either case, the Likelihood Ratio/Deviance Test statistic is:

$$D = 2\sum_i \sum_j y_{ij} \log\left(\frac{y_{ij}}{y_{i\bullet}y_{\bullet j}/y_{\bullet\bullet}}\right) \sim \chi^2_{(I-1)(J-1)} \text{ under } H_0.$$

Log Linear Models for 2-way Contingency Tables

- Basic Assumption: $Y_{ij} \sim \text{POI}(\mu_{ij})$, $\forall i, j$.

- Explanatory Variables: Factor V and W:

$$\begin{aligned} x_1 &= \mathbb{I}\{\text{Factor } V \text{ at level } 2\}, & x_I &= \mathbb{I}\{\text{Factor } W \text{ at level } 2\}, \\ x_2 &= \mathbb{I}\{\text{Factor } V \text{ at level } 3\}, & x_{I+1} &= \mathbb{I}\{\text{Factor } W \text{ at level } 3\}, \\ &\vdots & &\vdots \\ x_{I-1} &= \mathbb{I}\{\text{Factor } V \text{ at level I}\}, & x_{I+J-2} &= \mathbb{I}\{\text{Factor } W \text{ at level J}\}. \end{aligned}$$

- The main effects log-linear model would be:

$$\begin{aligned} \log(\mu_\ell) = \beta_0 + &\overbrace{\beta_1 x_{1\ell} + \beta_2 x_{2\ell} + \cdots + \beta_{I-1}x_{I-1\ell}}^{\text{Factor } V} + \\ &+ \underbrace{\beta_I x_{I\ell} + \beta_{I+1}x_{I+1\ell} + \cdots + \beta_{I+J-2}x_{I+J-2\ell}}_{\text{Factor } W} \qquad \ell = 1, \ldots, IJ. \end{aligned}$$

- Note: # parameters $= 1 + (I-1) + (J-1) = I + J - 1$.
- The $\boldsymbol{x}^\top \boldsymbol{\beta}$ is quite cumbersome when I and J are large.
- Consider the following expression for the model:

$$\mathsf{log}(\mu_{ij}) = u + u_i^V + u_j^W,\ i = 1, \dots, I,\ j = 1, \dots, J,$$

  where $u_1^V + u_1^W = 0$.
- Note: # parameters $= 1 + (I-1) + (J-1) = I + J - 1$.
- This notation suppresses the binary x variables.
- The relationship between the β and u is as follows:

$$u = \beta_0, \quad \begin{array}{ll} u_2^V = \beta_1, & u_2^W = \beta_I, \\ u_3^V = \beta_2, & u_3^W = \beta_{I+1}, \\ \vdots & \vdots \\ u_I^V = \beta_{I-1}, & u_J^W = \beta_{I+J-2}. \end{array}$$

- Testing independence in a 2-way table:

$$H_0\colon \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}\ \forall i, j$$
$$H_A\colon \pi_{ij} \neq \pi_{i\bullet}\pi_{\bullet j} \text{ for some } i, j.$$

- The corresponding log-linear models are:

$$H_0\colon\ \mathsf{log}(\mu_{ij}) = u + u_i^V + u_j^W$$
$$H_A\colon\ \mathsf{log}(\mu_{ij}) = u + u_i^V + u_j^W + u_{ij}^{VW}.$$

- Using corner-point constraints we require:

$$u_1^V = 0, \qquad u_1^W = 0, \qquad u_{1j}^{VW} = 0\ \forall j, \qquad u_{i1}^{VW}\ \forall i.$$

- The interaction model has $1 + (I-1) + (J-1) + (I-1)(J-1) = IJ$ parameters.
- Wait: We're using a Poisson model to fit data/test hypotheses from a Multinomial distribution?
- Examine the log-likelihood from the Poisson:

$$\ell(\boldsymbol{\mu}) = \sum_i \sum_j \big[y_{ij}\,\mathsf{log}(\mu_{ij}) - \mu_{ij} - \mathsf{log}(y_{ij}!)\big].$$

- Substitute in the log linear model $H_0\colon\ \mathsf{log}(\mu_{ij}) = u + u_i^V + u_j^W$:

$$\begin{aligned}
\ell(\boldsymbol{u}) &= \sum\sum\Big(y_{ij}(u + u_i^V + u_j^W) - \mathsf{exp}\{u + u_i^V + u_j^W\} - \mathsf{log}(y_{ij}!)\Big) \\
\frac{\partial \ell}{\partial u} &= \sum\sum\big(y_{ij} - \mathsf{exp}\{u + u_i^V + u_j^W\}\big) \\
&= \sum\sum\big(y_{ij} - \mu_{ij}\big) \\
&= y_{\bullet\bullet} - \mu_{\bullet\bullet} \qquad (\text{set} = 0) \implies \hat{\mu}_{\bullet\bullet} = y_{\bullet\bullet}. \\
\frac{\partial \ell}{\partial u_{i^\star}^V} &= \sum\sum\big(y_{i^\star j} - \mathsf{exp}\{u + u_{i^\star}^V + u_j^W\}\big) \\
&= y_{i^\star\bullet} - \mu_{i^\star\bullet} \qquad (\text{set} = 0) \implies \hat{\mu}_{i\bullet} = y_{i\bullet}\ \forall i. \\
\frac{\partial \ell}{\partial u_{j^\star}^W} &= \sum\sum\big(y_{ij^\star} - \mathsf{exp}\{u + u_i^V + u_{j^\star}^W\}\big) \\
&= y_{\bullet j^\star} - \mu_{\bullet j^\star} \qquad (\text{set} = 0) \implies \hat{\mu}_{\bullet j} = y_{\bullet j}\ \forall j.
\end{aligned}$$

- So the main effects log linear model reproduces the row, column and grand totals.
- If we do the same with the saturated model

$$H_A\colon\ \log(\mu_{ij}) = u + u_i^V + u_j^W + u_{ij}^{VW},$$

  we find it provides a perfect fit to the data: $\tilde{\mu}_{ij} = y_{ij}$ for all i, j.

- Recall the Deviance Test for the Poisson Distribution

$$\begin{aligned} D &= 2\big(\ell(\tilde{\boldsymbol{\mu}}) - \ell(\hat{\boldsymbol{\mu}})\big) \\ &= 2\sum\sum\Big(\big(y_{ij} - \log(\tilde{\mu}_{ij}) - \tilde{\mu}_{ij} - \log(y_{ij}!)\big) - \big(y_{ij} - \log(\hat{\mu}_{ij}) - \hat{\mu}_{ij} - \log(y_{ij}!)\big)\Big) \\ &= 2\sum\sum\left(y_{ij}\log\left(\frac{\tilde{\mu}_{ij}}{\hat{\mu}_{ij}}\right) - (\tilde{\mu}_{ij} - \hat{\mu}_{ij})\right) \\ &= 2\sum\sum y_{ij}\log\left(\frac{y_{ij}}{y_{i\bullet}y_{\bullet j}/y_{\bullet\bullet}}\right), \end{aligned}$$

  since

$$\begin{aligned} \hat{\mu}_{ij} &= y_{\bullet\bullet}\hat{\pi}_{i\bullet}\hat{\pi}_{\bullet j} \\ &= y_{\bullet\bullet}\left(\frac{\hat{\mu}_{i\bullet}}{\hat{\mu}_{\bullet\bullet}}\right)\left(\frac{\hat{\mu}_{\bullet j}}{\hat{\mu}_{\bullet\bullet}}\right) \\ &= y_{i\bullet}y_{\bullet j}/y_{\bullet\bullet}, \end{aligned}$$

  and

$$\begin{aligned} \sum\sum\tilde{\mu}_{ij} &= \sum\sum u_{ij} = y_{\bullet\bullet}, \\ \sum\sum\hat{\mu}_{ij} &= \hat{\mu}_{\bullet\bullet} = y_{\bullet\bullet}. \end{aligned}$$

$$D = 2\sum\sum y_{ij}\log\left(\frac{y_{ij}}{y_{i\bullet}y_{\bullet j}/y_{\bullet\bullet}}\right).$$

- We know $D \sim \chi^2_{(n-p)}$ under H_0. Here,

$$n - p = (I - J) - \big(1 + (I-1) + (J-1)\big) = (I-1)(J-1).$$

- Same as the Likelihood Ratio/Deviance Test statistic from the Multinomial and Product Multinomial last section.
- Use the Deviance Test from fitting Poisson models to conduct hypotheses tests for data from 2-way contingency tables!

Example: A Melanoma Study

- A cross-sectional study was conducted in which 400 patients with malignant melanoma were classified according to two factors: the site of the tumour and the histological type.

Melanoma Study Data

Tumour Type	Head and Neck	Trunk	Extremities	Total
Hutchinson's freckle	22	2	10	34
Superficial Spreading	16	54	115	185
Nodular	19	33	73	125
Indeterminate	11	17	28	56
Total	68	106	226	400

- Here we wish to investigate whether the different types of tumour appear equally likely in the different sites.
- That is, we are assessing whether there is an association between histological type and tumour site.
- We wish to test for independence:

$$H_0\colon \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}\ \forall i, j$$

$$H_A\colon \pi_{ij} \neq \pi_{i\bullet}\pi_{\bullet j} \text{ for some } i, j.$$

- Under H_0: $\mu_{ij} = \mathbb{E}[Y_{ij}] = y_{\bullet\bullet}\pi_{i\bullet}\pi_{\bullet j}$, meaning we will have to fit the row and column totals to allow estimation of $\pi_{i\bullet}$ and $\pi_{\bullet j}$.
- Thus, our log linear model under the null hypothesis is

$$\log(\mu_{ij}) = u + u_i^V + u_j^W,\ i = 1, 2, 3, 4,\ j = 1, 2, 3$$

- V corresponds to tumour type variable (i indicating the level).
- W corresponds to tumour site variable (j indicating the level).
- If the model fits the data well, then there's no evidence against the assumption that tumour type and site are independent.
- If the model does not fit the data well, then some tumour types appear more frequently in certain locations.

R Dataset

Melanoma Data Set

```
   type locat   y
1     1     1  22
2     1     2   2
3     1     3  10
4     2     1  16
5     2     2  54
6     2     3 115
7     3     1  19
8     3     2  33
9     3     3  73
10    4     1  11
11    4     2  17
12    4     3  28
```

R Code

```
derm.dat = read.table("derm.dat", header = T)
derm.dat$typef = factor(derm.dat$type)
derm.dat$sitef = factor(derm.dat$locat)
derm.dat
# fitting the model with both main effects
model1 = glm(y ~ typef + sitef, family = poisson, data = derm.dat)
summary(model1)
# creating deviance residuals for diagnostic plots
derm.dat$fitted.values = model1$fitted.values
derm.dat$rdeviance = residuals.glm(model1, type = "deviance")
derm.dat
# fitting the model with only the 'histological type' main effect
model2 = glm(y ~ typef, family = poisson, data = derm.dat)
1 - pchisq(model2$deviance - model1$deviance, model2$df.residual - model1$df.residual)
# fitting the model with only the 'site' main effect
model3 = glm(y ~ sitef, family = poisson, data = derm.dat)
1 - pchisq(model3$deviance - model1$deviance, model3$df.residual - model1$df.residual)
```

- One line per cell in the contingency table.
- $IJ = 12$ observations.
- `type` is tumour type (4 levels).
- `locat` is tumour location (3 levels).
- `y` is the count in the contingency table.

R output for Model 1: `type + site`

```
model1 = glm(y ~ typef + sitef, family = poisson, data = derm.dat)
summary(model1)


Call:
glm(formula = y ~ typef + sitef, family = poisson, data = derm.dat)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.0453  -1.0741   0.1297   0.5857   5.1354  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)   1.7544     0.2040   8.600  < 2e-16 ***
typef2        1.6940     0.1866   9.079  < 2e-16 ***
typef3        1.3020     0.1934   6.731 1.68e-11 ***
typef4        0.4990     0.2174   2.295  0.02173 *  
sitef2        0.4439     0.1554   2.857  0.00427 ** 
sitef3        1.2010     0.1383   8.683  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

```
    Null deviance: 295.203  on 11  degrees of freedom
Residual deviance:  51.795  on  6  degrees of freedom
AIC: 122.91

Number of Fisher Scoring iterations: 5
```

- Recall we are testing for independence

$$H_0\colon \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}\ \forall i, j$$

$$H_A\colon \pi_{ij} \neq \pi_{i\bullet}\pi_{\bullet j} \text{ for some } i, j.$$

- The Deviance test statistic $\chi^2_{(12-6)}$ under H_0.
- Here $D = 51.795$ which corresponds to a p-value of

$$p = \mathbb{P}\left(\chi^2_{(6)} > 51.795\right) < 0.001.$$

  Therefore, we reject the null hypothesis of independence.

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

[1] 2.050453e-09
```

- Examine the fitted values and residuals.

```
derm.dat

   type locat   y typef sitef fitted.values   rdeviance
1     1     1  22     1     1         5.780  5.13537787
2     1     2   2     1     2         9.010 -2.82829426
3     1     3  10     1     3        19.210 -2.31583297
4     2     1  16     2     1        31.450 -3.04533605
5     2     2  54     2     2        49.025  0.69899703
6     2     3 115     2     3       104.525  1.00813975
7     3     1  19     3     1        21.250 -0.49711084
8     3     2  33     3     2        33.125 -0.02173229
9     3     3  73     3     3        70.625  0.28104581
10    4     1  11     4     1         9.520  0.46798432
11    4     2  17     4     2        14.840  0.54787007
12    4     3  28     4     3        31.640 -0.66016102
```

- Can verify that the row and column totals are fit exactly.
- For example, sum the first three observations corresponding to the total number of Hutchinson freckle cases, and sum the corresponding fitted values.
- We conclude that the model does not provide a very good fit to the data since there are some rather large deviance residuals corresponding to the first two rows of the table.
- Therefore, our hypothesis that tumour type and site are independent does not seem plausible.
- Specifically, based on the fitted values and residuals we see that Hutchinson's freckle occurs more often on the head and neck than we would expect under the independence assumption, and less often on the trunk and extremities.

- Furthermore, superficial spreading melanoma occurs less often on the head and neck than we would expect.
- Can we use a smaller model?

R output for Model 2: `type`

```
model2 = glm(y ~ typef, family = poisson, data = derm.dat)
summary(model2)


Call:
glm(formula = y ~ typef, family = poisson, data = derm.dat)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-6.9398  -2.2986  -0.7009   2.2079   6.0553  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)   2.4277     0.1715  14.156  < 2e-16 ***
typef2        1.6940     0.1866   9.079  < 2e-16 ***
typef3        1.3020     0.1934   6.731 1.68e-11 ***
typef4        0.4990     0.2174   2.295   0.0217 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 295.2  on 11  degrees of freedom
Residual deviance: 150.1  on  8  degrees of freedom
AIC: 217.21

Number of Fisher Scoring iterations: 5
```

- Model 2: $\log(\mu_{ij}) = u + u_i^V$ for $i = 1, 2, 3, 4$ and $j = 1, 2, 3$ with $u_1^V = 0$.
- Now we are testing

$$H_0\colon \pi_{ij} = \pi_{i\bullet}/J \;\forall i,$$

$$H_A\colon \exists i \text{ such that } \pi_{ij} \neq \pi_{i\bullet}/J$$

- The Deviance test statistic $\Delta D = D_0 - D_A \sim \chi^2_{(J-1)}$ under H_0.
- Here $\Delta D = 150.1 - 51.795$ which corresponds to a p-value of

$$p = \mathbb{P}\left(\chi^2_{(2)} > 98.305\right) < 0.001$$

  Therefore, we reject the null hypothesis that all location occur with equal frequency.

```
1 - pchisq(model2$deviance - model1$deviance, model2$df.residual - model1$df.residual)

[1] 0
```

R output for Model 3: `site`

```
model3 = glm(y ~ sitef, family = poisson, data = derm.dat)
summary(model3)


Call:
glm(formula = y ~ sitef, family = poisson, data = derm.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-7.6398  -2.5337   0.1155   1.4367   6.8161  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)   2.8332     0.1213  23.363  < 2e-16 ***
sitef2        0.4439     0.1554   2.857  0.00427 ** 
sitef3        1.2010     0.1383   8.683  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 295.2  on 11  degrees of freedom
Residual deviance: 196.9  on  9  degrees of freedom
AIC: 262.01

Number of Fisher Scoring iterations: 5

1 - pchisq(model3$deviance - model1$deviance, model3$df.residual - model1$df.residual)

[1] 0
```

- Therefore, we reject the null hypothesis that different tumour types occur equally often when controlled for sites.

Summary: A Melanoma Study

- Row Percentages:

Tumour Type	Head and Neck	Trunk	Extremities	Total
Hutchinson's freckle	64.7	5.9	29.4	100
Superficial Spreading	8.6	29.2	62.2	100
Nodular	15.2	26.4	58.4	100
Indeterminate	19.6	30.4	50.0	100
Total	17.0	26.5	56.5	100

- Column Percentages:

Tumour Type	Head and Neck	Trunk	Extremities	Total
Hutchinson's freckle	32.4	1.9	4.4	8.5
Superficial Spreading	23.5	50.9	50.9	46.25
Nodular	27.9	31.1	32.3	31.25
Indeterminate	16.2	16.0	12.4	14.00
Total	100	100	100	100

- We rejected the null hypothesis that tumour type and site are independent.
- In addition, further investigation indicates that the different tumour types do not occur equally often, and melanoma does not occur equally often at the different sites of the body.
- See Course Notes for example of fitting model 1 with ANOVA constraints ($\sum_i u_i^V = 0$ and $\sum_j u_j^W = 0$) instead of corner-point constraints ($u_1^V = u_1^W = 0$).
- Coefficient estimates and correlation matrix change.
- Deviance, deviance residuals, and fitted values are unchanged.

Revisit the example from last section

Breast Self-Examination Contingency Table

		Frequency of breast self-examination			
		Monthly	Occasionally	Never	Total
Age	<45	91	90	51	232
	45–59	150	200	155	505
	≥60	109	198	172	479
	Total	350	488	378	1216

- Last class we rejected the null hypothesis that Age and Frequency of breast self-examination are independent:

$$D = 2\sum_i \sum_j y_{ij} \log\left(\frac{y_{ij}}{y_{i\bullet}y_{\bullet j}/y_{\bullet\bullet}}\right) = 25.19226.$$

$$p = \mathbb{P}\left(\chi^2_{(4)} > 25.19226\right) < 0.001.$$

R Code

```r
# Breast Self-Examination Contingency Table Analysis
y = c(91, 90, 51, 150, 200, 155, 109, 198, 172)
Age = as.factor(c(1, 1, 1, 2, 2, 2, 3, 3, 3))
Freq = as.factor(c(1, 2, 3, 1, 2, 3, 1, 2, 3))
Exam = data.frame(Age, Freq, y)
# Fit main effects log linear model
model1 = glm(y ~ Age + Freq, family = poisson)
summary(model1)
1 - pchisq(model1$deviance, model1$df.residual)
# Examine fitted values and deviance residuals
Exam$fv = model1$fitted.values
Exam$rd = residuals.glm(model1, type = "deviance")
Exam
```

R Output for Main Effects Model

```
# Fit main effects log linear model
model1 = glm(y ~ Age + Freq, family = poisson)
summary(model1)


Call:
glm(formula = y ~ Age + Freq, family = poisson)

Deviance Residuals: 
      1        2        3        4        5        6        7        8  
 2.8078  -0.3236  -2.6259   0.3834  -0.1876  -0.1585  -2.5530   0.4141  
      9  
 1.8471  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  4.20135    0.07966  52.743  < 2e-16 ***
Age2         0.77782    0.07931   9.807  < 2e-16 ***
Age3         0.72496    0.07999   9.063  < 2e-16 ***
Freq2        0.33238    0.07005   4.745 2.08e-06 ***
Freq3        0.07696    0.07418   1.037      0.3    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 173.944  on 8  degrees of freedom
Residual deviance:  25.192  on 4  degrees of freedom
AIC: 95.168

Number of Fisher Scoring iterations: 4

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

[1] 4.602407e-05

Exam$fv = model1$fitted.values
Exam$rd = residuals.glm(model1, type = "deviance")
Exam

  Age Freq   y        fv         rd
1   1    1  91  66.77632  2.8077823
2   1    2  90  93.10526 -0.3236329
3   1    3  51  72.11842 -2.6259260
4   2    1 150 145.35362  0.3833650
5   2    2 200 202.66447 -0.1875765
6   2    3 155 156.98191 -0.1585172
7   3    1 109 137.87007 -2.5530416
8   3    2 198 192.23026  0.4140893
9   3    3 172 148.89967  1.8470579
```

- Reject H_0 that main effects model is adequate, that is, we reject H_0 that age and frequency are independent.

- Same Deviance Test statistic as what we calculated based on the multinomial distribution.
- Compare the above fitted values to the expected data under H_0 (last lecture).

Topic 4g: A Generalization to Three-way Tables

Log Linear Models for 2-Way Tables

- Subjects are classified with respect to tow factor variables denoted V and W with I and J levels respectively.
- We are interested in testing for independence

$$H_0\colon \pi_{ij} = \pi_{i\bullet}\pi_{\bullet j}.$$

- The corresponding log linear model is:

$$\log(\mu_{ij}) = u + u_i^V + u_j^W$$

  with $u_1^V = u_1^W = 0$ (corner-point constraints).
- Number of model parameters $= 1 + (I-1) + (J-1) = I + J - 1$.
- Deviance test statistic:

$$D = 2\sum_i \sum_j y_{ij} \log\left(\frac{y_{ij}}{y_{i\bullet}y_{\bullet j}/y_{\bullet\bullet}}\right) \sim \chi^2_{(I-1)(J-1)} \text{ under } H_0.$$

- Residual df $= IJ - I - J - 1 = (I-1)(J-1)$.

3-way Contingency Tables

- Consider the general problem in which subjects are classified with respect to three factor variables denoted V, W, and Z with I, J, and K levels respectively.
- As with two-way tables, we initially assume

$$Y_{ijk} \sim \text{POI}(\mu_{ijk}),$$

  $i = 1, 2, \ldots, I$, $j = 1, 2, \ldots, J$, $k = 1, 2, \ldots, K$.
- As before, if $Y_{\bullet\bullet\bullet} = y_{\bullet\bullet\bullet}$ is fixed by design (as it usually would be), we condition on this to give the multinomial distribution:

$$\mathbb{P}(Y_{ijk} = y_{ijk} \forall (i,j,k) \mid Y_{\bullet\bullet\bullet} = y_{\bullet\bullet\bullet}) = \frac{y_{\bullet\bullet\bullet}!}{\prod_i \prod_j \prod_k y_{ijk}!} \prod_i \prod_j \prod_k \pi_{ijk}^{y_{ijk}}.$$

- $\pi_{ijk} = \mu_{ijk}/\mu_{\bullet\bullet\bullet} = \mathbb{P}(V = i, W = j, Z = k)$ are the parameters of interest ($\sum\sum\sum \pi_{ijk} = 1$).
- In the case of 2-way contingency tables we discussed the connection between log-linear models and questions about the association between the two factors.
- Main effects accommodated non-uniform distributions of the row and column totals, and the interaction terms allowed for association between the two factors of interest.
- In terms of an association, it was either present or absent.
- As we will see in what follows, with 3-way tables (contingency tables involving 3 factor variables) the nature of the associations present may be somewhat more complicated.

1. Mutual Independence.
2. Joint Independence.
3. Conditional Independence.
4. Homogeneous Association.

- The saturated model for a 3-way contingency table is:

$$\log(\mu_{ijk}) = u + u_i^V + u_j^W + u_k^Z + u_{ij}^{VW} + u_{ik}^{VZ} + u_{jk}^{WZ} + u_{ijk}^{VWZ}$$

  with corner-point constraints:

  - $u_1^V = u_1^W = u_1^Z = 0$.
  - $u_{1j}^{VW} = u_{i1}^{VW} = u_{1k}^{VZ} = u_{i1}^{VZ} = u_{1k}^{WZ} + u_{j1}^{WZ} = 0$ for all i, j, k.
  - $u_{1jk}^{VWZ} = u_{i1k}^{VWZ} = u_{ij1}^{VWZ}$ for all i, j, k.

- Shorthand notation: This model is denoted (VWZ) where we list the highest order terms involving each of the factors.

- It provides a perfect fit to the data

$$\tilde{\pi}_{ijk} = y_{ijk}/y_{\bullet\bullet\bullet},$$
$$\tilde{\mu}_{ijk} = y_{\bullet\bullet\bullet}\tilde{\pi}_{ijk} = y_{ijk}.$$

- To investigate the relationship between factors V, W, and Z we will consider simpler log-linear models.

1. Mutual Independence H_0: $\pi_{ijk} = \pi_{i\bullet\bullet}\pi_{\bullet j\bullet}\pi_{\bullet\bullet k}$

- H_0: All 3 factors V, W, and Z are independent of each other.

$$\pi_{ijk} = \pi_{i\bullet\bullet}\pi_{\bullet j\bullet}\pi_{\bullet\bullet k},$$

$$\mathbb{P}(V = i, W = j, Z = k) = \mathbb{P}(V = i)\,\mathbb{P}(W = j)\,\mathbb{P}(Z = k).$$

- The corresponding log-linear model is (V, W, Z)

$$\log(\mu)_{ijk} = u + u_i^V + u_j^W + u_k^Z$$

  with $u_1^V = u_1^W = u_1^Z = 0$ (with corner-point constraints).

- This model will fit the marginal totals exactly.

- The fitted values are:

$$\hat{\mu}_{ijk} = y_{\bullet\bullet\bullet}\hat{\pi}_{ijk} = y_{\bullet\bullet\bullet}\hat{\pi}_{i\bullet k}\hat{\pi}_{\bullet j\bullet} = y_{\bullet\bullet\bullet}\left(\frac{y_{i\bullet k}}{y_{\bullet\bullet\bullet}}\right)\left(\frac{y_{\bullet j\bullet}}{y_{\bullet\bullet\bullet}}\right)$$

- Number of model parameters $= 1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1)$.

- Residual df $= IJK - (IK + J - 1)$.

- Similar to ordinary 2-way independence between W and a new variable with IK levels of V and Z combined.

- The joint distribution of (V, Z) is the same at any level of W.

- For 3-way tables there are 3 possible joint independence hypotheses and models: (V, WZ), (VZ, W), and (VW, Z).

2. Joint Independence H_0: $\pi_{ijk} = \pi_{i\bullet k}\pi_{\bullet j\bullet}$

- H_0: Factor W is jointly independent of V and Z

$$\pi_{ijk} = \pi_{i\bullet k}\pi_{\bullet j\bullet},$$

$$\mathbb{P}(V=i, W=j, Z=k) = \mathbb{P}(V=i, Z=k)\,\mathbb{P}(W=j).$$

- The nature of the association between V and Z does not depend on the level of W.
- The corresponding log-linear model is (VZ, W)

$$\log(\mu_{ijk}) = u + u_i^V + u_j^W + u_k^Z + u_{ik}^{VZ}$$

  with $u_1^V = u_1^W = u_1^Z$, $u_{1k}^{VZ} = u_{i1}^{VZ} = 0$ for all i, k.
- This model will fit the marginal totals and VZ combination totals ($y_{i\bullet k}$) exactly.
- The fitted values are:

$$\hat{\mu}_{ijk} = y_{\bullet\bullet\bullet}\hat{\pi}_{ijk} = y_{\bullet\bullet\bullet}\hat{\pi}_{i\bullet k}\hat{\pi}_{\bullet j\bullet} = y_{\bullet\bullet\bullet}\left(\frac{y_{i\bullet k}}{y_{\bullet\bullet\bullet}}\right)\left(\frac{y_{\bullet j\bullet}}{y_{\bullet\bullet\bullet}}\right).$$

- Number of model parameters $= 1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1)$.
- Residual df $= IJK - (IK + J - 1)$.
- Similar to ordinary 2-way independence between W and a new variable with IK levels of V and Z combined.
- The joint distribution of (V, Z) is the same at any level of W.
- For 3-way tables there are 3 possible joint independence hypotheses and models: (V, WZ), (VZ, W), and (VW, Z).

3. Conditional Independence H_0: $\pi_{ij|k} = \pi_{i\bullet|k}\pi_{\bullet j|k}$

- Conditional probability notation: $\pi_{ij|k} = \pi_{ijk}/\pi_{\bullet\bullet k}$

$$\pi_{ijk} = \pi_{ij|k}\pi_{\bullet\bullet k},$$

$$\mathbb{P}(V=i, W=j, Z=k) = \mathbb{P}(V=i, W=j \mid Z=k)\,\mathbb{P}(Z=k).$$

- H_0: Factors V and W are conditionally independent given Z.

$$\pi_{ijk} = \pi_{ij|k}\pi_{\bullet\bullet k} = \pi_{i\bullet|k}\pi_{\bullet j|k},$$

$$\mathbb{P}(V=i, W=j, Z=k) = \mathbb{P}(V=i \mid Z=k)\,\mathbb{P}(W=j \mid Z=k)\,\mathbb{P}(Z=k).$$

- That is, the association between V and W can be *fully explained* by Z.
- The corresponding log-linear model is (VZ, WZ)

$$\log(\mu)_{ijk} = u_i^V + u_j^W + u_k^Z + u_k^Z + u_{ik}^{VZ} + u_{jk}^{WZ}.$$

- This model will fit all marginal totals and VWZ and WZ combination totals ($y_{i\bullet k}$ and $y_{\bullet jk}$ exactly).
- The fitted values are:

$$\hat{\mu}_{ijk} = y_{\bullet\bullet\bullet}\hat{\pi}_{ijk} = y_{\bullet\bullet\bullet}\frac{\hat{\pi}_{i\bullet k}\hat{\pi}_{\bullet jk}}{\hat{\pi}_{\bullet\bullet k}} = \frac{y_{i\bullet k}y_{\bullet jk}}{y_{\bullet\bullet k}}.$$

- Number of model parameters $= 1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1) + (J-1)(K-1)$.
- Residual df $= IJK - (IK + JK - K)$.
- Similar to ordinary 2-way independence between V and W at each level of Z.
- That is, make K 2-way tables $(I \times J)$ and test independence of each table.
- For 3-way tables there are 3 possible conditional independence hypotheses and models: (VZ, WZ), (VW, VZ), and (VW, WZ).

4. Homogeneous Association

- The remaining log-linear model is (VW, VZ, WZ)

$$\log(\mu_{ijk}) = u + u_i^V + u_j^W + u_k^Z + u_{ij}^{VW} + u_{ik}^{VZ} + u_{jk}^{WZ}.$$

- Let's examine the model at $k^\star$ an arbitrary fixed level of factor Z:

$$\begin{aligned}\log(\mu)_{ijk^\star} &= u + u_i^V + u_j^W + u_{k^\star}^Z + u_{ij}^{VW} + u_{ik^\star}^{VZ} + u_{jk^\star}^{WZ} \\ &= \left(u + u_{k^\star}^Z\right) + \left(u_i^V + u_{ik^\star}^{VZ}\right) + \left(u_j^W + u_{jk^\star}^{WZ}\right) + u_{ij}^{VW} \\ &= u^\star + u_i^{\star V} + u_j^{\star W} + u_{ij}^{VW}.\end{aligned}$$

  - This is a saturated model for the 2-way table of V and W at $Z = k^\star$.
  - V and W are not independent at level $Z = k^\star$.
  - However, at a different level $Z = k^\dagger$, the parameter u_{ij}^{VW} representing the association between V and W does not change.
- Homogeneous Association: There is a relationship between all pairs of factors, but the nature of the association is the same (i.e., homogeneous) for all levels of the third factor.
- The fitted values are not given by simple, intuitive formulas.
- Number of model parameters $= 1+(I-1)+(J-1)+(K-1)+(I-1)(J-1)+(I-1)(K-1)+(J-1)(K-1)$.
- Residual df $= (I-1)(J-1)(K-1)$.
- For 3-way tables there is only one homogeneous association hypothesis and model (UW, VZ, WZ).
- The relationship implied by this model is also sometimes referred to as All Pairs Conditionally Independent.

Testing Nested Models for 3-way Contingency Tables

These are called hierarchical log-linear models:

Type of Independence	Null Hypothesis	Log Linear Model
None	—	(VWZ)
Homogeneous Association	3x Conditional Independence H_0	(VW, VZ, WZ)
Conditional Independence	$\pi_{ij\|k} = \pi_{i\bullet k}\pi_{\bullet j\|k}$	(VZ, WZ), (VW, WZ), (VW, VZ)
Joint Independence	$\pi_{ijk} = \pi_{i\bullet k}\pi_{\bullet j\bullet}$	(VZ, W), (V, WZ), (VW, Z)
Mutual Independence	$\pi_{ijk} = \pi_{i\bullet\bullet}\pi_{\bullet j\bullet}\pi_{\bullet\bullet k}$	(V, W, Z)

Goodness of Fit Statistics for Log Linear Models

- The fit of a log linear model can be judged based on the deviance assuming an underlying Poisson distribution for the cell counts.
- We know from before that the deviance statistic has the form

$$D = 2\sum_i \sum_j \sum_k O_{ijk} \log\left(\frac{O_{ijk}}{E_{ijk}}\right).$$

- $D \sim \chi^2_{(IJK)-q}$ under H_0 where q is the number of parameters in the model under H_0.
- For nested models:

$$\Delta D = D_0 - D_A \sim \chi^2_{p-q}.$$

Application 1: General Social Survey

2008 US General Social Survey ($2 \times 5 \times 7$)

Gender (G)	Highest Degree (D)	Political Party Affiliation (P)						
		1	2	3	4	5	6	7
Males	< High school	32	20	18	29	11	12	9
	< High school	67	85	63	68	48	65	44
	Junior college	12	14	6	9	13	17	6
	Bachelor	23	21	29	20	19	32	20
	Graduate	16	9	12	13	7	14	13
Females	< High school	31	25	16	58	8	8	16
	High school	118	98	69	88	30	82	54
	Junior college	20	16	13	13	7	16	7
	Bachelor	33	23	28	11	16	44	23
	Graduate	38	20	8	13	3	13	9

- Note that there is no obvious response variable.
- Since we are interested in the association among all three variables, we consider methods based on log-linear models.
- Let G denote gender, D denote highest degree obtained, and P denote political party affiliation.
- We know the log linear model

$$\log(\mu_{ijk}) = u + u_i^G + u_j^D + u_k^P + u_{ij}^{GD} + u_{ik}^{GP} + u_{jk}^{DP} + u_{ijk}^{GDP}$$

  will provide a perfect fit to the data (since it is saturated).

- We seek to find a simpler model which describes the data well.
- In other words, we are looking for a simpler representation of the relationship between the gender, highest degree, and political party affiliation.

R Code

```
## Input the data for the 5 x 7 x 2 contingency table
freq = c(32, 67, 12, 23, 16, 20, 85, 14, 21, 9, 18, 63, 6, 29, 12, 29,
  68, 9, 20, 13, 11, 48, 13, 19, 7, 12, 65, 17, 32, 14, 9, 44, 6, 20,
  13, 31, 118, 20, 33, 38, 25, 98, 16, 23, 20, 16, 69, 13, 28, 8, 58,
  88, 13, 11, 13, 8, 30, 7, 16, 3, 8, 82, 16, 44, 13, 16, 54, 7, 23,
  9)
names = list(D = c("LT HSc", "HSc", "JunCol", "Bachelor", "Graduate"),
  P = c("1", "2", "3", "4", "5", "6", "7"), G = c("male", "female"))
party.3D = array(freq, c(5, 7, 2), dimnames = names)
## Flattened contingency table
library(plyr)
party = count(as.table(party.3D))
party = party[, 1:4]
names(party) = c("D", "P", "G", "Y")
# Fit the saturated model
model1 = glm(Y ~ G * D * P, family = poisson, data = party)
model1$df.residual
model1$deviance
# Fit the homogeneous association model
model2 <- glm(Y ~ G * D + G * P + D * P, family = poisson, data = party)
model2$df.residual
model2$deviance
1 - pchisq(model2$deviance - model1$deviance, model2$df.residual - model1$df.residual)
# Fit the three conditional independence models
model3 <- glm(Y ~ G * D + G * P, family = poisson, data = party)
model3$df.residual
model3$deviance
1 - pchisq(model3$deviance - model2$deviance, model3$df.residual - model2$df.residual)
model4 <- glm(Y ~ G * D + D * P, family = poisson, data = party)
model4$df.residual
model4$deviance
1 - pchisq(model4$deviance - model2$deviance, model4$df.residual - model2$df.residual)
model5 <- glm(Y ~ G * P + D * P, family = poisson, data = party)
model5$df.residual
model5$deviance
1 - pchisq(model5$deviance - model2$deviance, model5$df.residual - model2$df.residual)
# Fit the two joint independence models nested within model5
model6 <- glm(Y ~ G + D * P, family = poisson, data = party)
model6$df.residual
model6$deviance
1 - pchisq(model6$deviance - model5$deviance, model6$df.residual - model5$df.residual)
model7 <- glm(Y ~ G * P + D, family = poisson, data = party)
model7$df.residual
model7$deviance
1 - pchisq(model7$deviance - model5$deviance, model7$df.residual - model5$df.residual)
```

R Output: Models 1 (GDP) and 2 (GD, GP, DP)

```
# Fit the saturated model
model1 = glm(Y ~ G * D * P, family = poisson, data = party)
model1$df.residual
```

```
[1] 0

model1$deviance

[1] -9.547918e-15

# Fit the homogeneous association model
model2 <- glm(Y ~ G * D + G * P + D * P, family = poisson, data = party)
model2$df.residual

[1] 24

model2$deviance

[1] 28.81808

1 - pchisq(model2$deviance - model1$deviance, model2$df.residual - model1$df.residual)

[1] 0.2270527
```

- H_0: Homogeneous association model (2) is adequate

$$H_0\colon u_{ijk}^{GDP} = 0 \ \forall i, j, k \text{ versus } H_A\colon \exists i, j, k \text{ s.t. } u_{ijk}^{GDP} \neq 0.$$

$$\Delta D = D_0 - D_A = 28.818 - 0 \sim \chi^2_{(24)} \text{ under } H_0.$$

$$p = \mathbb{P}\left(\chi^2_{(24)} > 28.818\right) = 0.227.$$

- Do not reject H_0 that the fit of model 2 is adequate, as compared to model 1.

R Output: Models 3 (GD, GP), 4 (GD, DP), 5 (GP, DP)

```
model3 <- glm(Y ~ G * D + G * P, family = poisson, data = party)
model3$df.residual

[1] 48

model3$deviance

[1] 130.3407

1 - pchisq(model3$deviance - model2$deviance, model3$df.residual - model2$df.residual)

[1] 1.650369e-11

model4 <- glm(Y ~ G * D + D * P, family = poisson, data = party)
model4$df.residual

[1] 30

model4$deviance

[1] 52.76878

1 - pchisq(model4$deviance - model2$deviance, model4$df.residual - model2$df.residual)
```

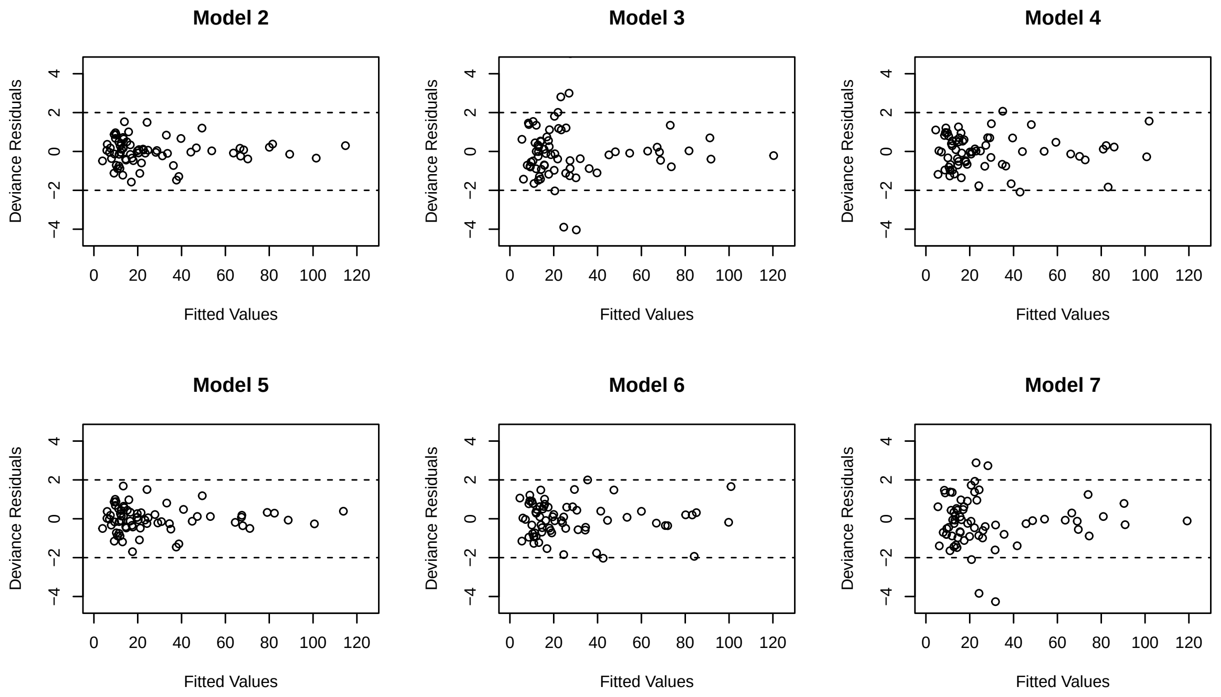
```
[1] 0.0005332749

model5 <- glm(Y ~ G * P + D * P, family = poisson, data = party)
model5$df.residual

[1] 28

model5$deviance

[1] 29.3232

1 - pchisq(model5$deviance - model2$deviance, model5$df.residual - model2$df.residual)

[1] 0.9730008
```

- Reject H_0 that the fit of models 3 and 4 are adequate, as compared to model 2.
- Do no reject H_0 that the fit of model 5 is adequate, as compared to model 2.

R Output: Models 6 (G, DP) and 7 (D, GP)

```
# Fit the two joint independence models nested within model5
model6 <- glm(Y ~ G + D * P, family = poisson, data = party)
model6$df.residual

[1] 34

model6$deviance

[1] 53.84259

1 - pchisq(model6$deviance - model5$deviance, model6$df.residual - model5$df.residual)

[1] 0.0004189688

model7 <- glm(Y ~ G * P + D, family = poisson, data = party)
model7$df.residual

[1] 52

model7$deviance

[1] 131.4145

1 - pchisq(model7$deviance - model5$deviance, model7$df.residual - model5$df.residual)

[1] 1.318701e-11
```

- Reject H_0 that the fit of models 6 and 7 are adequate, as compared to model 5. That is, we can conclude that model 5 is the “best” model.



Summary of Fitted Models

The following analysis of deviance table summarizes our findings.

Model	Form	Residual Deviance	Residual d.f.	p-value
1	(GDP)	0	0	NA
2	(GD, GP, DP)	28.82	24	0.228 (vs 1)
3	(GD, GP)	130.34	48	0.000 (vs 2)
4	(GD, DP)	52.77	30	0.001 (vs 2)
5	(GP, DP)	29.32	28	0.973 (vs 2)
6	(G, DP)	53.84	34	0.000 (vs 5)
7	(D, GP)	131.41	52	0.000 (vs 5)

- Conclude that Model 5 (GP, DP) is most appropriate.
- Conditional Independence: The responders educational level (D) is conditionally independent of his/her gender (G), given his/her party affiliation (P).
- We will return to this analysis in the next topic to discuss interpretation of the regression parameters.