Generalized Linear Models and their Applications

STAT 431/STAT 831[1]

Fall 2021 (1219)[2]

Cameron Roopnarine[3] Leilei Zeng[4]

2nd February 2022

[1] STAT 431 ≡ STAT 831

[2] Online Course

[3] LaTeXer

[4] Instructor

Contents

WEEK 1
8th to 10th September

Topic 1a: Review of Linear Regression

Review of Linear Regression (Stat 331/371)

The Model Fitting Process

0. Exploratory Data Analysis.
1. Model Specification — Select a probability distribution for the response variable and an equation linking the response to the explanatory variables.
2. Estimation of 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.

See Dunn & Smyth Chapters 2 & 3 or your Stat 331 notes for a thorough review.

Example: Dobson's Birthweight Data

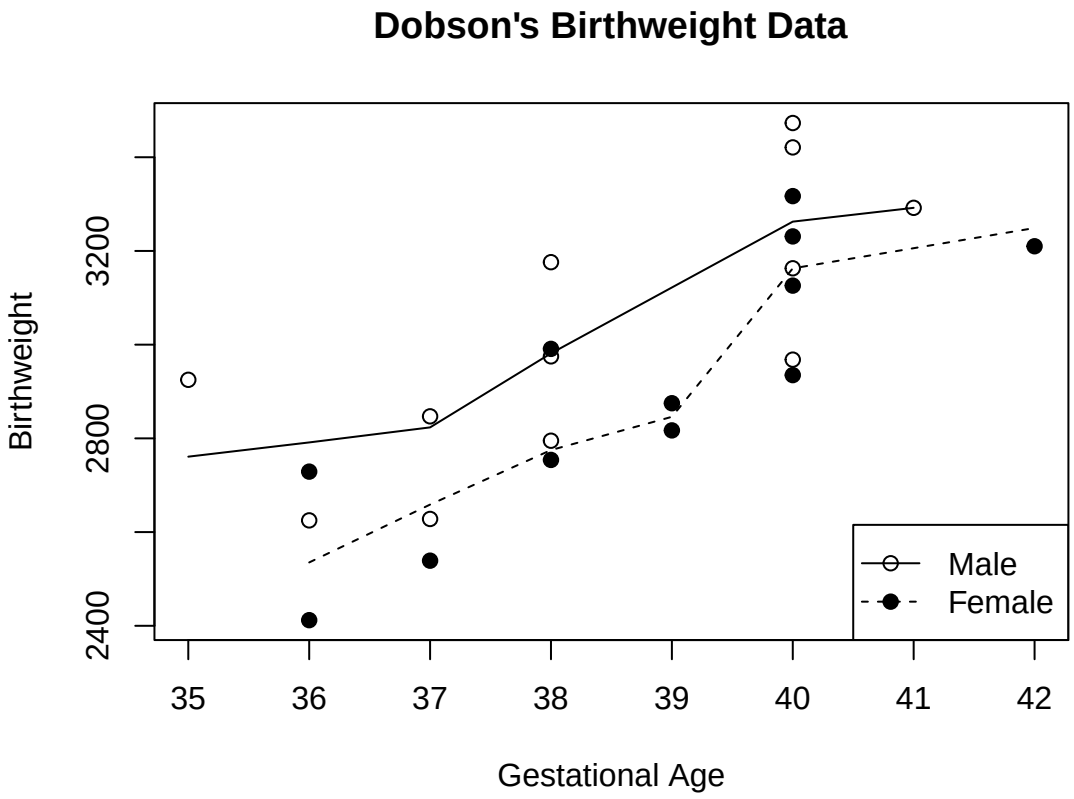
Dobson's Birthweight Data

For $n = 24$ babies, we have observed:

- $Y_i =$ birthweight for baby i (in grams).
- $x_{i1} =$ sex of baby i ($= 0$ female, $= 1$ male).
- $x_{i2} =$ gestational age (in weeks) of baby i.

We wish to model the relationship between the explanatory variables and the birthweight.

0. Exploratory Data Analysis



1. Model Specification

Notation

For each subject $i = 1, 2, \ldots, n$, we have:

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

Specification for Multiple Linear Regression

- Y_i are independent $\mathcal{N}(\mu_i, \sigma^2)$ random variables.
- Regression equation links the response to the explanatory variables:

$$\mathbb{E}[Y_i] = \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

- Putting these together, our regression model is:

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

- Alternatively, we can write our linear regression model 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/Model Fitting

Least Squares

We wish to minimize the expression:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\big)^2$$

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

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\big)$$

$$\frac{\partial S}{\partial \beta_j} = -2 \sum_{i=1}^{n} x_{ij} \big(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\big)$$

Maximum Likelihood Estimation

The likelihood function for $\boldsymbol{\beta}$ is:

$$\mathcal{L}(\boldsymbol{\beta}; \boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \big(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}\big)^2 \right\}$$

The log-likelihood function for $\boldsymbol{\beta}$ is therefore:

$$\begin{aligned}\ell(\boldsymbol{\beta};\boldsymbol{y}) &= \sum_{i=1}^{n}\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2\right)\\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}\right)^2\end{aligned}$$

We find the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ by maximizing $\ell(\boldsymbol{\beta};\boldsymbol{y})$ treating σ^2 as fixed.

- Regression estimates: For Linear Regression, the LSE and MLE of $\boldsymbol{\beta}$ are the same:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \qquad \text{provided } (\boldsymbol{X}^\top\boldsymbol{X}) \text{ is full rank}$$

- Fitted values: $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ or $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$.
- Residuals: $\hat{r}_i = (y_i - \hat{y}_i)$.
- Variance estimates:
  - An unbiased estimate of σ^2 is: $\hat{\sigma}^2 = \dfrac{1}{n-(p+1)}\displaystyle\sum_{i=1}^{n}\hat{r}_i^2$.
  - An estimate of the variance of $\hat{\boldsymbol{\beta}}$ is: $\widehat{\mathrm{V}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$.

Example: Dobson's Birthweight Data

Dobson's Birthweight Data

For $n = 24$ babies, we have observed:

- Y_i = birthweight for baby i (in grams).
- x_{i1} = sex of baby i ($= 0$ female, $= 1$ male).
- x_{i2} = gestational age (in weeks) of baby i.

We wish to model the relationship between the explanatory variables and the birthweight.

- Assume: $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_i, \sigma^2)$.
- Consider a multiple linear regression model:

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

Example: R Code

```r
# Input Dobson's Birthweight Data
age <- c(40, 38, 40, 35, 36, 37, 41, 40, 37, 38, 40, 38, 40,
  36, 40, 38, 42, 39, 40, 37, 36, 38, 39, 40)
birthw <- c(2968, 2795, 3163, 2925, 2625, 2847, 3292, 3473, 2628,
  3176, 3421, 2975, 3317, 2729, 2935, 2754, 3210, 2817, 3126,
  2539, 2412, 2991, 2875, 3231)
sex <- as.factor(c(rep("M", 12), rep("F", 12)))
```

```r
# Exploratory Data Analysis
plot(age, birthw, pch = 1 + 18 * as.numeric(sex == "F"), ylab = "Birthweight",
  xlab = "Gestational Age", main = "Dobson's Birthweight Data")
lines(lowess(age[sex == "M"], birthw[sex == "M"]))
lines(lowess(age[sex == "F"], birthw[sex == "F"]), lty = 2)
legend("bottomright", legend = c("Male", "Female"), pch = c(1,
  19), lty = c(1, 2))
# Model 1: Main Effects Linear Regression Model
m1 <- lm(birthw ~ sex + age)
summary(m1)
plot(age, birthw, pch = 1 + 18 * as.numeric(sex == "F"), ylab = "Birthweight",
  xlab = "Gestational's Birthweight Data", main = "Fitted Regression Lines (m1)")
abline(m1$coeff[1] + m1$coeff[2], m1$coeff[3])
abline(m1$coeff[1], m1$coeff[3], lty = 2)
# Residual Plots
plot(m1$fitted.values, rstandard(m1), main = "Residuals vs Fitted Values",
  ylim = c(-2.5, 2.5), ylab = "Standardized Residuals", xlab = "Fitted Values")
abline(h = 0)
abline(h = 1.96, lty = 3)
abline(h = -1.96, lty = 3)
lines(lowess(m1$fitted.values, rstandard(m1)), col = "red")
qqnorm(rstandard(m1))
abline(0, 1)
```

Example: R Output

```r
summary(m1)


Call:
lm(formula = birthw ~ sex + age)

Residuals:
    Min      1Q  Median      3Q     Max
-257.49 -125.28  -58.44  169.00  303.98

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1773.32     794.59  -2.232   0.0367 *
sexM          163.04      72.81   2.239   0.0361 *
age           120.89      20.46   5.908 7.28e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 177.1 on 21 degrees of freedom
Multiple R-squared:   0.64,Adjusted R-squared:  0.6057
F-statistic: 18.67 on 2 and 21 DF,  p-value: 2.194e-05
```

Interpretation of Regression Parameters

The main effect multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \qquad \text{where } \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$$

So the expected value of the response is:

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

- To interpret β_0, set $x_{i1} = x_{i2} = 0$:

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

  $\beta_0 =$ Expected birthweight of female baby ($x_{i1} = 0$) born a gestational age zero ($x_{i2} = 0$).

- To interpret β_1 consider the difference in the model with $x_{i1} = 1$ versus $x_{i1} = 0$ as seen in Table 1.

x_{i1}	x_{i2}	$\mathbb{E}[Y_i]$
1	x_2	$\beta_0 + \beta_1(1) + \beta_2 x_{i2}$
0	x_2	$\beta_0 + \beta_1(0) + \beta_2 x_{i2}$
		β_1

Table 1: Interpretation of β_1.

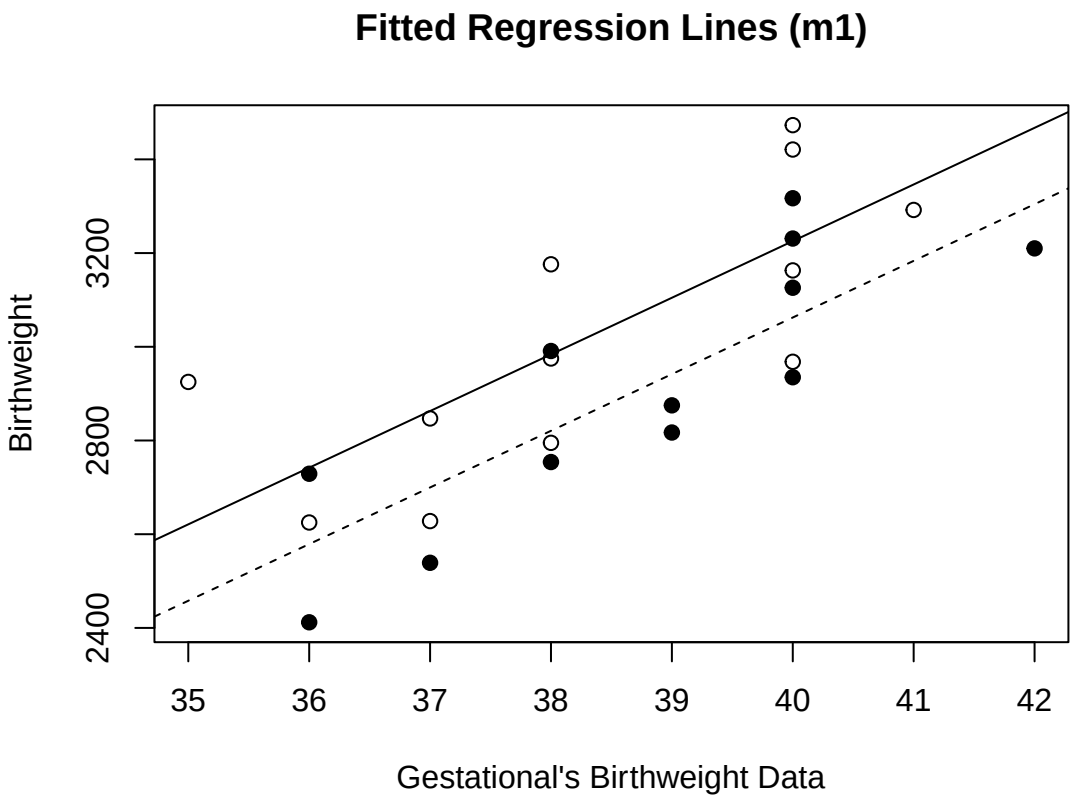
  $\beta_1 =$ Expected change in birthweight for male babies ($x_{i1} = 1$) versus female babies ($x_{i1} = 0$) at a fixed gestational age.

- To interpret β_2 consider the difference in the model with $x_{i2} + 1$ versus x_{i2} as seen in Table 2.

x_{i1}	x_{i2}	$\mathbb{E}[Y_i]$
x_1	$x_2 + 1$	$\beta_0 + \beta_1 x_{i1} + \beta_2(x_{i2} + 1)$
x_1	x_2	$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$
		β_2

Table 2: Interpretation of β_2.

  $\beta_2 =$ Expected change in birthweight associated with a one unit increase in gestational age (x_{i2}) adjusted for sex.



3. Model Checking

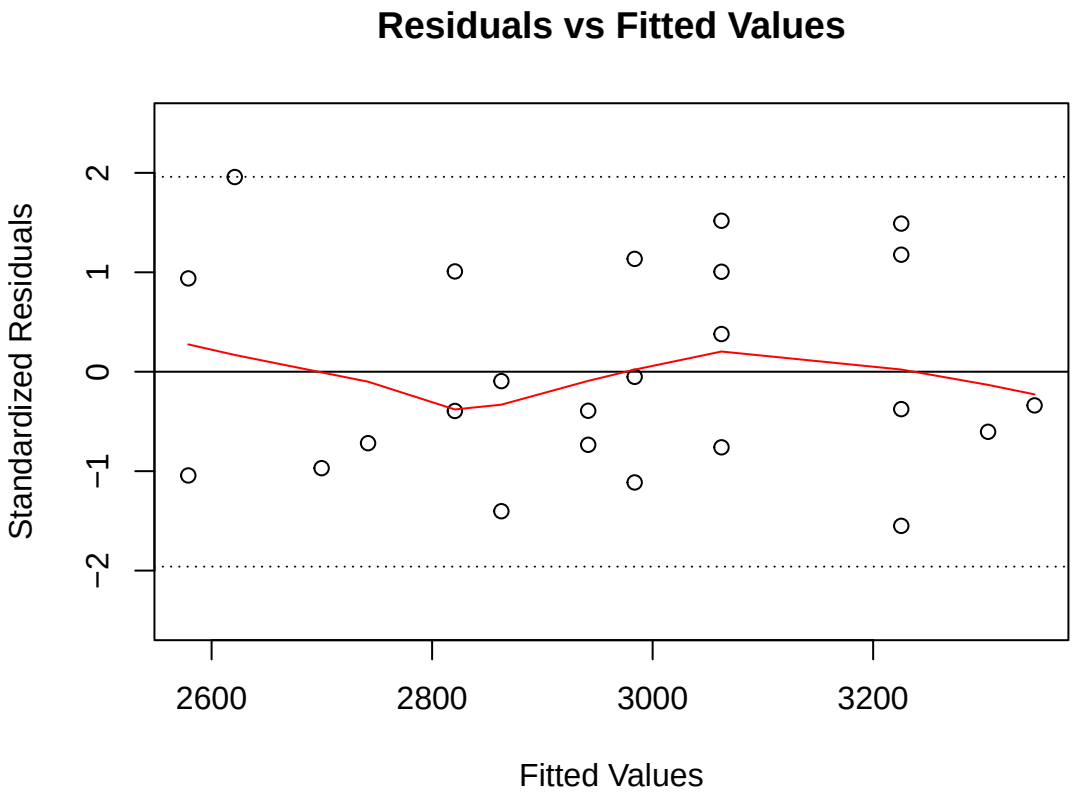
If the model provides a good fit to the data then asymptotic theory tells us that we should expect that the Standardized Residuals:

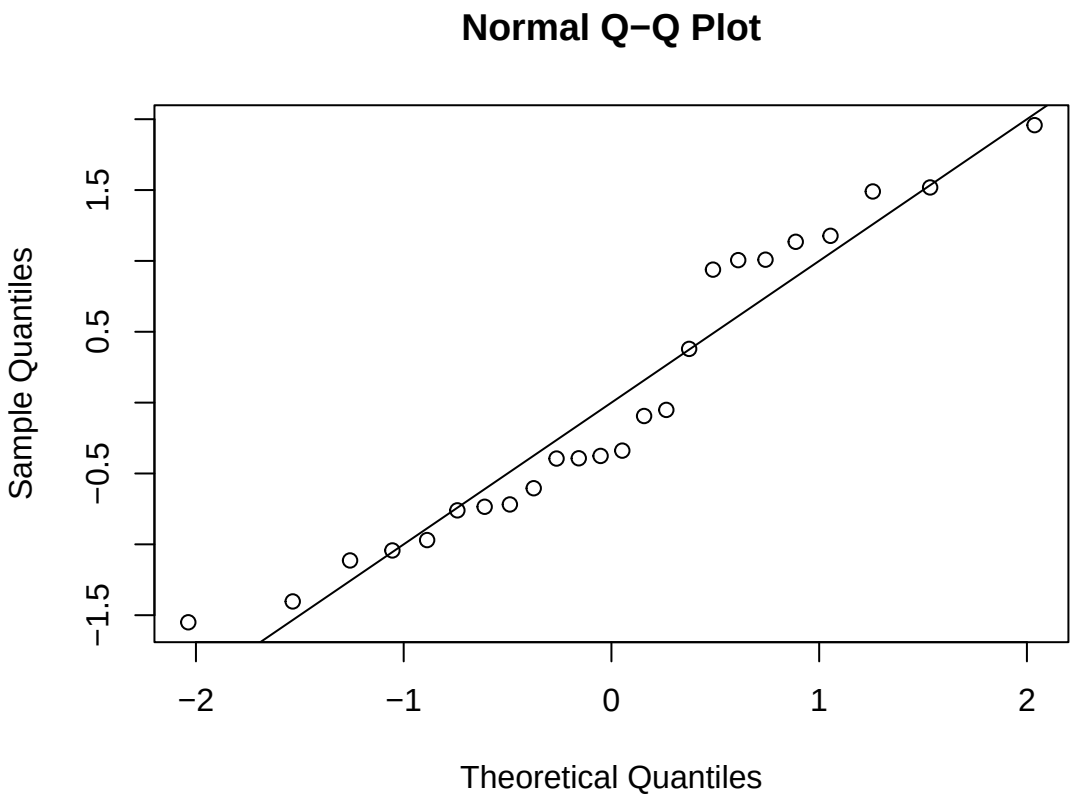
$$d_i = \frac{r_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}} \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \qquad \text{(approximately)}$$

Note that h_{ii} is the (i,i) element of $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$.

We visually check this by examining residual plots such as:

1. Standardized residuals versus the fitted values.
2. Standardized residuals versus the explanatory variable(s).
3. Normal probability plot (QQ plot) of the standardized residuals.
4. Added variable plots.


Residuals vs Fitted Values
Standardized Residuals
2
1
0
−1
−2
2600
2800
3000
3200
Fitted Values



4. Inference

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

$$\hat{\boldsymbol{\beta}} \sim \text{MVN}\big(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\big)$$

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 v_{jj})$$

Note that v_{jj} is the (j, j) element of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.

Confidence interval for β_j

$$\hat{\beta}_j \pm 1.96\sqrt{\sigma^2 v_{jj}}$$

Since σ^2 is generally unknown, we replace it with an unbiased estimate $\hat{\sigma}^2$ and use $\widehat{\mathsf{se}}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 v_{jj}}$.

$$\hat{\beta}_j \pm t_{n-p-1,\alpha/2}\widehat{\mathsf{se}}(\hat{\beta}_j)$$

Hypothesis Tests for β_j

To test:

$$H_0\text{: } \beta_j = \beta_j^\star \text{ vs } H_A\text{: } \beta_j \neq \beta_j^\star$$

we use the t-statistic:

$$t = \frac{\hat{\beta}_j - \beta_j^\star}{\widehat{\mathsf{se}}(\hat{\beta}_j)}$$

which has a t_{n-p-1} distribution when H_0 is true. That is, we reject H_0 if $|t| > t_{n-p-1,\alpha/2}$.

Example: Hypothesis Test and Confidence Interval

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1773.322    794.586  -2.232    0.037
sexM          163.039     72.808   2.239    0.036
age           120.894     20.463   5.908    0.000
```

$$\hat{\beta}_1 \pm t_{24-3}\,\mathsf{se}(\hat{\beta}_1) = 163.039 \pm 2.08(72.808) = (11.60, 314.48)$$

- After adjustment for gestational age, male babies are on average $163.04\,\mathrm{g}$ heavier than female babies. A $95\,\%$ confidence interval for this estimate is $(11.60, 314.48)$.
- We reject the null hypothesis that $\beta_1 = 0$:

$$p\text{-value} = \mathbb{P}\big(|t_{24-2-1}| > |t^\star|\big) = 2\,\mathbb{P}(t_{21} > 2.239) = 2\big(1 - \mathbb{P}(t_{21} < 2.239)\big) = 0.036 < 0.05$$

- We also reject the null hypothesis that $\beta_2 = 0$ since $p < 0.001$.

Example: Interaction Model

- Is the rate of increase of birthweight with gestational age the same for boys as for girls?

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

```
m2 <- lm(birthw ~ sex * age)
round(summary(m2)$coeff, 3)

             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2141.667   1163.601  -1.841    0.081
sexM          872.994   1611.331   0.542    0.594
age           130.400     29.998   4.347    0.000
sexM:age      -18.417     41.756  -0.441    0.664
```

- What is the interpretation of β_3?

Limitations of Linear Regression

Linear regression models can be very useful but may not be appropriate to use when:

- We cannot assume Y is normally distributed.
  - Binary data ($Y = 0$ or $Y = 1$).
  - Count data ($Y = 0, 1, 2, 3, \ldots$).
- The variance of Y depends on the mean μ.

Generalized Linear Models (GLM) extend the linear regression framework to address both of these issues.

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

Topic 1b: A Brief Review of Likelihood Methods

Likelihood Methods for Scalar Parameters

Setup

- Y is a random variable with a probability density or mass function $f(y \mid \theta)$, 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, or carry out tests of hypotheses regarding θ).

- Today's material: Appendix A1 & A2 of Stat 431 course notes, Dunn & Smyth Chapter 4, Stat 330 notes.

Likelihood Function

- The Likelihood function is any function which is proportional to the probability of observing the data you actually obtained:
$$\mathcal{L}(\theta \mid y) = c\,\mathbb{P}(Y = y \mid \theta) = cf(y \mid \theta)$$
- c is a *proportionality constant* which may be any positive function that does not depend on θ.
- $\mathcal{L}(\theta \mid y)$ contains all the information regarding θ from the data.
- $\mathcal{L}(\theta \mid y)$ ranks the parameter values of their consistency with the data.
- Since $\mathcal{L}(\theta \mid y)$ 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 the parameter value that makes the observed data the most likely (hence the term maximum likelihood).
- The maximum likelihood estimator (MLE) of θ is the value $\hat{\theta}$ that maximizes the likelihood function, that is:
$$\mathcal{L}(\hat{\theta} \mid y) \geq \mathcal{L}(\theta \mid y) \qquad \forall \theta \in \Omega$$
- Estimation is a simple optimization problem.
- Equivalently, since the log function is monotonic, $\hat{\theta}$ maximizes the log-likelihood function: $\ell(\theta \mid y) = \log\big(\mathcal{L}(\theta \mid y)\big)$.
- Often it is easier to work with $\ell(\theta \mid y)$ rather than $\mathcal{L}(\theta \mid y)$.
- For simplicity drop the y and use $\mathcal{L}(\theta) = \mathcal{L}(\theta \mid y)$.

Other Important Functions

- $\ell(\theta) = \log\big(\mathcal{L}(\theta)\big)$ be the log-likelihood function.
- $S(\theta) = \ell'(\theta)$ be the first derivative of the log-likelihood function which is called the Score function.
- $I(\theta) = -\ell''(\theta)$ be the negative second derivative of the log-likelihood function which is called the Information function.
- $\mathcal{I}(\theta) = \mathbb{E}[I(\theta)]$ be the Expected information function.

- $R(\theta) = \mathcal{L}(\theta)/\mathcal{L}(\hat{\theta})$ be the Relative likelihood function, which is the likelihood function standardized by its maximum value so that the relative likelihood will have a maximum value of 1 ($0 \leq R(\theta) \leq 1$).
- $r(\theta) = \log\big(\mathcal{L}(\theta)/\mathcal{L}(\hat{\theta})\big)$ be the log relative likelihood function which will have a maximum value of 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:
$$I(\hat{\theta}) = -\ell''(\hat{\theta}) > 0$$
- Should also check if there are any values of θ at the edges of Ω that give a local maxima of $\ell(\theta)$.
- Invariance property of MLEs: if $g(\theta)$ is any function of the parameter θ, then the MLE of $g(\theta)$ is $g(\hat{\theta})$.

Example: Poisson Distribution

Example: Poisson Distribution

Let $Y_1, Y_2, \ldots, Y_n$ be iid Poisson random variables with

$$f(y_i \mid \theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!}, \qquad \theta > 0, y_i = 0, 1, 2, \ldots$$

with unknown parameter θ. Find the MLE of θ.

- Likelihood function:
$$\mathcal{L}(\theta \mid y) = \prod_{i=1}^{n} f(y_i \mid \theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} \exp\{-\theta\}}{y_i!} = \frac{\theta^{\sum_i y_i} \exp\{-n\theta\}}{\prod_i y_i!}$$
- log-likelihood function:
$$\ell(\theta \mid y) = \Big(\sum y_i\Big) \log(\theta) - n\theta - \sum \big(\log(y_i!)\big)$$
- Score function:
$$S(\theta) = \frac{\mathrm{d}\ell}{\mathrm{d}\theta} = \frac{1}{\theta} \sum y_i - n$$
- Maximum likelihood estimate:
$$0 = \frac{1}{\hat{\theta}} \sum y_i - n \implies \hat{\theta} = \frac{\sum y_i}{n} = \bar{y}$$
- Second derivative test using Information function:
$$I(\theta) = -\frac{\mathrm{d}^2\ell}{\mathrm{d}\theta^2} = \frac{1}{\theta^2} \sum y_i > 0 \qquad \forall \theta > 0$$
Confirms that $\hat{\theta} = \bar{y}$ is the maximum likelihood estimate.
- See Appendix A2 for a Binomial example.

Example: Topical cyclones

Tropical cyclones

Number of tropical cyclones in Northeastern Australia for the 13 successive seasons 1956-57 through 1968-69 (Dobson §1.6.5)

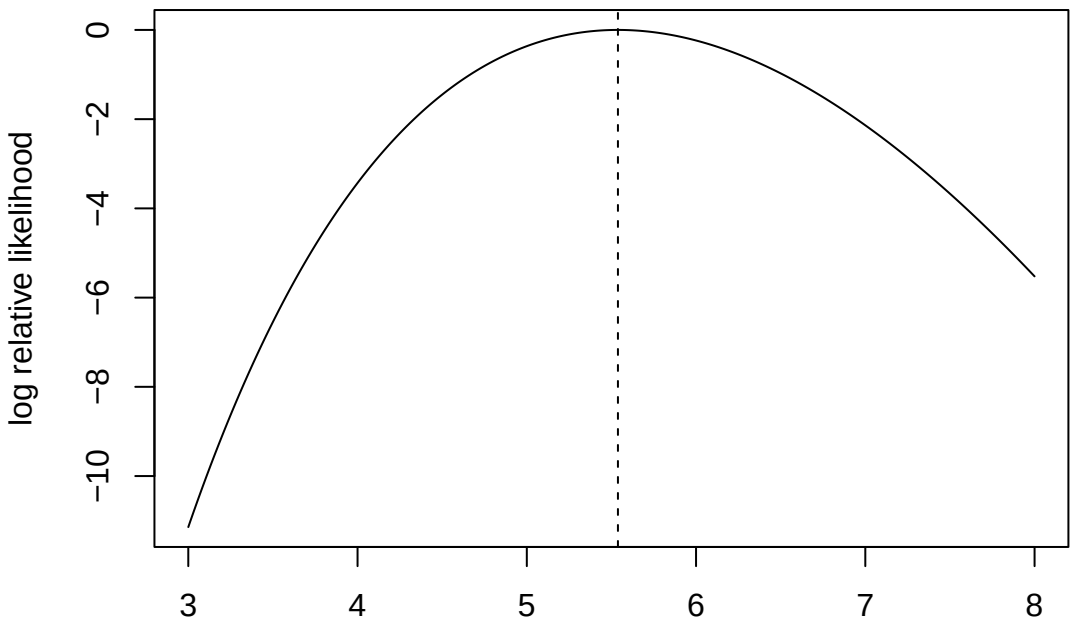
Season	1	2	3	4	5	6	7	8	9	10	11	12	13
Cyclones	6	5	4	6	6	3	12	7	4	2	6	7	4

- Let $Y_i =$ number of cyclones in season i.
- Assume the Y_i's are iid Poisson random variables with unknown parameter θ.
- The maximum likelihood estimate of θ is:

$$\hat{\theta} = \bar{y} = \frac{\sum y_i}{n} = \frac{72}{13} = 5.538$$

Example: Plot of log relative likelihood function

$$r(\theta) = \ell(\theta) - \ell(\hat{\theta}) = \left(\sum y_i\right) \log\left(\frac{\theta}{\hat{\theta}}\right) - n(\theta - \hat{\theta})$$



Newton Raphson Algorithm

- Sometimes we need to solve $S(\theta)$ iteratively.

- Taylor Series expansion of a differentiable function $f(x)$:

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

- Now suppose we wish to find $\hat{\theta}$, the root of $S(\theta) = 0$, and θ_0 is a guess that is "close to $\hat{\theta}$."
- Consider the Taylor series expansion of $S(\theta)$ about θ_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 $|\theta - \theta_0|$ small, we can drop the second and higher order terms and to a good approximation we have:

$$S(\theta) \simeq S(\theta_0) + S'(\theta_0)(\theta - \theta_0)$$
$$S(\theta) \simeq S(\theta_0) - I(\theta_0)(\theta - \theta_0)$$

- We are approximating $S(\theta)$ with a linear function that has the same value and slope as $S(\theta)$ at $\theta = \theta_0$. 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

We wish to maximize the function $\ell(\theta)$ by solving $S(\theta) = 0$.

- Begin with an initial estimate θ_0.
- Iteratively obtain estimates $\theta_1, \theta_2, \theta_3, \ldots$ using:

$$\theta_{i+1} = \theta_i + I^{-1}(\theta_i)S(\theta_i)$$

- Iteration should continue until $\theta_{i+1} \simeq \theta_i$. (i.e., $|\theta_{i+1} - \theta_i|$ is within a specified tolerance).
- Then set $\hat{\theta} = \theta_{i+1}$.
- To determine if it is a maxima of $\ell(\theta)$, check that $I(\hat{\theta}) > 0$.

Example: Newton Raphson for Cyclone Data

```
# Input the Cyclone data, define the score and information
# function
y <- c(6, 5, 4, 6, 6, 3, 12, 7, 4, 2, 6, 7, 4)
Score <- function(theta, y) {
  sum(y)/theta - length(y)
```

```
}
Info <- function(theta, y) {
  sum(y)/(theta^2)
}
# Run one-parameter Newton Raphson algorithm (track each
# iteration)
theta.old <- 0
theta.new <- 5
track <- c(theta.new, Score(theta.new, y))
while ((theta.new - theta.old)^2 > 10^(-3)) {
  theta.old <- theta.new
  theta.new <- theta.old + Score(theta.old, y)/Info(theta.old,
    y)
  track <- rbind(track, c(theta.new, Score(theta.new, y)))
}
track
```

```
          [,1]         [,2]
track 5.000000 1.400000e+00
      5.486111 1.240506e-01
      5.537967 1.161567e-03
      5.538461 1.037690e-07
```

```
mean(y)
```

```
[1] 5.538462
```

Inference for Scalar Parameters

- So far we have discussed estimation of $\hat{\theta}$.
- Next, we want to conduct inference (carry out hypothesis tests and construct confidence intervals).
- Several techniques are available, all based to varying degrees on the likelihood function.

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: $(\hat{\theta} - \theta)^2 I(\hat{\theta}) \sim \chi^2_{(1)}$.

(approximations improve as sample size increases)

Confidence Intervals

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

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

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

- The Wald-based pivotal gives an interval:

$$\left\{\theta : (\hat{\theta} - \theta) I(\hat{\theta}) < \chi^2_{(1)}(1-\alpha)\right\}$$

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

- The Wald-based interval should actually be familiar to you:
- Recall if $Z \sim \mathcal{N}(0,1)$ then $Z^2 \sim \chi^2_{(1)}$.
- So we have:

$$\begin{aligned}
\left\{\theta : (\hat{\theta} - \theta)^2 I(\hat{\theta}) < \chi^2_{(1)}(1-\alpha)\right\} &= \left\{\theta : \left|(\hat{\theta} - \theta) I^{1/2}(\hat{\theta})\right| < z_{(1-\alpha/2)}\right\} \\
&= \left\{\theta : \left|\hat{\theta} - \theta\right| < I^{-1/2} z_{(1-\alpha/2)}\right\} \qquad \text{since } I(\hat{\theta}) > 0 \\
&= \hat{\theta} \pm z_{(1-\alpha/2)} I^{-1/2}(\hat{\theta})
\end{aligned}$$

- This is the "standard" normal based confidence interval with $\mathsf{se}(\hat{\theta}) = I^{-1/2}(\hat{\theta})$.

Example: LR Confidence Interval for Cyclone Data

Likelihood Ratio based interval: $\left\{\theta : -2r(\theta) < \chi^2_{(1)}(1-\alpha)\right\}$.

- For the Poisson distribution $\hat{\theta} = \bar{y}$ so:

$$r(\theta) = \ell(\theta) - \ell(\hat{\theta}) = n\bar{y} \log\left(\frac{\theta}{\bar{y}}\right) - n(\theta - \bar{y})$$

- To find the interval find the roots of $-2r(\theta) - \chi^2_{(1)}(1-\alpha)$.

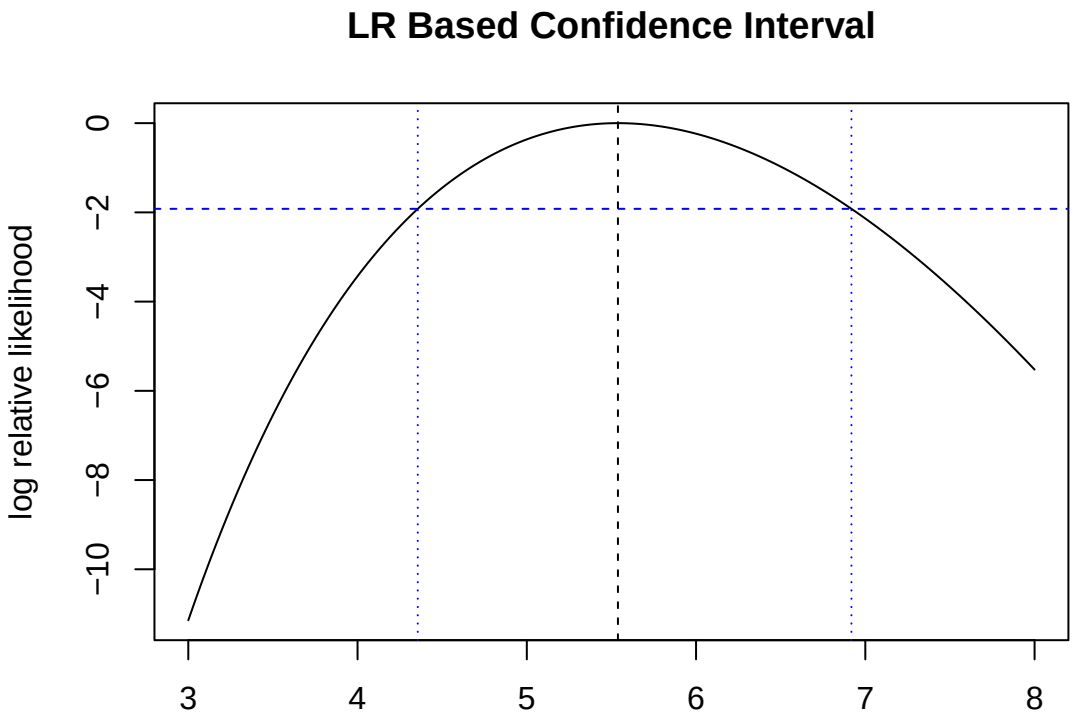
```
ybar <- mean(y)
n <- length(y)
LRest <- function(theta, ybar, n) {
  -2 * (n * ybar * log(theta/ybar) - n * (theta - ybar)) -
    qchisq(0.95, 1)
}
uniroot(LRest, interval = c(3, ybar), ybar = ybar, n = n)$root
```

```
[1] 4.355715
```

```
uniroot(LRest, interval = c(ybar, 8), ybar = ybar, n = n)$root
```

```
[1] 6.918103
```

The likelihood ratio based $95\,\%$ confidence interval is $(4.36, 6.92)$.



Example: Wald Confidence Interval for Cyclone Data

Wald based interval: $\left\{\theta : (\hat{\theta} - \theta) I(\hat{\theta}) < \chi^2_{(1)}(1-\alpha)\right\}$.

- For the Poisson distribution $\hat{\theta} = \bar{y}$ and

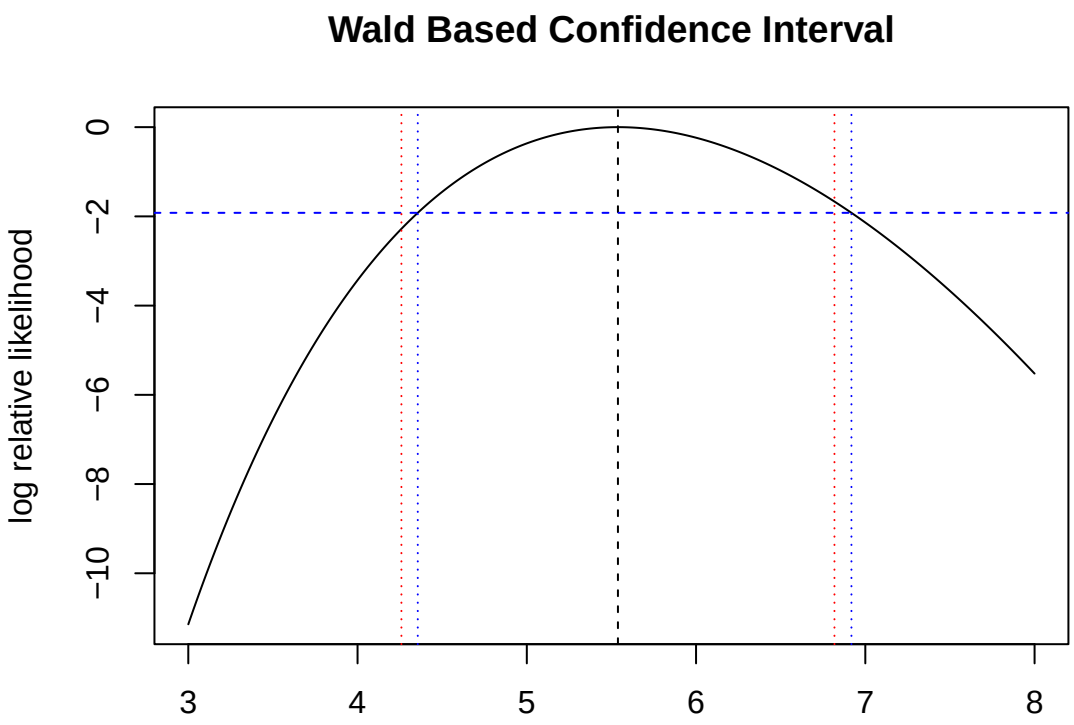
$$I(\hat{\theta}) = \frac{1}{\hat{\theta}^2} \sum y_i = \frac{n\bar{y}}{\bar{y}^2} = \frac{n}{\bar{y}}$$

- So we solve:

$$\begin{aligned} \hat{\theta} \pm 1.96\big(I(\hat{\theta})\big)^{-1/2} &= \bar{y} \pm 1.96(n/\bar{y})^{-1/2} \\ &= 5.538462 \pm 1.96(0.652714) \\ &= (4.2591, 6.8178) \end{aligned}$$

The likelihood ratio based $95\,\%$ confidence interval is $(4.36, 6.92)$.

The Wald based $95\,\%$ confidence interval is $(4.26, 6.82)$.



Testing Hypotheses

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

- Likelihood ratio (LR) test:

$$p = \mathbb{P}\Big(\chi^2_{(1)} > -2r(\theta_0)\Big)$$

- Score test:

$$p = \mathbb{P}\Big(\chi^2_{(1)} > \big(S(\theta_0)\big)^2 / I(\theta_0)\Big)$$

- Wald test:

$$p = \mathbb{P}\Big(\chi^2_{(1)} > (\hat{\theta} - \theta_0)^2 I(\hat{\theta})\Big)$$

  or

$$p = \mathbb{P}\Big(|Z| > |\theta - \theta_0|\sqrt{I(\hat{\theta})}\Big)$$

Example: Hypothesis Tests for Cyclone Data

Suppose we wish to test whether there were an average of 5 cyclones per year

$$H_0\text{: } \theta = 5 \text{ vs. } H_A\text{: } \theta \neq 5.$$

- Likelihood Ratio based test:

$$r(\theta_0 = 5) = n\bar{y}\log\left(\frac{5}{\bar{y}}\right) - n(5 - \bar{y}) = -0.3641$$

The p-value for this test is:

$$p = \mathbb{P}\Big(\chi^2_{(1)} > -2r(5)\Big) = \mathbb{P}\Big(\chi^2_{(1)} > 0.7282\Big) = 0.3934$$

Therefore we *do not reject* H_0.

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 (A3)

Suppose $\boldsymbol{\theta} \in \Omega$ is a continuous $p \times 1$ parameter vector indexing a probability density or mass function $f(\boldsymbol{y} \mid \boldsymbol{\theta})$.

- $\mathcal{L}(\boldsymbol{\theta})$ is the Likelihood function.
- $\ell(\boldsymbol{\theta}) = \log\big(\mathcal{L}(\boldsymbol{\theta})\big)$ is the log-likelihood function.
- $\boldsymbol{S}(\boldsymbol{\theta}) = \frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is the $p \times 1$ Score vector.
- $\boldsymbol{I}(\boldsymbol{\theta}) = -\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top \partial \boldsymbol{\theta}}$ is the $p \times p$ Information matrix.
- $R(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta})/\mathcal{L}(\hat{\boldsymbol{\theta}})$ is the Relative likelihood function.
- $r(\boldsymbol{\theta}) = \log\big(\mathcal{L}(\boldsymbol{\theta})/\mathcal{L}(\hat{\boldsymbol{\theta}})\big) = \ell(\boldsymbol{\theta}) - \ell(\hat{\boldsymbol{\theta}})$ is the log relative likelihood function.
- The Newton Raphson algorithm applies as before, but with vectors and matrices as follows:

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

- Again, we apply iteratively until we obtain convergence, but now check to see if $\boldsymbol{I}(\boldsymbol{\theta})$ is a positive definite matrix.
- Analogs to the LR, Score and Wald results apply based on partitioning the Information matrix by $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta})^\top$, where $\boldsymbol{\alpha}$ is a $p \times 1$ vector of nuisance parameters and $\boldsymbol{\beta}$ is a $q \times 1$ vector of parameters of interest:

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

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

Topic 1c: Likelihood for Generalized Linear Models

Likelihood for Generalized Linear Models

Recall Stat 331/371: Assume $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ independently. For linear regression:

$$\mathbb{E}[Y_i] = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

How can we do regression analysis if the distribution of Y_i is not Normal?

1. Definition of the Exponential Family.
    - Derivation of general likelihood results for the Score and Information.
    - Application of general results to the Exponential Family.
    - Definition of the canonical link.
    - Poisson example.
2. Definition of a Generalized Linear Model.

The Exponential Family

Definition (Exponential Family)

Consider a random variable Y_i with p.d.f. $f(y_i \mid \theta_i, \phi)$, θ_i unknown, ϕ known. We say that the distribution is a member of the exponential family if we can write the p.d.f. in the form:

$$f(y_i \mid \theta_i, \phi) = \exp\left\{ \frac{\big(y_i\theta_i - b(\theta_i)\big)}{a_i(\phi)} + c(y_i \mid \phi) \right\}$$

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

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

Likelihood for the Exponential Family

Consider a single observation y_i from the exponential family.

- Likelihood:
$$\mathcal{L}_i(\theta_i, \phi \mid y_i) = f(y_i \mid \theta_i, \phi) = \exp\left\{ \frac{\big(y_i\theta_i - b(\theta_i)\big)}{a_i(\phi)} + c(y_i \mid \phi) \right\}$$
- Log-likelihood:
$$\ell_i(\theta_i, \phi \mid y_i) = \log\big(f(y_i \mid \theta_i, \phi)\big) = \frac{\big(y_i\theta_i - b(\theta_i)\big)}{a_i(\phi)} + c(y_i \mid \phi)$$
- Score:
$$S_i(\theta_i) = \frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)}$$
- Observed Information:
$$I_i(\theta_i) = \frac{\partial^2 \ell_i}{\partial \theta_i^2} = \frac{b''(\theta_i)}{a_i(\phi)}$$
- Fisher/Expected Info:
$$\mathcal{I}_i(\theta_i) = \mathbb{E}\left[-\frac{\partial^2 \ell_i}{\partial \theta_i^2} \right]$$

Aside: General Results for the Score and Information

Fact: Probability density functions integrate to 1. Using this,

$$\begin{aligned}\int f(y_i \mid \theta_i, \phi)\,\mathrm{d}y_i &= 1\\ \frac{\partial}{\partial\theta_i}\int f(y_i \mid \theta_i, \phi)\,\mathrm{d}y_i &= \frac{\partial 1}{\partial\theta_i}\\ \int \frac{\partial}{\partial\theta_i} f(y_i \mid \theta_i, \phi)\,\mathrm{d}y_i &= 0 \end{aligned} \tag{1}$$

When differentiating the log-likelihood we have:

$$\begin{aligned}\frac{\partial}{\partial\theta_i}\log\big(f(y_i \mid \theta_i, \phi)\big) &= \frac{1}{f(y_i \mid \theta_i, \phi)}\frac{\partial}{\partial\theta_i} f(y_i \mid \theta_i, \phi)\\ f(y_i \mid \theta_i, \phi)\frac{\partial}{\partial\theta_i}\log\big(f(y_i \mid \theta_i, \phi)\big) &= \frac{\partial}{\partial\theta_i} f(y_i \mid \theta_i, \phi)\end{aligned} \tag{2}$$

Substituting (2) into (1) we get:

$$\begin{aligned}\int f(y_i \mid \theta_i, \phi)\frac{\partial}{\partial\theta_i}\log\big(f(y_i \mid \theta_i, \phi)\big)\,\mathrm{d}y_i &= 0\\ \int f(y_i \mid \theta_i, \phi) S_i(\theta_i)\,\mathrm{d}y_i &= 0\\ \mathbb{E}[S_i(\theta_i)] &= 0\end{aligned} \tag{3}$$

since by definition $\mathbb{E}[g(X)] = \int g(x) f(x \mid \theta)\,\mathrm{d}x$.

Result # 1

The expectation of the score function is zero.

$$\mathbb{E}[S_i(\theta_i)] = 0$$

Differentiate (3) again:

$$\begin{aligned}0 &= \int f(y_i \mid \theta_i, \phi)\frac{\partial}{\partial\theta_i}\log\big(f(y_i \mid \theta_i, \phi)\big)\,\mathrm{d}y_i\\ \frac{\partial 0}{\partial\theta_i} &= \frac{\partial}{\partial\theta_i}\int f(y_i \mid \theta_i, \phi)\frac{\partial}{\partial\theta_i}\log\big(f(y_i \mid \theta_i, \phi)\big)\,\mathrm{d}y_i\\ 0 &= \int\left[\frac{\partial^2}{\partial\theta_i^2}\log\big(f(y_i \mid \theta_i, \phi)\big)\right] f(y_i \mid \theta_i, \phi)\,\mathrm{d}y_i + \int \frac{\partial}{\partial\theta_i}\log\big(f(y_i \mid \theta_i, \phi)\big)\left[\frac{\partial}{\partial\theta_i} f(y_i \mid \theta_i, \phi)\right]\mathrm{d}y_i\\ 0 &= \int \frac{\partial^2}{\partial\theta_i^2}\log\big(f(y_i \mid \theta_i, \phi)\big) f(y_i \mid \theta_i, \phi)\,\mathrm{d}y_i + \int\left(\frac{\partial}{\partial\theta_i}\log\big(f(y_i \mid \theta_i, \phi)\big)\right)^2 f(y_i \mid \theta_i, \phi)\,\mathrm{d}y_i \qquad \text{Sub (2)}\\ 0 &= \mathbb{E}\left[\frac{\partial^2}{\partial\theta_i^2}\log\big(f(y_i \mid \theta_i, \phi)\big)\right] + \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log\big(f(y_i \mid \theta_i, \phi)\big)\right)^2\right]\end{aligned} \tag{4}$$

Examining (4) we get:

$$\begin{aligned}\mathbb{E}\left[\frac{\partial^2}{\partial\theta_i^2}\log\big(f(y_i \mid \theta_i, \phi)\big)\right] + \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log\big(f(y_i \mid \theta_i, \phi)\big)\right)^2\right] &= 0\\ \mathbb{E}[-I_i(\theta_i)] + \mathbb{E}\big[S_i(\theta_i)^2\big] &= 0\\ \mathbb{E}\big[S_i(\theta_i)^2\big] &= \mathbb{E}[I_i(\theta_i)] = \mathcal{I}_i(\theta_i)\end{aligned}$$

Result # 2

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

$$\mathbb{E}\left[S_i(\theta_i)^2\right] = \mathbb{E}[I_i(\theta_i)] = \mathcal{I}_i(\theta_i)$$

Recall that $\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \left(\mathbb{E}[X]\right)^2$. Using Results #1 and #2 we have:

$$\begin{aligned}\mathsf{Var}(S_i(\theta_i)) &= \mathbb{E}\left[S_i(\theta_i)^2\right] - \mathbb{E}[S_i(\theta_i)]^2 \\ &= \mathcal{I}_i(\theta_i) - 0^2 \\ &= \mathcal{I}_i(\theta_i)\end{aligned}$$

Result # 3

The variance of the score function is the expected information:

$$\mathsf{Var}(S_i(\theta_i)) = \mathcal{I}_i(\theta_i)$$

Applying these Results to the Exponential Family

$$\begin{aligned}\mathbb{E}[S_i(\theta_i)] &= 0 \\ \mathbb{E}\left[\frac{Y_i - b'(\theta_i)}{a_i(\phi)}\right] &= 0 \\ \mathbb{E}[Y_i] &= b'(\theta_i)\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[S_i(\theta_i)^2\right] &= \mathbb{E}[I_i(\theta_i)] \\ \mathbb{E}\left[\left(\frac{Y_i - b'(\theta_i)}{a_i(\phi)}\right)^2\right] &= \mathbb{E}\left[\frac{b''(\theta_i)}{a_i(\phi)}\right] \\ \frac{1}{a_i(\phi)^2}\mathbb{E}\left[(Y_i - \mathbb{E}[Y_i])^2\right] &= \frac{b''(\theta_i)}{a_i(\phi)} \\ \mathsf{Var}(Y_i) &= b''(\theta_i)a_i(\phi)\end{aligned}$$

Properties of the Exponential Family

For a random variable Y_i with a distribution in the exponential family, θ_i unknown, ϕ known:

$$\mathcal{L}_i(\theta_i, \phi \mid y_i) = f(y_i \mid \theta_i, \phi) = \exp\left\{\frac{\left(y_i\theta_i - b(\theta_i)\right)}{a_i(\phi)} + c(y_i \mid \phi)\right\}$$

Mean and Variance for the Exponential Family

- Mean: $\mathbb{E}[Y_i] = b'(\theta_i) = \mu_i$.
- Variance: $\mathsf{Var}(Y_i) = b''(\theta_i)a_i(\phi)$

- $\mathsf{V}(\mu_i) = b''(\theta_i)$ is called the Variance function.
- $b''(\theta_i)$ is a function of the canonical parameter θ_i and hence a function of the mean (mean-variance relationship)
- $a_i(\phi)$ is a known function of the dispersion parameter ϕ.
- Often we can write $a_i(\phi) = \phi/w_i$ where w_i is a weight.

Link Functions

Definition (Link Function)

The link function $g(\mu_i)$ relates the linear predictor $\eta_i = \boldsymbol{x}^\top \boldsymbol{\beta}$ to the expected value μ_i of the random variable Y_i.

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

Definition (Canonical Link Function)

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

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

(i.e., canonical parameter = linear predictor)

Examples

Many well known distributions belong to the exponential family:

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

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

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

$$f(y \mid \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}, \qquad y = 0, 1, 2, \ldots$$

- Binomial Distribution: $Y \sim \text{BIN}(m, \pi)$.

$$f(y \mid \pi) = \binom{m}{y} \pi^y (1-\pi)^{m-y}, \qquad y = 0, 1, \ldots, m$$

The Poisson Distribution

Let Y_i, $i = 1, 2, \ldots, n$ be iid $\text{POI}(\lambda_i)$.

$$f(y_i \mid \lambda_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, \qquad y_i = 0, 1, 2, \ldots$$

Show that the distribution of Y_i is a member of the exponential family and find the mean, variance, variance function and canonical link function.

Exponential Family: Full Disclosure

The definition of the exponential family used in the Stat 431 course notes is actually a special case of:

Definition (General Exponential Family)

A distribution is a member of the General Exponential Family if it can be expressed as:

$$f(y \mid \boldsymbol{\theta}) = \exp\left\{\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) t_i(y) + b(\boldsymbol{\theta}) + h(y)\right\}$$

for $t_1(y), \ldots, t_k(y)$ real-valued function of y, and $w_1(\boldsymbol{\theta}), \ldots, w_k(\boldsymbol{\theta})$ real-valued functions of the possibly vector-valued parameter $\boldsymbol{\theta}$.

Random Sample from the Exponential Family

Now suppose Y_i, $i = 1, 2, \ldots, n$ are iid with a distribution that is a member of the exponential family. Then:

$$\mathcal{L}(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}) = \prod_{i=1}^{n} f(y_i \mid \theta_i, \phi) = \prod_{i=1}^{n} \exp\left\{\frac{\big(y_i\theta_i - b(\theta_i)\big)}{a_i(\phi)} + c(y_i \mid \phi)\right\}$$

$$\ell(\boldsymbol{\theta}, \phi \mid \boldsymbol{y}) = \sum_{i=1}^{n} \log\big(f(y_i \mid \theta_i, \phi)\big) = \sum_{i=1}^{n} \left(\frac{\big(y_i\theta_i - b(\theta_i)\big)}{a_i(\phi)} + c(y_i \mid \phi)\right)$$

In a regression context, we are interested in estimating $\boldsymbol{\beta}$ under the link function:

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

where $\boldsymbol{x}_i$ is a vector of explanatory variables for subject $i = 1, 2, \ldots, n$.

Generalized Linear Models

Definition (Generalized Linear Model (GLM))

A Generalized Linear Model (GLM) is composed of:

- The Random Component: The distribution of the iid response variables Y_i is assumed to come from a parametric distribution that is a member of the exponential family.
- The 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 $\boldsymbol{\beta}$.
- The Link function that relates the mean of the distribution of Y_i to the linear predictor through:

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

Topic Summary: Likelihood for Generalized Linear Models

1. Definition of the Exponential Family.
   - Derivation of general likelihood results for the Score and Information.
   - Application of general results to the Exponential Family.
   - Definition of the canonical link.
   - Poisson example.
2. Definition of a Generalized Linear Model.

Next Topic: Estimation for Generalized Linear Models.

Estimation of $\boldsymbol{\beta}$ from a GLM through Iteratively Reweighted Least Squares (IRWLS).

Topic 1d: Estimation for GLMs

Generalized Linear Models

Definition

A Generalized Linear Model (GLM) is composed of:

1. The Random Component: The distribution of the response variables 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 $\boldsymbol{\beta}$ that relates the mean of the distribution of Y_i to the linear predictor through

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

Estimation of β from a GLM through IRWLS

Consider the log-likelihood for a single observation from the exponential family:

$$\ell_i(\theta_i, \phi \mid y_i) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i; \phi)$$

- ℓ_i is a function of θ_i (assume that ϕ is known).
- μ_i can be expressed in terms of θ_i through the mean:

$$\mu_i = b'(\theta_i)$$

- η_i can be expressed in terms of μ_i through the link function:

$$\eta_i = g(\mu_i)$$

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

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

Thus, $\ell_i(\theta_i, \phi \mid y_i)$ depends on θ_i, so θ_i depends on μ_i, so μ_i depends on η_i, and so η_i depends on β_j. Therefore, we will use the chain rule on:

$$\ell_i(\beta_j) = f\Big(\theta_i\Big(\mu_i\big(\eta_i(\beta_j)\big)\Big)\Big)$$

The Score Vector

Using Maximum Likelihood to estimate $\boldsymbol{\beta}$, we must solve $\boldsymbol{S}(\boldsymbol{\beta}) = \mathbf{0}_p$. Consider the j^{th} element of the score vector:

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$$

where

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)} \qquad \text{since } b'(\theta_i) = \mu_i$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1} = \frac{1}{b''(\theta_i)} = \frac{a_i(\phi)}{\mathsf{Var}(\mu_i)} \qquad \text{since } \mu_i = b'(\theta_i),\ \mathsf{Var}(\mu_i) = b''(\theta_i)a_i(\phi)$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \eta_i} \qquad \text{(depends on selected link)}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij} \qquad \text{since } \boldsymbol{x}_i^\top \boldsymbol{\beta} = \eta_i = \sum_{j=0}^{p-1} x_{ij}\beta_j, \text{ for } i = 1, \ldots, n$$

So we have

$$\begin{aligned}\frac{\partial \ell_i}{\partial \beta_j} &= \frac{y_i-\mu_i}{a_i(\phi)}\frac{a_i(\phi)}{\mathsf{Var}(\mu_i)}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} \\ &= \frac{y_i-\mu_i}{\mathsf{Var}(y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2\frac{\partial \eta_i}{\partial \mu_i}x_{ij} && \text{multiply by } 1 = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \mu_i} \\ &= (y_i-\mu_i)w_i\frac{\partial \eta_i}{\partial \mu_i}x_{ij}\end{aligned}$$

where $w_i = \frac{1}{\mathsf{Var}(Y_i)(\frac{\partial \eta_i}{\partial \mu_i})^2}$. Note that generally $\frac{\partial \eta_i}{\partial \mu_i}$ is easier to calculate than $\frac{\partial \mu_i}{\partial \eta_i}$ since we define the link as $\eta_i = g(\mu_i)$.

With n iid observations, the j^{th} element of the score vector is:

$$\big[\boldsymbol{S}(\boldsymbol{\beta})\big]_j = \sum_{i=1}^n \frac{\partial \ell_i}{\partial \beta_j} = \sum_{i=1}^n (y_i-\mu_i)w_i\frac{\partial \eta_i}{\partial \mu_i}x_{ij} \quad \text{for } j = 0,1,\ldots,p-1$$

In vector form we can write:

$$\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{\mu}) \circ \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}}$$

where $\boldsymbol{y} = (y_1,\ldots,y_n)^\top$ and $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_n)^\top$ are $n \times 1$ vectors, $\boldsymbol{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ is a $p \times n$ matrix, $\boldsymbol{W}$ denotes the $n \times n$ diagonal matrix with $\boldsymbol{W} = \mathsf{diag}(w_1, w_2, \ldots, w_n)$, $\circ$ denotes an element-wise product, and $\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}} = \big(\frac{\partial \eta_1}{\partial \mu_1}, \frac{\partial \eta_2}{\partial \mu_2}, \ldots, \frac{\partial \eta_n}{\partial \mu_n}\big)^\top$.

Example: The Poisson Distribution (Problem 1.4)

Let Y_i, $i = 1,\ldots,n$ be independent Poisson random variables with $\mathbb{E}[Y_i] = \lambda_i$. Suppose that associated with each y_i is a $p \times 1$ vector of explanatory variables $\boldsymbol{x}_i$. A Poisson regression model with the canonical link takes the form:

$$\mathsf{log}(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i(p-1)} = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

To answer the following you may either calculate the derivatives using standard methods, or use the general results derived in class for the exponential family.

a. Write down the score vector for the regression coefficients β.

Newton Raphson and Fisher Scoring

To solve $\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{0}_p$, the Newton Raphson update equation is:

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

where $\boldsymbol{I}(\,\cdot\,)$ is the observed information matrix.

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

The Fisher Scoring update equation is:

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

The Information Matrix

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

$$\begin{aligned}
I_{jk} &= -\frac{\partial^2 \ell_i}{\partial\beta_j\,\partial\beta_k} \\
&= -\frac{\partial}{\partial\beta_k}\frac{\partial\ell_i}{\partial\beta_j} \\
&= -\frac{\partial}{\partial\beta_k}\left[(y_i-\mu_i)w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)x_{ij}\right] \\
&= -(y_i-\mu_i)\left\{\frac{\partial}{\partial\beta_k}\left[w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)x_{ij}\right]\right\} - w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)x_{ij}\left[\frac{\partial}{\partial\beta_k}(y_i-\mu_i)\right] \\
&= -(y_i-\mu_i)\left\{\frac{\partial}{\partial\beta_k}\left[w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)x_{ij}\right]\right\} + w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)x_{ij}\frac{\partial\mu_i}{\partial\eta_i}x_{ik} \\
&= -(y_i-\mu_i)\frac{\partial}{\partial\beta_k}\left[w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)x_{ij}\right] + x_{ij}w_ix_{ik}
\end{aligned}$$

Where the above holds since

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} \implies \frac{\partial\mu_i}{\partial\beta_k} = \frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_k} = \frac{\partial\mu_i}{\partial\eta_i}x_{ik}$$

Fisher Scoring

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

$$\begin{aligned}
\mathcal{I}_{jk} &= \mathbb{E}\left[-\frac{\partial^2\ell_i}{\partial\beta_j\,\partial\beta_k}\right] \\
&= \mathbb{E}\left[-(y_i-\mu_i)\frac{\partial}{\partial\beta_k}\left[w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)x_{ij}\right] + x_{ij}w_ix_{ik}\right] \\
&= \frac{\partial}{\partial\beta_k}\left[w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)x_{ij}\right]\mathbb{E}[(y_i-\mu_i)] + x_{ij}w_ix_{ik} \\
&= x_{ij}w_ix_{ik} && \text{since } \mathbb{E}[(y_i-\mu_i)] = 0
\end{aligned}$$

Therefore, for n observations we can write:

$$\mathcal{I}_{jk} = \sum_{i=1}^{n} x_{ij}w_ix_{ik} = \left(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}^\top\right)_{jk}$$

where again, $\boldsymbol{W} = \mathsf{diag}(w_1, w_2, \ldots, w_n)$ and $w_i = \frac{1}{\mathsf{Var}(Y_i)(\frac{\partial\eta_i}{\partial\mu_i})^2}$.

When is Fisher Scoring Equivalent to Newton Raphson?

Fisher Scoring is equivalent to Newton Raphson when the expected information matrix is equal to the observed information matrix. Recall:

$$I_{jk} = -(y_i-\mu_i)\frac{\partial}{\partial\beta_k}\left[w_i\left(\frac{\partial\eta_i}{\partial\mu_i}\right)x_{ij}\right] + x_{ij}w_ix_{ik}$$

Now examine:

$$\begin{aligned} w_i &= \frac{1}{\mathsf{Var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \\ &= \frac{1}{a_i(\phi) b''(\theta_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\left(\frac{\partial \mu_i}{\partial \eta_i}\right) && \text{since } \mathsf{Var}(Y_i) = b''(\theta_i) a_i(\phi) \\ &= \frac{1}{a_i(\phi)}\frac{\partial \theta_i}{\partial \mu_i}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\left(\frac{\partial \mu_i}{\partial \eta_i}\right) && \text{and } b''(\theta_i) = \frac{\partial b'(\theta_i)}{\partial \theta_i} = \frac{\partial \mu_i}{\partial \theta_i} \\ &= \frac{1}{a_i(\phi)}\frac{\partial \mu_i}{\partial \eta_i} && \text{under the canonical link } \theta_i = \eta_i \end{aligned}$$

So under the canonical link:

$$\begin{aligned} \frac{\partial}{\partial \beta_k}\left[w_i\left(\frac{\partial \eta_i}{\partial \mu_i}\right)x_{ij}\right] &= \frac{\partial}{\partial \beta_k}\left[\frac{1}{a_i(\phi)}\frac{\partial \mu_i}{\partial \eta_i}\left(\frac{\partial \eta_i}{\partial \mu_i}\right)x_{ij}\right] \\ &= \frac{\partial}{\partial \beta_k}\left(\frac{x_{ij}}{a_i(\phi)}\right) \\ &= 0 \end{aligned}$$

We then have:

$$I_{jk} = -(y_i - \mu_i)\underbrace{\frac{\partial}{\partial \beta_k}\left[w_i\left(\frac{\partial \eta_i}{\partial \mu_i}\right)x_{ij}\right]}_{=0} + x_{ij} w_i x_{ik} = x_{ij} w_i x_{ik} = \mathcal{I}_{jk}$$

Therefore under the canonical link, the expected information matrix equals the observed information matrix and Fisher Scoring is equivalent to Newton Raphson.

Iteratively Reweighted Least Squares (IRWLS)

Why is this called the iteratively reweighted least squares? The Fisher Scoring update equation:

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

can actually be rewritten as:

$$\hat{\boldsymbol{\beta}}^{(r+1)} = \Big[\boldsymbol{X}\boldsymbol{W}\big(\hat{\boldsymbol{\beta}}^{(r)}\big)\boldsymbol{X}^\top\Big]^{-1}\boldsymbol{X}\boldsymbol{W}\big(\hat{\boldsymbol{\beta}}^{(r)}\big)\boldsymbol{z}\big(\hat{\boldsymbol{\beta}}^{(r)}\big)$$

- See manipulation in Section 1.2.3 of course notes with: $\boldsymbol{z} = \boldsymbol{\eta} + (\boldsymbol{y} - \boldsymbol{\mu}) \circ \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}}$.
- Same form as weighted LS estimate of $\boldsymbol{\beta}$ with dependent variable $\boldsymbol{z}\big(\hat{\boldsymbol{\beta}}^{(r)}\big)$ and weight matrix $\boldsymbol{W}\big(\hat{\boldsymbol{\beta}}^{(r)}\big)$.

Summary

- 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 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: Assignment 1 & Chapter 1 review problems.

Example: The Poisson Distribution (Problem 1.4)

Let Y_i, $i = 1, \ldots, n$ be independent Poisson random variables with $\mathbb{E}[Y_i] = \lambda_i$. Suppose that associated with each y_i is a $p \times 1$ vector of explanatory variables $\boldsymbol{x}_i$. A Poisson regression model with the canonical link takes the form:

$$\mathsf{log}(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i(p-1)} = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

To answer the following you may either calculate the derivatives using standard methods, or use the general results derived in class for the exponential family.

a. Write down the score vector for the regression coefficients $\boldsymbol{\beta}$.

b. Write down the observed and expected information matrix for $\boldsymbol{\beta}$. Are they the same or different? Why?

c. What is the form of the weight function? What types of observations will have the largest and smallest weights?

Topic 2a: Binary Data: Estimation of the Odds Ratio

1. Definition of the Odds Ratio as a measure of association.
2. Likelihood based estimation of the Odds Ratio.
3. Inference for the Odds Ratio (Wald based confidence interval).
4. Example: Prenatal Care.

2.1 Introduction to the Analysis of Binary Data

- Outcome/Response: Binary (yes/no, diseased/healthy).
- Explanatory Variable: Binary (yes/no, treatment/control).
- Use a 2×2 table to summarize the data:

	Disease		
	Present	Absent	
Treatment	y_1	$m_1 - y_1$	m_1
Control	y_2	$m_2 - y_2$	m_2
	$y_\bullet$	$m_\bullet - y_\bullet$	$m_\bullet$

- Treat m_1 and m_2 as fixed.
- Assume Y_k are independent binomial random variables:

$$Y_k \sim \text{BIN}(m_k, \pi_k) \qquad \text{where } 0 < \pi_k < 1 \text{ for } k = 1, 2$$

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

Definition: Odds

How do we measure the association between treatment and response?

Definition

The Odds is the ratio of the probability that an event occurs (π) to the probability that it does not occur:

$$\text{Odds} = \frac{\pi}{1-\pi}$$

The odds is a one-to-one monotonically increasing function of π which takes on values on the non-negative real line.

Measures of Association

Definition

The Odds Ratio is the ratio of the odds of an event occurring in one group to the odds of the event occurring in another group:

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

Definition

The Relative Risk 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), then:

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

- This can be seen by noting that:

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

- Interpretation of OR:

$$\begin{array}{lcll} \pi_1 = \pi_2 & \implies & \text{OR} = 1 & \implies \text{equal risk} \\ \pi_1 > \pi_2 & \implies & \text{OR} > 1 & \implies \text{higher risk in group 1} \\ \pi_1 < \pi_2 & \implies & 0 < \text{OR} < 1 & \implies \text{higher risk in group 2} \end{array}$$

Odds Ratio Example Calculations

- $\pi_1 = 0.50$, $\pi_2 = 0.25$, so RR $= 0.50/0.25 = 2$ and OR $= \big(0.50/0.50\big)/\big(0.25/0.75\big) = 3$.
- $\pi_1 = 0.10$, $\pi_2 = 0.05$, so RR $= 0.10/0.05 = 2$ and OR $= \big(0.10/0.90\big)/\big(0.05/0.95\big) = 2.11$.
- $\pi_1 = 0.25$, $\pi_2 = 0.10$, so RR $= 0.25/0.10 = 2.5$ and OR $= \big(0.25/0.75\big)/\big(0.10/0.90\big) = 3$.

2.2 (Likelihood Based) Estimation of the Odds Ratio

- Goal: Use Likelihood Theory to estimate OR $= \psi = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$.

- Assumption: $Y_k \sim \text{BIN}(m_k, \pi_k)$, $k = 1, 2$ independently.

$$\begin{aligned}
\mathcal{L}(\pi_1, \pi_2) &= \mathbb{P}(Y_1 = y_1, Y_2 = y_2 \mid \pi_1, \pi_2) \\
&= \mathbb{P}(Y_1 = y_1 \mid \pi_1)\, \mathbb{P}(Y_2 = y_2 \mid \pi_2) \\
&= \binom{m_1}{y_1} \pi_1^{y_1} (1-\pi_1)^{m_1 - y_1} \binom{m_2}{y_2} \pi_2^{y_2} (1-\pi_2)^{m_2 - y_2} \\
&\propto \left(\frac{\pi_1}{1-\pi_1}\right)^{y_1} (1-\pi_1)^{m_1} \left(\frac{\pi_2}{1-\pi_2}\right)^{y_2} (1-\pi_2)^{m_2} \left(\frac{\pi_2/(1-\pi_2)}{\pi_2/(1-\pi_2)}\right)^{y_1} \\
&\propto \left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right)^{y_1} \left(\frac{\pi_2}{1-\pi_2}\right)^{y_2+y_1} (1-\pi_1)^{m_1} (1-\pi_2)^{m_2}
\end{aligned}$$

Estimation of the Odds Ratio

- Since we want to estimate ψ, we can reparameterize using:

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

- Note that $\pi_1, \pi_2 \in (0, 1)$ but $\theta_1, \theta_2 \in (-\infty, \infty)$.
- Our reparameterization implies:

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

- Now the likelihood becomes:

$$\begin{aligned}
\mathcal{L}(\pi_1, \pi_2) &\propto \left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right)^{y_1} \left(\frac{\pi_2}{1-\pi_2}\right)^{y_2+y_1} (1-\pi_1)^{m_1} (1-\pi_2)^{m_2} \\
\mathcal{L}(\theta_1, \theta_2) &= (e^{\theta_1})^{y_1} (e^{\theta_2})^{y_1+y_2} (1 + e^{\theta_1+\theta_2})^{-m_1} (1 + e^{\theta_2})^{-m_2}
\end{aligned}$$

- Recall our goal was to estimate OR $= \psi = e^{\theta_1}$.

$$\begin{aligned}
\mathcal{L}(\theta_1, \theta_2) &= (e^{\theta_1})^{y_1} (e^{\theta_2})^{y_1+y_2} (1 + e^{\theta_1+\theta_2})^{-m_1} (1 + e^{\theta_2})^{-m_2} \\
\ell(\theta_1, \theta_2) &= y_1\theta_1 + (y_1 + y_2)\theta_2 - m_1 \log(1 + e^{\theta_1+\theta_2}) - m_2 \log(1 + e^{\theta_2}) \\
S_1(\theta_1, \theta_2) &= y_1 - m_1 \left(\frac{e^{\theta_1+\theta_2}}{1 + e^{\theta_1+\theta_2}}\right) \\
S_2(\theta_1, \theta_2) &= y_1 + y_2 - m_1 \left(\frac{e^{\theta_1+\theta_2}}{1 + e^{\theta_1+\theta_2}}\right) - m_2 \left(\frac{e^{\theta_2}}{1 + e^{\theta_2}}\right)
\end{aligned}$$

- 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:

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

Example: Prenatal Care Data from Two Clinics

Consider the data below describing the relationship between the level of prenatal care and fetal mortality.

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

$$\hat{\theta}_1 = \log\left(\frac{y_1/(m_1-y_1)}{y_2/(m_2-y_2)}\right) = \log\left(\frac{y_1(m_2-y_2)}{y_2(m_1-y_1)}\right) = \log\left(\frac{(20)(373)}{(46)(316)}\right) = -0.6670729$$

$$\widehat{\text{OR}} = \hat{\psi} = \frac{y_1(m_2-y_2)}{y_2(m_1-y_1)} = \frac{(20)(373)}{(46)(316)} = 0.5132086$$

Inference for the Odds Ratio

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

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

- Differentiating we have:

$$I_{11} = m_1\left(\frac{e^{\theta_1+\theta_2}}{(1+e^{\theta_1+\theta_2})^2}\right) = m_1\pi_1(1-\pi_1)$$

$$I_{12} = I_{21} = m_1\left(\frac{e^{\theta_1+\theta_2}}{(1+e^{\theta_1+\theta_2})^2}\right) = m_1\pi_1(1-\pi_1)$$

$$I_{22} = m_1\left(\frac{e^{\theta_1+\theta_2}}{(1+e^{\theta_1+\theta_2})^2}\right) = m_1\pi_1(1-\pi_1) + m_2\left(\frac{e^{\theta_2}}{(1+e^{\theta_2})^2}\right) = m_1\pi_1(1-\pi_1) + m_2\pi_2(1-\pi_2)$$

Asymptotic Distribution of a Multidimensional MLE (A.3)

- We are interested in doing inference on $\theta_1 = \log(\psi)$ while θ_2 can be viewed as a nuisance parameter.
- Recall the Wald Result for a scalar parameter θ is $(\hat{\theta}-\theta)I(\hat{\theta}) \sim \chi^2_1$.

Wald Result for a scalar parameter from a vector

For the vector $\boldsymbol{\theta} = (\theta_1,\theta_2)^\top$, where θ_1 is a scalar parameter of interest:

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

asymptotically, where I^{11} is the $(1,1)$ element of $\boldsymbol{I}^{-1}(\hat{\theta}_1,\hat{\theta}_2)$ (i.e., the inverse of the information at the MLE) given by:

$$I^{11} = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}$$

- General result $\boldsymbol{I}$ is a $p \times p$ partitioned matrix.
  - Information Matrix:

$$\boldsymbol{I}(\theta_1,\theta_2) = \begin{bmatrix} \boldsymbol{I}_{11} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_{21} & \boldsymbol{I}_{22} \end{bmatrix}$$

- Inverse Information Matrix:

$$\boldsymbol{I}^{-1}(\theta_1, \theta_2) = \begin{bmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{bmatrix}$$

where $I^{11} = (\boldsymbol{I}_{11} - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\boldsymbol{I}_{21})^{-1}$.

- Consider the 2×2 matrix case:

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \implies \boldsymbol{A}^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} A^{11} & A^{12} \\ A^{21} & A^{22} \end{bmatrix}$$

where $A^{11} = \frac{d}{ad-bc} = \frac{1}{a-(bc)/d} = (a - bd^{-1}c)^{-1}$.

Confidence Interval for the Odds Ratio

- We will use this result to find the confidence interval for $\theta_1 = \log(\psi)$.
- First, we need to find $I^{11}(\theta_1, \theta_2)$.

$$\begin{aligned} I^{11} &= (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1} \\ &= \left(m_1\pi_1(1-\pi_1) - \frac{\big(m_1\pi_1(1-\pi_1)\big)^2}{m_1\pi_1(1-\pi_1) + m_2\pi_2(1-\pi_2)}\right)^{-1} \\ &= \left(\frac{m_1\pi_1(1-\pi_1)m_2\pi_2(1-\pi_2)}{m_1\pi_1(1-\pi_1) + m_2\pi_2(1-\pi_2)}\right)^{-1} \\ &= \frac{1}{m_1\pi_1(1-\pi_1)} + \frac{1}{m_2\pi_2(1-\pi_2)} \\ &= \frac{1}{m_1\pi_1} + \frac{1}{m_1(1-\pi_1)} + \frac{1}{m_2\pi_2} + \frac{1}{m_2(1-\pi_2)} \end{aligned}$$

Confidence Interval for the Odds Ratio

- Now we can calculate $I^{11}(\hat{\theta}_1, \hat{\theta}_2)$ using the invariance property of MLEs:

$$\begin{aligned} I^{11}(\hat{\theta}_1, \hat{\theta}_2) &= \frac{1}{m_1\hat{\pi}_1} + \frac{1}{m_1(1-\hat{\pi}_1)} + \frac{1}{m_2\hat{\pi}_2} + \frac{1}{m_2(1-\hat{\pi}_2)} \\ &= \frac{1}{y_1} + \frac{1}{m_1 - y_1} + \frac{1}{y_2} + \frac{1}{m_2 - y_2} \end{aligned}$$

- 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 Data from Two Clinics

Example: Prenatal Care Data from Two Clinics

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

$$I^{11}(\hat{\theta}_1, \hat{\theta}_2) = \mathsf{Var}(\hat{\theta}_1) = \frac{1}{20} + \frac{1}{316} + \frac{1}{46} + \frac{1}{373} = 0.07758465$$

$95\,\%$ confidence interval for $\theta_1 = \mathsf{log}(\psi)$:

$$\hat{\theta}_1 \pm 1.96\sqrt{I^{11}(\hat{\theta}_1, \hat{\theta}_2)} = -0.6671 \pm 1.96\sqrt{0.07758} = (-1.2130, -0.1211)$$

$95\,\%$ confidence interval for the Odds Ratio ψ:

$$\mathsf{exp}\left\{\hat{\theta}_1 \pm 1.96\sqrt{I^{11}(\hat{\theta}_1, \hat{\theta}_2)}\right\} = \mathsf{exp}\{-1.2130, -0.1211\} = (0.2973, 0.8859)$$

- Outcome: Fetal death vs Survival.
- Explanatory Variable: Level of Care: Intensive vs Regular.
  - Using results from the previous section we have: $\hat{\psi} = 0.51$, and a $95\,\%$ confidence interval for ψ was $(0.30, 0.89)$.
- 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

- $\hat{\psi}_{\mathrm{A}} = 0.80$, and a $95\,\%$ confidence interval for ψ_{A} is $(0.37, 1.73)$.
- $\hat{\psi}_{\mathrm{B}} = 1.01$, and a $95\,\%$ confidence interval for ψ_{B} is $(0.33, 3.10)$.
- 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

$\hat{\psi} = 14.06$, and a $95\,\%$ confidence interval for ψ is $(9.12, 21.76)$.

Association Between Clinic and Mortality

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

$\hat{\psi} = 0.35$, and a 95 % confidence interval for ψ is $(0.21, 0.58)$.

- The initial strong association between Level of Care and Fetal Mortality ($\hat{\psi} = 0.51$) disappeared when we stratified by clinic ($\hat{\psi}_{\text{A}} = 0.80$ and $\hat{\psi}_{\text{B}} = 1.01$).
- Instead of having to examine multiple 2×2 tables we'd like to estimate the OR and compute associations using a regression model.
- I.e., OR for the association between Level of Care and Mortality adjusted for Clinic.
- One way to do this by fitting a Binomial GLM to the data.

2.3 Multiple Regression (GLM) for Binary Responses

- Our previous derivations held for a binary response with a single binary explanatory variable.
- More often we need multiple regression methodology since we may:
  a. Want to be able to control for confounding variables and hence want to examine the effect of several (possibly related collinear) variables simultaneously.
  b. Want to examine the effect of categorical covariates (> 2 levels) or continuous covariates.
  c. Want to develop sophisticated models that describe complex relationships.

WEEK 4
0927 to 1st October

Topic 2b: Binomial (Logistic) Regression Models

2.4 Setting Up a Binomial Regression Model

1. Introduction and Notation.
2. Interpretation of β from logistic regression models as log odds ratios.
3. Logistic regression analysis of to the Prenatal Care example.
   - R Data and Code for fitting GLMs.
   - Hypothesis tests for β_k.
   - Confidence Intervals for the OR $\exp\{\beta_k\}$.

Introduction and Notation

- Outcome/Response variable: $Y_i \sim \text{BIN}(m_i, \pi_i)$, $i = 1, 2, \ldots, n$ independently.
- Explanatory variables: $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{i(p-1)})^\top$ with $x_{i0} = 1$.
- Regression parameters: $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$.
- Linear predictor: $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{i(p-1)}$.

- Recall multiple linear regression ($Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$):

$$\mathbb{E}[Y_i] = \mu_i = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

- Now with the Binomial data this would suggest we use:

$$\mathbb{E}\left[\frac{Y_i}{m_i}\right] = \pi_i = \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$

- But this is a bad idea because $0 < \pi_i < 1$ and we'd have to do constrained maximization to find $\hat{\pi}_i$.
- We want a link function: $g(\pi_i) = g(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ that maps:

$$g\colon (0,1) \to (-\infty, \infty)$$

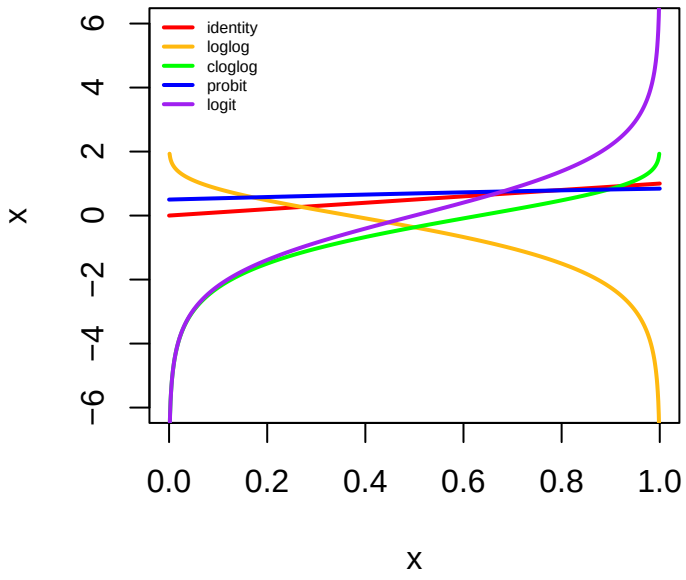
- Here are some link functions we might consider:

Identity	$g(\pi_i) = \pi_i$
log-log	$g(\pi_i) = \log\big(-\log(\pi_i)\big)$
complementary log-log	$g(\pi_i) = \log\big(-\log(1-\pi_i)\big)$
Probit†	$g(\pi_i) = \Phi^{-1}(\pi_i)$
Logit*	$g(\pi_i) = \log\big(\pi_i/(1-\pi_i)\big)$

†: Φ is the cdf for a standard normal random variable.

*: the canonical link for the Binomial (see Chapter 1).

Link Functions for the Binomial Distribution



The Logit Link and Odds Ratios

- The Logit link is the canonical link for the Binomial (see Chapter 1).
- This leads us to a Logistic Regression Model:

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

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

$$\text{logit}(a) = \log\left(\frac{a}{1-a}\right) = b \iff a = \frac{\exp\{b\}}{1+\exp\{b\}} = \text{expit}(b)$$

- Next: What is the interpretation of the $\boldsymbol{\beta}$ parameters in this model?

Simple Logistic Regression

- Consider a simple case of a binomial outcome $Y_i \sim \text{BIN}(m_i, \pi_i)$ for $i = 0, 1$ and a single binary explanatory variable:

$$x_{i1} = \begin{cases} 0 & \text{group 0} \\ 1 & \text{group 1} \end{cases}$$

- The simple logistic regression model equation is:

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

- When $x_{i1} = 0$ for $i = 0$, the model becomes:

$$\log\left(\frac{\pi_0}{1-\pi_0}\right) = \beta_0 + \beta_1(0) = \beta_0$$

- $\beta_0 =$ log odds of response for subjects with $x_{i1} = 0$.
- Now let's compare the model with $x_{i1} = 1$ versus $x_{i1} = 0$.

Group	$(1, x_{i1})^\top$	η_i	$= \log(\pi_i/(1-\pi_i))$
1	$(1,1)^\top$	$\beta_0 + \beta_1$	$= \log(\pi_1/(1-\pi_1))$
0	$(1,0)^\top$	β_0	$= \log(\pi_0/(1-\pi_0))$
		β_1	$= \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}\right)$

- 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 $x_{i1} = 0$.

Logistic Regression Models for Prenatal Care Example

- Response = Fetal mortality

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

- Explanatory Variables:

$$x_{i1} = \begin{cases} 1 & \text{Clinic A} \\ 0 & \text{Clinic B} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{Intensive level of care} \\ 0 & \text{Regular level of care} \end{cases}$$

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

- We will use the context of this example to interpret regression parameters from multiple logistic regression models.
- See Section 2.4.2 for general interpretations.

Model 1: Clinic only model

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

Clinic	Level of Care	$(1, x_{i1})^\top$	$\log(\pi_i/(1-\pi_i))$
A	—	$(1,1)^\top$	$\beta_0 + \beta_1$
B	—	$(1,0)^\top$	β_0

Table 3: Clinic only model

- β_0 is the log odds of infant mortality for babies born to mothers treated at Clinic B.
- β_1 is the log odds ratio of mortality for babies born to mothers treated at Clinic A versus Clinic B.

Model 2: Main effects model

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

Clinic	Level of Care	$(1, x_{i1}, x_{i2})^\top$	$\log(\pi_i/(1-\pi_i))$
A	Intensive	$(1,1,1)^\top$	$\beta_0 + \beta_1 + \beta_2$
A	Regular	$(1,1,0)^\top$	$\beta_0 + \beta_1$
B	Intensive	$(1,0,1)^\top$	$\beta_0 + \beta_2$
B	Regular	$(1,0,0)^\top$	β_0

Table 4: Main effects model

- β_0 is the log odds of infant mortality for babies born to mothers treated at Clinic B with Regular care.
- β_1 is the log odds ratio of mortality for babies born to mothers treated at Clinic A versus Clinic B at the same level of care.
- β_2 is the log odds ratio of mortality for babies born to mothers treated with Intensive versus Regular care at the same clinic (*OR of interest*).

Model 3: Interaction model

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

Clinic	Level of Care	$(1, x_{i1}, x_{i2}, x_{i3})^\top$	$\log(\pi_i/(1-\pi_i))$
A	Intensive	$(1,1,1,1)^\top$	$\beta_0+\beta_1+\beta_2+\beta_3$
A	Regular	$(1,1,0,0)^\top$	$\beta_0+\beta_1$
B	Intensive	$(1,0,1,0)^\top$	$\beta_0+\beta_2$
B	Regular	$(1,0,0,0)^\top$	β_0

Table 5: Interaction model

- β_1 is the log odds ratio of mortality for babies born to mothers treated at Clinic A versus Clinic B at Regular care.
- $\beta_1+\beta_3$ is the log odds ratio of mortality for babies born to mothers treated at Clinic A versus Clinic B at Intensive care.
- β_2 is the log odds ratio of mortality for babies born to mothers treated with Intensive versus Regular care at Clinic B.
- $\beta_2+\beta_3$ is the log odds ratio of mortality for babies born to mothers treated with Intensive versus Regular care at Clinic A.
- β_3 is a difference is log ratio odds.
- If $\beta_3 = 0$, then the association between mortality and level of care does not depend on Clinic.
- Equivalently, if $\beta_3 = 0$, then the association between mortality and Clinic does not depend on level of care.

Prediction from Logistic Regression

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i \iff \pi_i = \frac{\exp\{\eta_i\}}{1+\exp\{\eta_i\}} = \text{expit}(\eta_i)$$

- Assume we have found $\hat{\boldsymbol{\beta}}$ using Fisher scoring (R `glm()` function).
- The fitted value for the probability of response $\pi_i = \mathbb{E}[Y_i/m_i]$ for explanatory variable(s) $\boldsymbol{x}_i$ is:

$$\hat{\pi}_i = \hat{\pi}(\boldsymbol{x}_i) = \frac{\exp\{\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}\}}{1+\exp\{\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}\}} = \text{expit}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})$$

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

Logistic Regression Analysis of Prenatal Care Data

- Previously: Analysis using likelihood for 2×2 tables:

Odds Ratio (outcome = mortality)	Estimate and 95 % CI
Intensive vs Regular	$0.51\ (0.30, 0.89)$
Intensive vs Regular at Clinic A	$0.80\ (0.37, 1.73)$
Intensive vs Regular at Clinic B	$1.01\ (0.33, 3.10)$

- Now:
  1. Use `glm()` function in R to fit logistic regression models and estimate $\hat{\beta}$.
  2. Extract estimates β_k with log(OR) interpretations.
  3. Conduct hypothesis tests for H_0: $\beta_k = \beta_{k0}$.
  4. Calculate $95\,\%$ confidence intervals for β_k and $\psi = \exp\{\beta_k\}$.
  5. Try to find best fitting model with fewest parameters.

R Data and Code

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.
  - $x_{i1} =$ `clinic` $= \mathbb{I}\{\text{Clinic A}\}$.
  - $x_{i2} =$ `loc` $= \mathbb{I}\{\text{Intensive care}\}$.
- The response variable `y` is the number of events (deaths).
- `m` is the number of binomial trials (number of mothers).

```r
# R program for analysis of prenatal care data
prenatal.dat <- read.table("prenatal.dat", header = T)
# here we construct the response variable for the logistic
# regression analysis
prenatal.dat$resp <- cbind(prenatal.dat$y, prenatal.dat$m - prenatal.dat$y)
prenatal.dat
# now we fit the model using the glm function and store the
# result in 'model 1' we indicate 'resp' contains a
# binomial response and that we are using the logistic link
# function
model1 <- glm(resp ~ loc, family = binomial(link = logit), data = prenatal.dat)
summary(model1)
# the 'names' function lists the contents of the object
# 'model1' and following this statement we examine some of
# the contents of these objects (try it)
names(model1)
model1$family
model1$formula
model1$coefficients
model1$deviance
model1$fitted.values
model1$residuals
# now we fit a model to examine the relationship between
```

```
# level of care and mortality adjusting for clinic
model2 <- glm(resp ~ clinic + loc, family = binomial(link = logit),
  data = prenatal.dat)
summary(model2)
# here we examine whether the association between loc and
# mortality depends on the clinic
model3 <- glm(resp ~ loc + clinic + loc * clinic, family = binomial(link = logit),
  data = prenatal.dat)
summary(model3)
# now we examine the marginal relationship between
# mortality and clinic
model4 <- glm(resp ~ clinic, family = binomial(link = logit),
  data = prenatal.dat)
summary(model4)
```

Selected R Output

Print the augmented dataframe to see what the resp variable $(Y_i, m_i - Y_i)$ looks like:

```
# here we construct the response variable 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:

```
# now we fit the model using the glm function and store the
# result in 'model 1' we indicate 'resp' contains a
# binomial response and that we are using the logistic link
# function
model1 <- glm(resp ~ loc, family = binomial(link = logit), data = prenatal.dat)
summary(model1)
```

Fit of Model 1: Level of Care Model

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

```
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 $\hat{\beta}_k$.
- `Std Error`: estimated standard errors based on the inverse of the information.

$$\text{se}(\hat{\beta}_k) = \sqrt{\left(\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})\right)_{kk}} = \sqrt{I^{kk}(\hat{\boldsymbol{\beta}})}$$

- `z value`: Wald-based test statistics for the hypothesis test:

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

- `Pr(>|z|)`: p-value for the above test.

For this model:

- β_2 is the log odds ratio of mortality for babies born to mothers treated with Intensive versus Regular care.

$$\hat{\psi} = \exp\{\hat{\beta}_2\} = \exp\{-0.6670729\} = 0.51$$

Hypothesis test for β_k

- We may wish to test:

$$H_0\text{: } \beta_k = \beta_{k0} \text{ versus } H_A\text{: } \beta_k \neq \beta_{k0}$$

- The general Wald Result for scalar β_k is:

$$(\hat{\beta}_k - \beta_{k0})^2 \left(I^{kk}(\hat{\boldsymbol{\beta}})\right)^{-1} \sim \chi^2_1$$

  equivalently $\dfrac{\hat{\beta}_k - \beta_{k0}}{\text{se}(\hat{\beta}_k)} \sim \mathcal{N}(0,1)$ where $\text{se}(\hat{\beta}_k) = \sqrt{I^{kk}(\hat{\boldsymbol{\beta}})}$.

- And we can find the p-value of this test using

$$p = 2\,\mathbb{P}\left(Z > \frac{|\hat{\beta}_k - \beta_{k0}|}{\text{se}(\hat{\beta}_k)}\right) \qquad \text{where } Z \sim \mathcal{N}(0,1)$$

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

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

Hypothesis test for β_2 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:

$$H_0\text{: } \beta_2 = 0 \text{ vs } H_A\text{: } \beta_2 \neq 0$$

- The Wald-based test statistic is:

$$z^\star = \frac{\hat{\beta}_2 - 0}{\mathsf{se}(\hat{\beta}_2)} = \frac{-0.6671}{0.2785} = -2.3949$$

- And we can find the p-value of this test using:

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

- Therefore, we reject the null hypothesis that $\beta_2 = 0$.
- Equivalently, we reject the null hypothesis that OR $= 1$.

Confidence Interval for the OR

- Calculate CI for $\beta_k = \log(\psi)$ and then exponentiate.
- Recall the Wald-based confidence interval:

$$\hat{\beta}_k \pm 1.96\,\mathsf{se}(\hat{\beta}_k)$$

- The `Std Error` from the `summary()` output is the square root of the diagonal of the inverse of the Information matrix.

```
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

summary(model1)$cov.unscaled  # The inverse of the Information Matrix

            (Intercept)         loc
(Intercept)  0.02442007 -0.02442007
loc         -0.02442007  0.07758452

sqrt(diag(summary(model1)$cov.unscaled))  # The se of the betas

(Intercept)         loc
  0.1562692   0.2785400
```

Confidence Interval for exp$\{\beta_2\}$ 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
```

- The 95 % confidence interval for the OR is:

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

- Note: The estimate and $95\,\%$ confidence interval here match those found previously from the 2×2 table analysis.

Fit of Model 2: Main Effects Model

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

```
model2 <- glm(resp ~ clinic + loc, 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
clinic      -0.9862793  0.3089322 -3.1925427 1.410261e-03
loc         -0.1503053  0.3301670 -0.4552402 6.489365e-01
```

- Odds Ratio for mortality for Intensive versus Regular care, controlling for clinic:

$$\exp\{\hat{\beta}_2\} = \exp\{-0.1503\} = 0.860$$

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
```

Interpretation of Model 3: Interaction Model

Clinic	Level of Care	$(1, x_{i1}, x_{i2}, x_{i3})^\top$	$\log(\pi_i/(1-\pi_i))$
A	Intensive	$(1,1,1,1)^\top$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$
A	Regular	$(1,1,0,0)^\top$	$\beta_0 + \beta_1$
B	Intensive	$(1,0,1,0)^\top$	$\beta_0 + \beta_2$
B	Regular	$(1,0,0,0)^\top$	β_0

- Odds Ratio for mortality for Intensive vs Regular care, Clinic A:

$$\exp\{\hat{\beta}_2 + \hat{\beta}_3\} = \exp\{0.007643 - 0.229650\} = 0.80 = \hat{\psi}_{\text{A}}$$

- Odds Ratio for mortality for Intensive vs Regular care, Clinic B:

$$\exp\{\hat{\beta}_2\} = \exp\{0.007643\} = 1.01 = \hat{\psi}_{\text{B}}$$

Fit of Model 4: Clinic Only Model

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

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

             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -1.756041  0.1756737 -9.996041 1.586117e-23
clinic      -1.062357  0.2621216 -4.052916 5.058308e-05
```

- Odds Ratio for mortality in Clinic A versus Clinic B:

$$\exp\{\hat{\beta}_1\} = \exp\{-1.0624\} = 0.35$$

Prenatal Care Wrap-up

- Model 4 provides the best fit to the data with the fewest parameters.
- However, the original research question was about the level of care therefore we select Model 2 as our final model.
- Odds Ratio for mortality for Intensive versus Regular care, controlling for clinic:

$$\exp\{\hat{\beta}_2\} = \exp\{-0.1503\} = 0.860$$

- Exercises:
  1. Conduct a formal hypothesis test of H_0: $\beta_2 = 0$ and confirm p-value in the R output.
  2. Show that the $95\,\%$ confidence interval for the OR is $(0.450, 1.643)$.
  3. Show that the Odds Ratio for mortality for Clinic B versus Clinic A, controlling for level of care is:

$$\exp\{-\hat{\beta}_1\} = \exp\{0.9863\} = 2.68$$

Topic 2c: Likelihood Ratio (Deviance) Tests

Major Developments From Last Topic: Logistic Regression Models

Binomial GLM / Logistic Regression Model

$Y_i \sim \text{BIN}(m_i, \pi_i)$, $i = 1, \ldots, n$ independently with explanatory variables $\boldsymbol{x}_i$:

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

- Estimation: $\hat{\boldsymbol{\beta}}$ come from Fisher Scoring using R function `glm()`.
- Interpretation: β_k have log OR interpretations ($k > 0$).
- Hypothesis Tests of H_0: $\beta_k = \beta_{k0}$ versus H_A: $\beta_k \neq \beta_{k0}$.
- Confidence Intervals: $\hat{\beta}_k \pm z_{1-\alpha/2}\,\text{se}(\hat{\beta}_k)$ where $\text{se}(\hat{\beta}_k) = \sqrt{I^{kk}(\hat{\boldsymbol{\beta}})}$.

Topic 2c: Logistic Regression: Likelihood Ratio (Deviance) Tests

1. Likelihood for Binary (Logistic) Regression:

$$\ell(\boldsymbol{\beta} \mid \boldsymbol{y}) = \sum_{i=1}^{n} \Big(y_i(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - m_i \log\big(1 + \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\}\big) \Big)$$

2. Likelihood Ratio Tests:

$$\begin{aligned} -2\log\big(R(\theta)\big) &= -2r(\theta) \sim \chi^2_1 && \text{Scalar } \theta \\ -2\log\big(R(\boldsymbol{\theta})\big) &= -2r(\boldsymbol{\theta}) \sim \chi^2_{n-p} && p\text{-dim Vector } \boldsymbol{\theta} \end{aligned}$$

3. Testing Nested Non-saturated Models:

$$H_0\text{: } \beta_p = \cdots = \beta_{q-1} = 0 \text{ vs } H_A\text{: at least one of } \beta_p, \ldots, \beta_{q-1} \neq 0$$

2.5 Likelihood for Binary (Logistic) Regression

- Outcome/Response variable: $Y_i \sim \text{BIN}(m_i, \pi_i)$, $i = 1, \ldots, n$.

$$\mathcal{L}(\boldsymbol{\pi} \mid \boldsymbol{y}) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i}$$

$$\ell(\boldsymbol{\pi} \mid \boldsymbol{y}) = \sum_{i=1}^{n} \big(y_i \log(\pi_i) + (m_i - y_i) \log(1 - \pi_i) \big)$$

- Explanatory variables: $\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{i(p-1)})^\top$.
- Regression parameters: $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$.
- Link function: logistic link

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

Likelihood for Logistic Regression

- Log likelihood for Binomial distribution:

$$\ell(\boldsymbol{\pi} \mid \boldsymbol{y}) = \sum_{i=1}^{n} \left(y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + m_i \log(1 - \pi_i) \right)$$

- Using logit link we can reparameterize the log-likelihood in terms of $\boldsymbol{\beta}$:

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

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

- Log likelihood for logistic regression:

$$\ell(\boldsymbol{\beta} \mid \boldsymbol{y}) = \sum_{i=1}^{n} \Big(y_i(\boldsymbol{x}_i^\top \boldsymbol{\beta}) - m_i \log\big(1 + \exp\{\boldsymbol{x}_i^\top \boldsymbol{\beta}\}\big) \Big)$$

- If $\dim(\boldsymbol{\beta}) = p < n$, the model is said to be unsaturated.
- If $p = n$, we have a saturated model.
- Maximize $\ell(\boldsymbol{\beta} \mid \boldsymbol{y})$ to obtain the MLE's $\hat{\boldsymbol{\beta}}$ (i.e., using Fisher Scoring in R `glm()`).
- Transform to get estimates $\hat{\pi}_i = \text{expit}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})$.
- For a saturated model, $\text{expit}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})$ will equal the Binomial MLE y_i/m_i, and we will have $m_i\hat{\pi}_i = y_i$ (i.e., a perfect fit).

Hypothesis Tests for Logistic Regression

- We want to ask: how good is the model?
- How well do the $m_i\hat{\pi}_i$ approximate the data y_i?
- How much worse is the fit of a particular unsaturated model versus the saturated model?
- Previously: Wald-based tests of:

$$H_0\text{: } \beta_k = \beta_{k0} \text{ versus } H_A\text{: } \beta_k \neq \beta_{k0}.$$

- Today: Likelihood Ratio based tests for:

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

- This will allow us to test the overall fit of nested models.

Likelihood Ratio Tests — General Setting

- Suppose $\mathcal{L}(\boldsymbol{\theta})$ is the likelihood of a q-dim parameter vector $\boldsymbol{\theta}$.
  - Let $\tilde{\boldsymbol{\theta}}$ be the q-dim MLE (unconstrained/saturated, $q = n$).
  - Let $\hat{\boldsymbol{\theta}}$ be the p-dim MLE (constrained/unsaturated, $p < q$).
- H_0: the unsaturated p-dim model is adequate.
- H_A: the p-dim model is not adequate.
- Recall the Likelihood Ratio result. Under H_0,

$$-2\log\big(R(\boldsymbol{\theta})\big) = -2\log\left(\frac{\mathcal{L}(\hat{\boldsymbol{\theta}})}{\mathcal{L}(\tilde{\boldsymbol{\theta}})}\right) = -2\big(\ell(\hat{\boldsymbol{\theta}}) - \ell(\tilde{\boldsymbol{\theta}})\big) \sim \chi^2_{q-p}$$

- This is often referred to as the Deviance: $D = -2\log\big(R(\boldsymbol{\theta})\big)$.
- Reject H_0 at significance level θ if:

$$\mathbb{P}\Big(\chi^2_{q-p} > -2\log\big(R(\boldsymbol{\theta})\big)\Big) < \alpha$$

Likelihood Ratio Tests — Logistic Regression

- Saturated model MLEs: $\tilde{\pi}_i = y_i/m_i$, $i = 1, \ldots, n$.
- Unsaturated model MLEs: $\hat{\pi} = \text{expit}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})$.
  - Regression models are a way of imposing constraints on the estimation of $\boldsymbol{\pi}$ (through p-dim $\boldsymbol{\beta}$).
- H_0: The p-dim model $\text{logit}(\pi_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ is adequate.
- H_A: The p-dim model is not adequate compared to the n-dim saturated model.
- For the binomial with logit link the LR/Deviance test statistic is:

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

- Aside: Note that D has the general form: $D = 2\sum\sum O_{ij}\log\left(\frac{O_{ij}}{E_{ij}}\right)$.

Likelihood Ratio Tests — Logistic Regression

- We expect $D \sim \chi^2_{n-p}$ under H_0.
  - Unfortunately, this is not a good approximation.
  - Approximation is much better for tested nested unsaturated models though.
- In R, the D is reported as the `Residual Deviance`.

```
model3$deviance

[1] -4.352074e-14

model4$deviance

[1] 0.3148411
```

- H_0: Model 4 is adequate, H_A: Model 4 is not adequate compared to Model 3.

$$p = \mathbb{P}(\chi^2_2 > 0.3148) = 0.85$$

- Therefore do not reject the null hypothesis that Model 4 (Clinic only) is adequate.

Pearson Statistic — Logistic Regression

- The Pearson statistic is another statistic one can use for assessing "overall" fit of a 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 D, $P \sim \chi^2_{n-p}$ under H_0: the model provides a reasonable fit to the data.
- Note: P has the general form $P = \sum \frac{(O_i - E_i)^2}{V_i}$.
- The χ^2 approximation is a bit better than for deviance statistics.
- Both are poor if sample size (m_i) are small though.

2.6 Testing Nested Non-saturated Models

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

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i(p-1)} \tag{1}$$
$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i(p-1)} + \cdots + \beta_{q-1}x_{i(q-1)} \tag{2}$$
$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1}x_{i(p-1)} + \cdots + \beta_{q-1}x_{i(q-1)} + \cdots + \beta_{n-1}x_{i(n-1)} \tag{3}$$

- Model (1) is nested within Model (2).
- H_0: The p-dim 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 $\beta_p, \ldots, \beta_{q-1} \neq 0$.

Testing Nested Non-saturated Models

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

- Previously, found the LR/Deviance test vs saturated models.
- LR/Deviance test of (1) vs (3):

$$D_0 = -2\big(\ell(\hat{\boldsymbol{\pi}}) - \ell(\tilde{\tilde{\boldsymbol{\pi}}})\big) \sim \chi^2_{n-p}$$

- LR/Deviance test of (2) vs (3):

$$D_A = -2\big(\ell(\tilde{\boldsymbol{\pi}}) - \ell(\tilde{\tilde{\boldsymbol{\pi}}})\big) \sim \chi^2_{n-q}$$

- Now we wish to conduct an LR/Deviance test of (1) vs (2):

$$\begin{aligned}\Delta D &= -2\big(\ell(\hat{\boldsymbol{\pi}}) - \ell(\tilde{\boldsymbol{\pi}})\big) \\ &= -2\big(\ell(\hat{\boldsymbol{\pi}}) - \ell(\tilde{\tilde{\boldsymbol{\pi}}})\big) + 2\big(\ell(\tilde{\boldsymbol{\pi}}) - \ell(\tilde{\tilde{\boldsymbol{\pi}}})\big) \\ &= D_0 - D_A\end{aligned}$$

- Fact: If $X_i \sim \chi^2_{r_i}$, then $X_1 + X_2 \sim \chi^2_{r_1+r_2}$.
- Therefore, under H_0: $\Delta D \sim \chi^2_{q-p}$.
  - This approximation is much better than when testing an unsaturated model versus the saturated model.
- If $p = \mathbb{P}(\chi^2_{q-p} > \Delta D) < \alpha$ then reject H_0: This implies:
  - Reduced model does not fit the data as well as Full model.
  - One or more of $x_{ip}, \ldots, x_{i(q-1)}$ is important (i.e., associated with the outcome).

Testing Non-Saturated Models in Prenatal Care Example

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

Model	Variables	Deviance	Parameters
1	`loc`	10.81438	2
2	`clinic + loc`	0.1069281	3
3	`clinic + loc + loc*clinic`	≈ 0	4
4	`clinic`	0.3148411	2

- Is level of care associated with fetal mortality?

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

```
# Deviance/LR test H_0: Model 4 is adequate compared to
# Model 2
model4$deviance - model2$deviance

[1] 0.207913

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

[1] 0.6484081

# Reprint the fitted Model 2: Main Effects Only model
model2 <- glm(resp ~ clinic + loc, family = binomial(link = logit),
  data = prenatal.dat)
summary(model2)


Call:
glm(formula = resp ~ clinic + loc, 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 ***
clinic       -0.9863     0.3089  -3.193  0.00141 **
loc          -0.1503     0.3302  -0.455  0.64894
---
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: Likelihood Ratio (Deviance) Tests for Logistic Regression

- (log) Relative Likelihood Result: $D = -2\log\big(R(\boldsymbol{\theta})\big) = -2r(\boldsymbol{\theta}) \sim \chi^2_{n-p}$.
- For the binomial with logit link the LR/Deviance test statistic is:

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

- This is reported as the "`Residual Deviance`" in R `glm` summary output.
- Used to test the fit of nested non-saturated models ($q > p$):

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

  - H_0: $\beta_p = \cdots = \beta_{q-1} = 0$.
  - H_A: at least one of $\beta_p, \ldots, \beta_{q-1} \neq 0$.
- Test statistic: $\Delta D \sim \chi^2_{q-p}$ and p-value $= \mathbb{P}(\chi^2_{q-p} > \Delta D)$.

WEEK 5
4th to 8th October

Topic 2d: Logistic Regression: Residuals & CIs

Binomial GLM / Logistic Regression Model

$Y_i \sim \text{BIN}(m_i, \pi_i)$, $i = 1, \ldots, n$ independently with explanatory variables $\boldsymbol{x}_i$:

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

- Estimation: $\hat{\boldsymbol{\beta}}$ come from Fisher scoring using R function `glm()`.
- Interpretation: β_k have log OR interpretations ($k > 0$).
- Wald based Hypothesis Tests of H_0: $\beta_k = \beta_{k0}$ versus H_A: $\beta_k \neq \beta_{k0}$.
- Confidence Intervals: $\hat{\beta}_k \pm z_{1-\alpha/2}\,\text{se}(\hat{\beta}_k)$ where $\text{se}(\hat{\beta}_k) = \sqrt{I^{kk}(\hat{\boldsymbol{\beta}})}$.
- Deviance/LR based Hypothesis Tests for nested models:

$$H_0\text{: } \beta_p = \cdots = \beta_{q-1} = 0 \text{ vs } H_A\text{: at least one of } \beta_p, \ldots, \beta_{q-1} \neq 0$$

  using $\Delta D = D_0 - D_A \sim \chi^2_{q-p}$ under H_0.

Topic 2d: Logistic Regression: Residuals & Confidence Intervals

1. Residuals for Binomial Data — Deviance Residuals:

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

2. Neuroblastoma Example:
   - Categorical explanatory variables: use of dummy variables.
   - Residual plots using Deviance Residuals.
   - Finding requested Odds Ratios.
3. Confidence Intervals for non-linear functions of η_i:
   - How to get a CI for $\exp\{\beta_2 - \beta_1\}$ or $\pi_i = \text{expit}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$.

Residuals for Normal Linear Regression Models

- The raw residuals were:

$$r_i = y_i - \hat{\mu}_i$$

- The standardized residuals were:

$$d_i = \frac{(y_i - \hat{\mu}_i)}{\hat{\sigma}\sqrt{1 - h_{ii}}} \sim t_{n-p} \to \mathcal{N}(0, 1)$$

- The overall fit of the model and appropriateness of its underlying assumptions can be assessed using various types of Residual Plots. For example:
  - Residuals versus covariate x_j (checks linearity assumption).
  - Residuals versus fitted values $\hat{\mu}_i$ (check normality and constant variance).
  - Normal QQ (quantile-quantile) plots of residuals (checks normality).

2.7 Residuals for Binomial Data — Deviance Residuals

- Recall the LR/Deviance test statistic:

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

- Define the Deviance Residuals to be:

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

- Under H_0: the model is adequate,

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

- Use the plots of the deviance residuals to assess whether the r_i^D looks like independent $\mathcal{N}(0, 1)$ observations.

Residuals for Binomial Data — Pearson Residuals

- Define the Pearson Residuals to be:

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

- Under H_0: the model is adequate, $r_i^P \sim \mathcal{N}(0,1)$.
- Note: if $m_i\hat{\pi}_i < 5$ for one or more i, we should be concerned about the validity of the approximation (χ^2 or $\mathcal{N}(0,1)$) and hence your conclusions (the same holds for $m_i(1-\hat{\pi}_i)$).

2.8 Estimation of Prognosis for Children with Neuroblastoma

Purpose of Study: To investigate the relationship between the probability of surviving 2 years free of disease following diagnosis and treatment for neuroblastoma, and age at diagnosis and stage of disease at diagnosis.

	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

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.

	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

Setting up the Regression Models

- Outcome: Let Y_i be the number of children in group i who survived 2 years out of m_i total children in group i. Assume $Y_i \sim \text{BIN}(m_i, \pi_i)$ independently $i = 1, \ldots, 15$.
- Explanatory variables: Use dummy variables to represent Age and Stage levels:

$$x_{i1} = \begin{cases} 1 & \text{if age 12-23 months} \\ 0 & \text{o.w.} \end{cases} \quad x_{i2} = \begin{cases} 1 & \text{if age 24+ months} \\ 0 & \text{o.w.} \end{cases} \quad x_{i3} = \begin{cases} 1 & \text{stage II} \\ 0 & \text{o.w.} \end{cases}$$

$$x_{i4} = \begin{cases} 1 & \text{if stage III} \\ 0 & \text{o.w.} \end{cases} \quad x_{i5} = \begin{cases} 1 & \text{if stage IV} \\ 0 & \text{o.w.} \end{cases} \quad x_{i6} = \begin{cases} 1 & \text{if stage V} \\ 0 & \text{o.w.} \end{cases}$$

- Now consider the models:

1. Age & Stage:

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

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

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

R Data and Code

Data file 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

- First line contains the variable labels.
- 15 observations.
- Recall **age** and **stage** are categorical (ordinal) variables, so we will have to compute the indicator variables in the R program.

```
# Read in the Neuroblastoma dataset
neuro.dat <- read.table("neuro.dat", header = T)
# here we create indicator variables for age and stage
neuro.dat$agef <- factor(neuro.dat$age)
neuro.dat$stagef <- factor(neuro.dat$stage)
# here we construct the response variable for logistic
# regression: (y, m-y)
neuro.dat$resp <- cbind(neuro.dat$y, neuro.dat$m - neuro.dat$y)
neuro.dat
# here we fit the model with age and stage and print out
# summary statistics
model1 <- glm(resp ~ agef + stagef, family = binomial(link = logit),
  data = neuro.dat)
summary(model1)
summary(model1, corr = T)$correlation
# record deviance residuals (rd1), linear predcitor (lp1),
# and fitted values (fv1)
rd1 <- residuals.glm(model1, "deviance")
lp1 <- model1$linear.predictors
```

```
fv1 <- model1$fitted.values
# here we compute the Pearson residual as an exercise
rp1 <- (neuro.dat$y - neuro.dat$m * fv1)/sqrt(neuro.dat$m * fv1 *
  (1 - fv1))
# here we verify that the fitted values agree with what we
# expect from the linear predictor
fv2 <- exp(lp1)/(1 + exp(lp1))
cbind(rd1, rp1, lp1, fv1, fv2)
# plotting the deviance and Pearson residuals
pdf("neuro-residual.pdf", height = 6, width = 8)
plot(fv1, rd1, ylim = c(-3, 3), xlab = "FITTED VALUES", ylab = "RESIDUALS",
  pch = 1)
points(fv1, rp1, pch = 2)
abline(h = -2, lty = 2)
abline(h = 2, lty = 2)
legend(0, 3, c("Deviance Residual", "Pearson Residual"), pch = c(1,
  2), bty = "n")
dev.off
# here we fit two reduced models to enable us to test the
# importance of age and stage
model2 <- glm(resp ~ agef, family = binomial(link = logit), data = neuro.dat)
summary(model2)
model3 <- glm(resp ~ stagef, family = binomial(link = logit),
  data = neuro.dat)
summary(model3)
```

Selected R Output

The final data object `neuro.dat` is given by:

```
neuro.dat

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

Here is the summary of model 1 including both Age & Stage:

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

```
model1 <- glm(resp ~ agef + stagef, family = binomial(link = logit),
  data = neuro.dat)
summary(model1)
```

```

Call:
glm(formula = resp ~ agef + stagef, 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 ***
agef2        -2.1181     0.5736  -3.693 0.000222 ***
agef3        -2.6130     0.5017  -5.208 1.91e-07 ***
stagef2      -1.2529     0.7837  -1.599 0.109860
stagef3      -1.7759     0.8003  -2.219 0.026478 *
stagef4      -4.3678     0.7902  -5.528 3.25e-08 ***
stagef5      -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.

```
rd1 <- residuals.glm(model1, "deviance")
rp1 <- (neuro.dat$y - neuro.dat$m * fv1)/sqrt(neuro.dat$m * fv1 *
  (1 - fv1))
lp1 <- model1$linear.predictors
fv1 <- model1$fitted.values
fv2 <- exp(lp1)/(1 + exp(lp1))
cbind(rd1, rp1, lp1, fv1, fv2)
```

```
           rd1         rp1         lp1        fv1        fv2
1  -0.77808711 -0.91184050  3.31753053 0.96502534 0.96502534
2   0.68559153  0.63381666  2.06460478 0.88741505 0.88741505
3  -1.47407847 -1.69888561  1.54162565 0.82370092 0.82370092
4   0.17884403  0.18019371 -1.05030016 0.25916747 0.25916747
5   0.63431439  0.58779486  2.29528941 0.90848616 0.90848616
6  -0.08658336 -0.08736144  1.19944586 0.76842619 0.76842619
7  -0.30801258 -0.30734393 -0.05347989 0.48663321 0.48663321
8   1.52114028  1.56325351 -0.57645902 0.35974778 0.35974778
9  -1.43545385 -1.02556686 -3.16838483 0.04037295 0.04037295
10 -0.73520283 -0.73328264  0.17720474 0.54418562 0.54418562
```

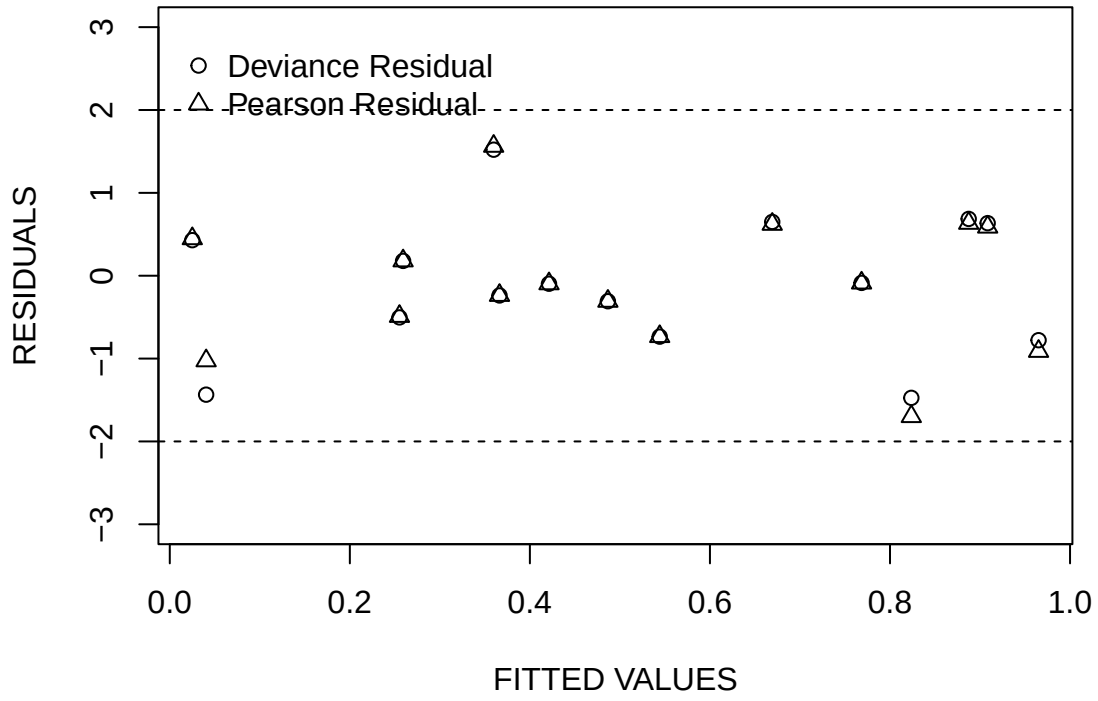


Figure 1: Plot of Residuals by Fitted Values for Neuroblastoma Model with Age and Stage

```
11  0.64949774  0.62163765  0.70456102 0.66919823 0.66919823
12 -0.23825133 -0.23663531 -0.54836473 0.36624389 0.36624389
13 -0.50305728 -0.48993834 -1.07134385 0.25514760 0.25514760
14  0.42894854  0.44782015 -3.66326967 0.02500712 0.02500712
15 -0.09643089 -0.09619454 -0.31768010 0.42124123 0.42124123
```

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).

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

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


Call:
glm(formula = resp ~ agef, 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 ***
agef2        -2.1119     0.4325  -4.883 1.05e-06 ***
agef3        -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
```

Now we fit the model excluding the age variable to examine the drop in the quality of fit from model one (with age and stage).

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

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


Call:
glm(formula = resp ~ stagef, 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 **
stagef2      -1.2657     0.7150  -1.770  0.07671 .
stagef3      -2.3224     0.7401  -3.138  0.00170 **
stagef4      -4.5643     0.7223  -6.319 2.63e-10 ***
stagef5      -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 consider testing nested models using Deviance/Likelihood Ratio Tests. Recall:

Model	Factors In Model	Deviance	p	$n-p$
1	Age + Stage	9.625	7	8
2	Age	83.583	3	12
3	Stage	42.446	5	10
4	Intercept only	162.832	1	14

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

where $\hat{\boldsymbol{\pi}}$ represents the MLEs from the reduced (nested) model and $\tilde{\boldsymbol{\pi}}$ are the MLEs from the full model.

Task #1: Pick the model that best represents the important associations between the outcome and explanatory variables.

1. Is Stage important?

H_0: $\beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ (Model 2 is adequate vs Model 1)
H_A: $\beta_3 \neq 0$ or $\beta_4 \neq 0$ or $\beta_5 \neq 0$ or $\beta_6 \neq 0$ (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(model2$deviance - model1$deviance, model2$df.residual -
  model1$df.residual)

[1] 3.330669e-15
```

Therefore we reject the null hypothesis that stage is unimportant.

2. Is Age important?

H_0: $\beta_1 = \beta_2 = 0$ (Model 3 is adequate vs Model 1)
H_A: $\beta_1 \neq 0$ or $\beta_2 \neq 0$ (Model 3 is not adequate)

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

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

[1] 7.464321e-08
```

Therefore we reject the null hypothesis that age is unimportant.

3. Do we need an Age$*$Stage interaction?

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

[1] 0.292341
```

So we select Model 1 for interpretation. Here's the fitted R `summary()` again for reference.

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

```
model1 <- glm(resp ~ agef + stagef, family = binomial(link = logit),
  data = neuro.dat)
summary(model1)


Call:
glm(formula = resp ~ agef + stagef, 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 ***
agef2        -2.1181     0.5736  -3.693 0.000222 ***
agef3        -2.6130     0.5017  -5.208 1.91e-07 ***
stagef2      -1.2529     0.7837  -1.599 0.109860
stagef3      -1.7759     0.8003  -2.219 0.026478 *
stagef4      -4.3678     0.7902  -5.528 3.25e-08 ***
stagef5      -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
```

Model Interpretation

Task #2: Interpret the selected model through estimated ORs.

A. What is the odds ratio of surviving two years for a patient with disease in stage IV versus stage I?

$$\widehat{\text{OR}} = \exp\{\hat{\beta}_5\} = \exp\{-4.368\} = 0.013$$

When controlling for age, the odds of surviving two years among those diagnosed in state IV is 0.013 times the odds among subjects diagnosed in stage I.

Age	Stage	$(1, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6})^\top$	$\log\big(\pi_i/(1-\pi_i)\big)$
NA	IV	$(1, x_{i1}, x_{i2}, 0, 0, 1, 0)^\top$	$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_5$
NA	I	$(1, x_{i1}, x_{i2}, 0, 0, 0, 0)^\top$	$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$

Age	Stage	$(1, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6})^\top$	$\log\big(\pi_i/(1-\pi_i)\big)$
24+	NA	$(1, 0, 1, x_{i3}, x_{i4}, x_{i5}, x_{i6})^\top$	$\beta_0 + \beta_2 + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}$
12-23	NA	$(1, 1, 0, x_{i3}, x_{i4}, x_{i5}, x_{i6})^\top$	$\beta_0 + \beta_1 + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}$

B. What is the odds ratio of surviving for a patient aged 24+ months versus 12-23 months?

$$\widehat{\text{OR}} = \exp\{\hat\beta_2 - \hat\beta_1\} = \exp\{-2.613 - (-2.118)\} = 0.61$$

When controlling for stage, the odds of surviving two years among those diagnosed at 24+ months of age is 0.61 times the odds of surviving two years among subjects diagnosed at 12-23 months of age.

Constructing Confidence Intervals

Task #3: Get Confidence Intervals for our estimated ORs.

A. $\text{OR} = \exp\{\beta_5\}$, so we can use a Wald-based CI,

$$\exp\big\{\hat\beta_5 \pm 1.96\,\text{se}(\hat\beta_5)\big\} = \exp\big\{-4.368 \pm 1.96(0.7902)\big\} = (0.003, 0.060)$$

B. What about a CI for $\text{OR} = \exp\{\beta_2 - \beta_1\}$?

- In order to calculate a CI we need to obtain $\text{se}(\hat\beta_2 - \hat\beta_1)$.
- This is not directly available from R `summary()`.

$$\text{Var}(aX + bY) = a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2ab\,\text{Cov}(X, Y)$$
$$\text{Var}(\hat\beta_2 - \hat\beta_1) = \text{Var}(\hat\beta_2) + \text{Var}(\hat\beta_1) - 2\,\text{Cov}(\hat\beta_2, \hat\beta_1)$$

- The covariance matrix ($\boldsymbol{I}^{-1}$) is available from `summary(model1)$cov.unscaled`.

2.9 Confidence Intervals for non-linear functions of η_i

Recall that since $\hat{\boldsymbol\beta}$ is an MLE, $\hat{\boldsymbol\beta} \sim \text{MVN}\big(\beta, \boldsymbol{I}^{-1}(\hat{\boldsymbol\beta})\big)$ approximately. This means that:

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

and

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

1. An approximate $95\,\%$ CI for $\eta_i = \boldsymbol{x}_i^\top \boldsymbol\beta$ is then given by:

$$\boldsymbol{x}_i^\top \hat{\boldsymbol\beta} \pm 1.96\sqrt{\boldsymbol{x}_i^\top \boldsymbol{I}^{-1}(\hat{\boldsymbol\beta})\boldsymbol{x}_i} = (\hat\eta_{\text{L}}, \hat\eta_{\text{U}})$$

2. If the OR of interest is expressed as $\exp\{\boldsymbol{c}^\top \boldsymbol\beta\}$ where $\boldsymbol{c}$ is a column vector defining the contrast of

the regression coefficients, then an approximate $95\,\%$ CI for this OR is:

$$\exp\left\{\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} \pm 1.96\sqrt{\boldsymbol{c}^\top \boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{c}}\right\}$$

3. An approximate $95\,\%$ CI for $\pi_i = \exp\{\boldsymbol{x}_i^\top\boldsymbol{\beta}\}/\big(1+\exp\{\boldsymbol{x}_i^\top\boldsymbol{\beta}\}\big) = \text{expit}(\boldsymbol{x}_i^\top\boldsymbol{\beta})$. is:

$$\Big(\text{expit}(\hat{\eta}_\text{L}), \text{expit}(\hat{\eta}_\text{U})\Big).$$

Back to the Neuroblastoma Example

B. Find a confidence interval for OR $= \exp\{\beta_2 - \beta_1\}$.

- The vector defining the contrast of interest is $\boldsymbol{c} = (0,-1,1,0,0,0,0)^\top$:

$$\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} = \begin{bmatrix}0 & -1 & 1 & 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_6\end{bmatrix} = \hat{\beta}_2 - \hat{\beta}_1$$

$$\begin{aligned}\boldsymbol{c}^\top\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{c} &= \begin{bmatrix}0 & -1 & 1 & 0 & 0 & 0 & 0\end{bmatrix}\begin{bmatrix} I^{00} & I^{01} & I^{02} & \dots & I^{0(p-1)}\\ I^{10} & I^{11} & I^{12} & \dots & I^{1(p-1)}\\ I^{21} & I^{21} & I^{22} & \dots & I^{2(p-1)}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ I^{(p-1)1} & I^{(p-1)1} & I^{(p-1)2} & \dots & I^{(p-1)(p-1)}\end{bmatrix}\begin{bmatrix}0\\-1\\1\\0\\0\\0\\0\end{bmatrix}\\ &= I^{11} + I^{22} - I^{12} - I^{21}\\ &= \text{Var}(\hat{\beta}_2) + \text{Var}(\hat{\beta}_1) - 2\,\text{Cov}(\hat{\beta}_2,\hat{\beta}_1)\end{aligned}$$

- The program used to compute the variance is as follows:

```
# use the summary.glm function and store the result in the
# tmp object
tmp <- summary.glm(model1)
# examine the contents of the tmp objects and store the
# covariance matrix in v
names(tmp)
v <- tmp$cov.unscaled
v
# create x vector to get contrast of regression
# coefficients
x <- c(0, -1, 1, 0, 0, 0, 0)
x <- as.matrix(x, 7, 1)
dim(x)
# compute the variance estimate of difference in estimates
# for age parameters
t(x) %*% v %*% x
```

- The resulting output is as follows.

```r
# use the summary.glm function and store the result in the
# tmp object
tmp <- summary.glm(model1)
# examine the contents of the tmp objects and store the
# covariance matrix in v
names(tmp)
```

```
 [1] "call"           "terms"          "family"         "deviance"
 [5] "aic"            "contrasts"      "df.residual"    "null.deviance"
 [9] "df.null"        "iter"           "deviance.resid" "coefficients"
[13] "aliased"        "dispersion"     "df"             "cov.unscaled"
[17] "cov.scaled"
```

```r
v <- tmp$cov.unscaled
v
```

```
            (Intercept)       agef2       agef3     stagef2     stagef3
(Intercept)   0.5960933 -0.18392302 -0.17583962 -0.46378468 -0.43679598
agef2        -0.1839230  0.32901185  0.15792411  0.01837858 -0.01638010
agef3        -0.1758396  0.15792411  0.25170695  0.01648743 -0.01538075
stagef2      -0.4637847  0.01837858  0.01648743  0.61411717  0.44845241
stagef3      -0.4367960 -0.01638010 -0.01538075  0.44845241  0.64043301
stagef4      -0.5098913  0.08279176  0.06769368  0.45553801  0.44437841
stagef5      -0.4969598  0.06047881  0.05606300  0.45425149  0.44552862
                stagef4     stagef5
(Intercept) -0.50989130 -0.49695978
agef2        0.08279176  0.06047881
agef3        0.06769368  0.05606300
stagef2      0.45553801  0.45425149
stagef3      0.44437841  0.44552862
stagef4      0.62439906  0.46920985
stagef5      0.46920985  0.74722885
```

```r
x <- c(0, -1, 1, 0, 0, 0, 0)
x <- as.matrix(x, 7, 1)
dim(x)
```

```
[1] 7 1
```

```r
# compute the variance estimate of difference in estimates
# for age parameters
t(x) %*% v %*% x
```

```
          [,1]
[1,] 0.2648706
```

- We previously found that:

$$\widehat{\text{OR}} = \exp\{\hat{\beta}_2 - \hat{\beta}_1\} = \exp\big\{-2.613 - (-2.118)\big\} = 0.61$$

- From the new R output, we have calculated:

$$\widehat{\text{Var}}(\hat{\beta}_2 - \hat{\beta}_1) = 0.2649$$

- An approximate $95\,\%$ CI for $\beta_2 - \beta_1$ is therefore:

$$-0.495 \pm 1.96\sqrt{0.2649} = (-1.504, 0.514)$$

- The corresponding interval for the odds ratio is:

$$\exp\{(-1.504, 0.514)\} = (0.22, 1.67)$$

Topic 2e: Bioassay and Dose Response Models

- Previously: Logistic regression analysis of Binomial data with binary and categorical explanatory variables.
- Today: Explore different link functions and continuous explanatory variables.

1. Modelling the Dose Response Relationship.
   - Tolerance distributions and their associated links.
   - Finding the median lethal/effective dose.
2. Beetle Mortality Example.

2.10 Bioassay and Dose Response Models

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

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

- Response: As a result of the stimulus, subjects will manifest a binary response (often of the form died/survived).
- 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?

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

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

Modelling the Dose Response Relationship

For each group $j = 1, \ldots, J$ let:

- $m_j =$ number of subjects in group j.
- $x_j =$ dose applied to subjects in group j.
- $y_j =$ the number of subjects with response in group j.

Dose x_j	Responders y_j	Total m_j	y_j/m_j
1.6907	6	59	0.10
1.7242	13	60	0.22
1.7552	18	62	0.29

Assume

$$Y_j \sim \text{BIN}(m_j, \pi_j), \qquad j = 1, \ldots, J \text{ independently}$$

where $\pi_j =$ probability of response in group j (i.e., at dose x_j).

- Goal: To model $\pi_j = \pi_j(x_j)$ as a function of the continuous stimulus/dose covariate x_j.
- Since $0 \leq \pi \leq 1$, the usual setup is to model using:

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

  where $g(\,\cdot\,)$ is a link function.
- Then we have:

  $$\pi(x) = g^{-1}(\beta_0 + \beta_1 x)$$
- What link function should we select?

Typical Dose Response Curve

TODO figure

- This suggests selecting $g(\,\cdot\,)$ such that $g^{-1}(\,\cdot\,)$ is a cdf:

  $$\pi(x) = g^{-1}(\beta_0 + \beta_1 x) = F^*(\beta_0 + \beta_1 x)$$

The Link Function and the Tolerance Distribution

- Now we have an inverse link function that is a cdf:

  $$\pi(x) = g^{-1}(\beta_0 + \beta_1 x) = F^*(\beta_0 + \beta_1 x)$$
- Recall our original definition of the tolerance distribution:

  $$\pi(x) = \int_{-\infty}^{x} f(s)\,\mathrm{d}s$$
- So if we select a tolerance distribution that will determine the link function through:

  $$\pi(x) = g^{-1}(\beta_0 + \beta_1 x) = F^*(\beta_0 + \beta_1 x) = \int_{-\infty}^{\infty} f(s)\,\mathrm{d}s$$
- $f(x)$ determines how the "probability of a positive response" changes with the value of the dose.

  $$\frac{\partial \pi(x)}{\partial x} = (F^*)'(\beta_0 + \beta_1 x)(\beta_1) = f(x)$$

Some Choices for the Tolerance Distribution

1. Normal Tolerance Distribution ($f(s)$ is Normal pdf):

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

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

$$\begin{aligned} g^{-1}(\beta_0+\beta_1 x) &= \Phi\left(\frac{x-\mu}{\sigma}\right) \\ \pi(x) = g^{-1}(\beta_0+\beta_1 x) &= \Phi\left(\frac{x-\mu}{\sigma}\right) \\ \Phi^{-1}\big(\pi(x)\big) = \beta_0+\beta_1 x &= \frac{x-\mu}{\sigma}\end{aligned}$$

We call this the Probit link $g(\,\cdot\,) = \Phi^{-1}(\,\cdot\,)$.

1. 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 tolerance distribution.

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

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

Median lethal/effective dose

Let δ be the median lethal/effective dose.

- The dose δ at which $50\,\%$ of the population has the response i.e., $\pi(\delta) = 0.50$.
- Find an expression for δ in terms of β_0 and β_1:

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

- Can also find other quantiles of the tolerance distribution i.e., $\pi(\delta_p) = p$, $0 < p < 1$.

A Dose Response Example

Name	Tolerance Distribution $\pi = g^{-1}(\eta)$	Link Function $\eta = g(\pi)$
Normal	$\pi(x) = \int_{-\infty}^{x} \frac{\beta_1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(\beta_0+\beta_1 s)^2\right\} ds$ $= \Phi(\beta_0+\beta_1 x)$ $= \Phi(\eta)$	$\eta = \Phi^{-1}(\pi)$ probit
Logistic	$\pi(x) = \int_{-\infty}^{x} \frac{\beta_1 \exp\{\beta_0+\beta_1 s\}}{(1+\exp\{\beta_0+\beta_1 s\})^2} ds$ $= \frac{\exp\{\beta_0+\beta_1 x\}}{1+\exp\{\beta_0+\beta_1 x\}}$ $= \frac{\exp\{\eta\}}{1+\exp\{\eta\}}$	$\eta = \log\left(\frac{\pi}{1-\pi}\right)$ logistic
Extreme Value	$\pi(x) = \int_{-\infty}^{x} \beta_1 \exp\{\beta_0+\beta_1 s - \exp\{\beta_0+\beta_1 s\}\} ds$ $= \int_{-\infty}^{\eta} \exp\{\nu - \exp\{\nu\}\} d\nu$ $= 1 - \exp\{-\exp\{\eta\}\}$	$\eta = \log(-\log(1-\pi))$ complementary log-log

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.

Dose (x_i)	# of insects killed (x_i)	# of insects m_i	y_j/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

R Data and Code

Data file `beetle.dat`

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

- Recall we are interested in modelling the dose-response relationship:

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

  where $x =$ dose.

- We will fit several binomial regression models to this data.
- Use various link functions to find the best model.

```r
# R program for analysis of dose-response data
beetle.dat <- read.table("beetle.dat", header = T)
# here we construct the response variable for logistic
# regression
beetle.dat$resp <- cbind(beetle.dat$y, beetle.dat$m - beetle.dat$y)
beetle.dat
# here we fit a logistic model involving dose
model1 <- glm(resp ~ dose, family = binomial(link = logit), data = beetle.dat)
summary(model1)
# here we record deviance residuals in rd1
rd1 <- residuals.glm(model1, "deviance")
fv1 <- model1$fitted.values
# plotting the deviance residuals by dose and by fitted
# values
pdf("beetle-residuals.pdf", width = 10, height = 8)
par(mfrow = c(3, 2))
plot(beetle.dat$dose, rd1, ylim = c(-5, 5), xlab = "DOSE", ylab = "DEVIANCE RESIDUALS")
abline(h = -2)
abline(h = 2)
title("Model 1 - logit link")
plot(fv1, rd1, ylim = c(-5, 5), xlab = "FITTED VALUE", ylab = "DEVIANCE RESIDUALS")
abline(h = -2)
abline(h = 2)
title("Model 1 - logit link")
# here we fit a probit model involving dose
model2 <- glm(resp ~ dose, family = binomial(link = probit),
  data = beetle.dat)
summary(model2)
rd2 <- residuals.glm(model2, "deviance")
fv2 <- model2$fitted.values
# here we fit a complementary log-log model involving dose
model3 <- glm(resp ~ dose, family = binomial(link = cloglog),
  data = beetle.dat)
summary(model3)
rd3 <- residuals.glm(model3, "deviance")
fv3 <- model3$fitted.values
plot(beetle.dat$dose, rd2, ylim = c(-5, 5), xlab = "DOSE", ylab = "DEVIANCE RESIDUALS")
abline(h = -2)
abline(h = 2)
title("Model 2 - probit link")
plot(fv2, rd2, ylim = c(-5, 5), xlab = "FITTED VALUE", ylab = "DEVIANCE RESIDUALS")
abline(h = -2)
abline(h = 2)
title("Model 2 - probit link")
```

```
plot(beetle.dat$dose, rd3, ylim = c(-5, 5), xlab = "DOSE", ylab = "DEVIANCE RESIDUALS")
abline(h = -2)
abline(h = 2)
title("Model 3 - log-log link")
plot(fv3, rd3, ylim = c(-5, 5), xlab = "FITTED VALUE", ylab = "DEVIANCE RESIDUALS")
abline(h = -2)
abline(h = 2)
title("Model 3 - log-log link")
```

Selected R Output

Fit of the logistic link model:

```
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 link model:

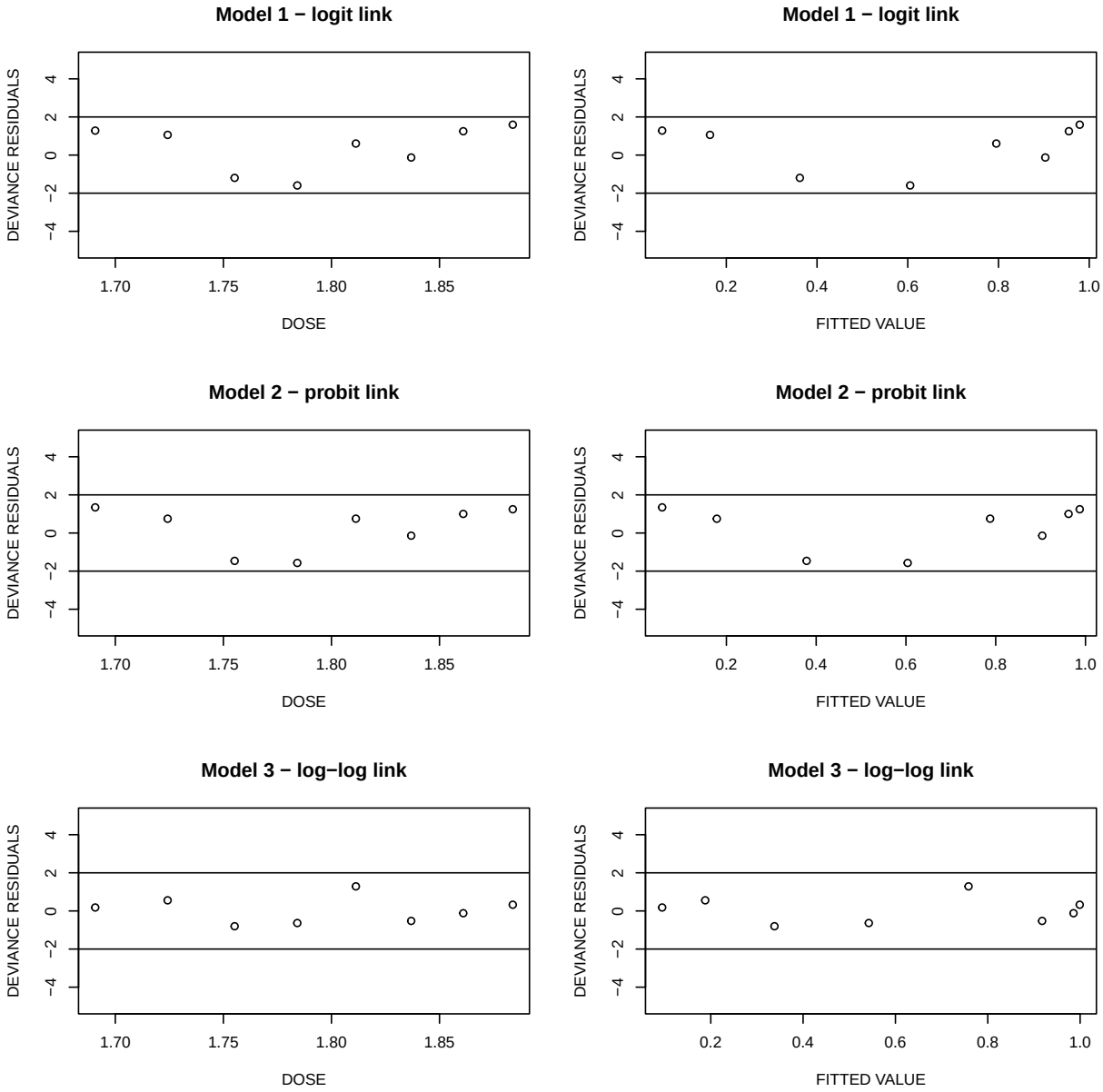
```
summary(model2)


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 link model:

```
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

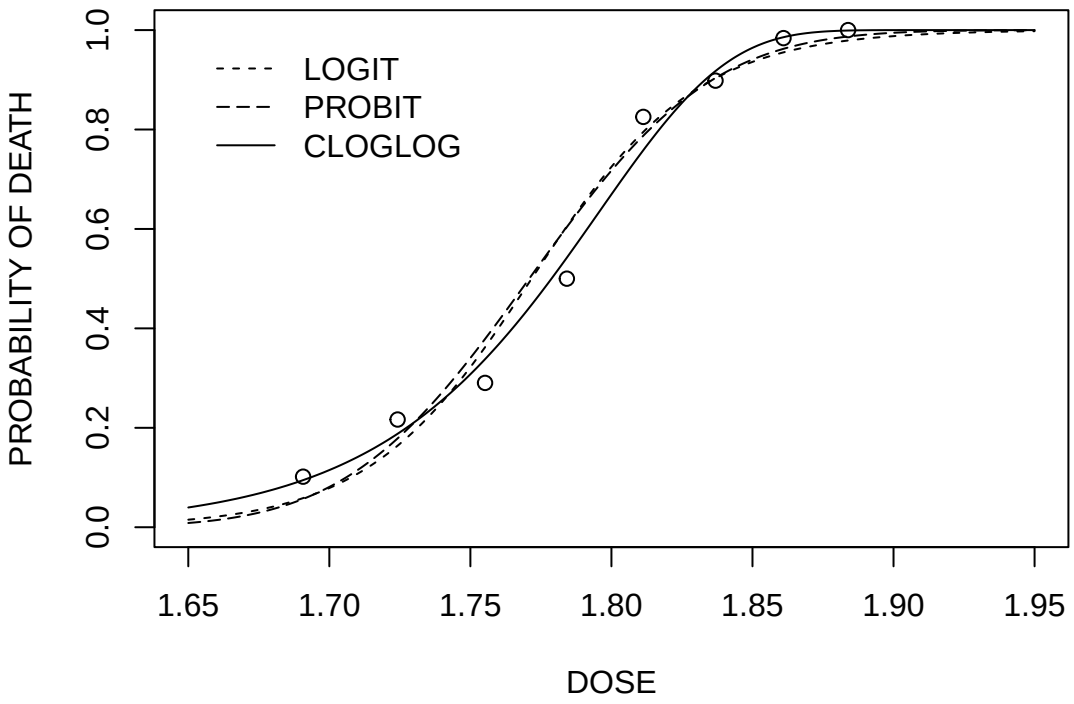


We can plot the actual data (as y_i/m_i) against dose x_i, and see how well the dose-response curves $\hat{\pi}(x) = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1)$ fit the data.

Fitted Dose-Response Curves

```
# Plot the dose-response curves
plot(beetle.dat$dose, beetle.dat$y/beetle.dat$m, xlim = c(1.65,
  1.95), ylim = c(0, 1), xlab = "DOSE", ylab = "PROBABILITY OF DEATH")
x <- seq(1.65, 1.95, by = 0.001)
prob <- as.vector(rep(1, length(x)))
beta <- as.vector(model1$coefficients)  # logistic model
for (i in 1:length(x)) {
  prob[i] <- exp(beta[1] + beta[2] * x[i])/(1 + exp(beta[1] +
```

```
    beta[2] * x[i]))
}
lines(x, prob, lty = 2)
beta <- as.vector(model2$coefficients)  # probit model
for (i in 1:length(x)) {
  prob[i] <- pnorm(beta[1] + beta[2] * x[i])
}
lines(x, prob, lty = 5)
beta <- as.vector(model3$coefficients)  # cloglog model
for (i in 1:length(x)) {
  prob[i] <- 1 - exp(-exp(beta[1] + beta[2] * x[i]))
}
lines(x, prob, lty = 1)
legend(1.65, 1, c("LOGIT", "PROBIT", "CLOGLOG"), lty = c(2, 5,
  1), bty = "n")
```



The curve for the complementary log-log link fits the data better than the other two, as one would expect from the residual plots and the deviance statistics.

Interpretation of Dose-Response Models: `Logistic` Link

$$\text{logit}\big(\pi(x)\big) = \beta_0 + \beta_1 x$$

- $\beta_0 =$ log odds of response at dose of zero.
- Now let's compare the model with $x_1 = 1$ versus $x_1 = 0$.

Dose	$\boldsymbol{x}_i$	η_i	$= \log\big(\pi_i/(1-\pi_i)\big)$
$x+1$	$(1, x+1)^\top$	$\beta_0 + \beta_1(x+1)$	$= \log\big(\pi_1/(1-\pi_1)\big)$
x	$(1, x)^\top$	$\beta_0 + \beta_1 x$	$= \log\big(\pi_0/(1-\pi_0)\big)$
		β_1	$= \log\Big(\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}\Big)$

- We subtract line 2 from line 1 to isolate β_1 and find its interpretation.
- $\beta_1 =$ log odds ratio for response associated with a one unit increase in dose.

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

```
             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -60.71745   5.180701 -11.71993 1.007549e-31
dose         34.27033   2.912134  11.76811 5.698445e-32
```

- What is the OR of response associated with a 0.001 increase in dose?

$$\widehat{\text{OR}} = \exp\{0.001\hat{\beta}\} = \exp\{34.27/1000\} = 1.41$$

- An expression for the median lethal/effective dose:

$$\pi(\delta) = 0.50 \implies \text{logit}(0.5) = \beta_0 + \beta_1\delta \implies \delta = -\beta_0/\beta_1$$

- Here $\hat{\delta} = 60.7175/34.2703 = 1.772$.
- Can also find an expression for the $100p$th percentile of the tolerance distribution ($0 < p < 1$):

$$\pi(\delta) = p \implies \text{logit}(p) = \beta_0 + \beta_1\delta_p$$

Interpretation of Dose-Response Models: `Probit` Link

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

where Φ is the CDF of a $\mathcal{N}(0, 1)$ random variable.

- Interpretation of β in terms of (μ, σ) parameters of the tolerance distribution:

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

```
summary(model2)$coefficients
```

```
             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -34.93527   2.647879 -13.19368 9.541285e-40
dose         19.72794   1.487213  13.26504 3.692396e-40
```

- An expression for the median lethal/effective dose:

$$\pi(\delta) = 0.50 \implies \delta = \frac{-\beta_0}{\beta_1}$$

- Here $\hat{\delta} = 34.9353/19.7279 = 1.771$.

- Can also find an expression for the $100p$th percentile of the tolerance distribution:

$$\Phi^{-1}(p) = \beta_0 + \beta_1 \delta_p \implies \delta_p = \frac{\Phi^{-1}(p) - \beta_0}{\beta_1}$$

- Exercise: What are $\delta_{0.25}$ and $\delta_{0.75}$ the 25th and 75th percentiles of the tolerance distribution from the probit model?

```
qnorm(0.25)

[1] -0.6744898

qnorm(0.75)

[1] 0.6744898
```

$$\hat{\delta}_{0.25} = \frac{-0.6745 + 34.9353}{19.7279} = 1.737, \qquad \hat{\delta}_{0.75} = \frac{0.6745 + 34.9353}{19.7279} = 1.805$$

Interpretation of Dose-Response Models: `cloglog` Link

$$\log\Big(-\log\big(1 - \pi(x)\big)\Big) = \beta_0 + \beta_1 x$$

- Interpretation of β parameters is not as natural as in other two models:

$$\beta_0 = \log\Big(-\log\big(1 - \pi(0)\big)\Big), \qquad \beta_1 = \log\left(\frac{-\log\big(1 - \pi(x+1)\big)}{-\log\big(1 - \pi(x)\big)}\right)$$

```
summary(model3)$coefficients

             Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -39.57231   3.240290 -12.21258 2.662986e-34
dose         22.04117   1.799365  12.24942 1.692092e-34
```

- An expression for the median lethal/effective dose:

$$\pi(\delta) = 0.50 \implies \delta = \frac{\log\big(-\log(1 - 0.5)\big) - \beta_0}{\beta_1}$$

- Here $\hat{\delta} = (-0.3665 + 39.5723)/22.0412 = 1.779$.

Dose-Response Models: Summary

- Comparison of models with different links must be done through plots of the deviance residuals or fitted dose response curves.
- Interpretation of regression parameters β_j depend on the link function.
- Consider estimating δ_p where $\pi(\delta_p) = p$, $0 < p < 1$ to learn about the underlying tolerance distribution.
- Prediction: $\hat{\pi}(x) = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1)$.

- Multiple explanatory variables can be included in dose response models.

WEEK 6
11th to 15th October

Reading week.

WEEK 7
18th to 22nd October

Topic 2f: Topic 2f: Binomial Regression Wrap-Up

1. Summary of Chapter 2.
2. Example: Birdkeeping and Lung Cancer.
3. Example: Birdkeeping and Lung Cancer (continued).

Summary of Chapter 2

Binomial GLM / Logistic Regression Model

$Y_i \sim \text{BIN}(m_i, \pi_i)$, $i = 1, \ldots, n$ independently, with explanatory variables $\boldsymbol{x}_i$:

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

- Estimation: $\hat{\boldsymbol{\beta}}$ come from Fisher Scoring using R function `glm()`.
- Interpretation: β_k have log OR interpretations ($k > 0$).
- Wald based Hypothesis Tests of H_0: $\beta_k = \beta_{k0}$ versus H_A: $\beta_k \neq \beta_{k0}$. Under H_0:

  $$(\hat{\beta}_k - \beta_{k0})^2 \left(I^{kk}(\hat{\boldsymbol{\beta}})\right)^{-1} \sim \chi^2(1)$$

  equivalently, $\dfrac{\hat{\beta}_k - \beta_{k0}}{\mathsf{se}(\hat{\beta}_k)} \sim \mathcal{N}(0,1)$ where $\mathsf{se}(\hat{\beta}_k) = \sqrt{I^{kk}(\hat{\boldsymbol{\beta}})}$.
- Confidence Interval for a single β_k:

  $$\hat{\beta}_k \pm z_{1-\alpha/2}\, \mathsf{se}(\hat{\beta}_k) \qquad \text{where } \mathsf{se}(\hat{\beta}_k) = \sqrt{I^{kk}(\hat{\boldsymbol{\beta}})}$$

- Deviance/LR based Hypothesis Tests for nested models:

  $$H_0\text{: } \beta_p = \cdots = \beta_{q-1} = 0 \text{ vs } H_A\text{: at least one of } \beta_p, \ldots, \beta_{q-1} \neq 0$$

  using

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

- Deviance Residuals (should be iid $\mathcal{N}(0,1)$ for a well-fitting model):

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

  where

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

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

$$\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}} \pm 1.96\sqrt{\boldsymbol{x}_i^\top I^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i} = (\hat{\eta}_{\text{L}}, \hat{\eta}_{\text{U}})$$

  then transform ends of the interval to get a CI for OR, π, etc.

- Bioassay experiments:
  - β interpretation depends on link function.
  - Calculation of δ_p: dose that gives pth percentile of response.

The Model Fitting Process

The Model Fitting Process

0. Exploratory Data Analysis.
1. Model Specification — Select a probability distribution for the response variable and an equation linking the response to the explanatory variables.
2. Estimation of 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.

Let's apply this process to an example using logistic regression.

Example: Birdkeeping and Lung Cancer

Birdkeeping and Lung Cancer

A 1972 to 1981 health survey in The Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer. To investigate birdkeeping as a risk factor, researchers conducted a case-control study of patients in 1985 at four hospitals in The Hague (population 450,000). They identified 49 cases of lung cancer among the patients who were registered with a general practice, who were age 65 or younger and who had resided in the city since 1965. They also selected 98 controls from a population of residents having the same general age structure.
From Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Methods of Data Analysis (2nd ed) https://cran.r-project.org/web/packages/Sleuth3/Sleuth3.pdf

Birdkeeping and Lung Cancer Dataset

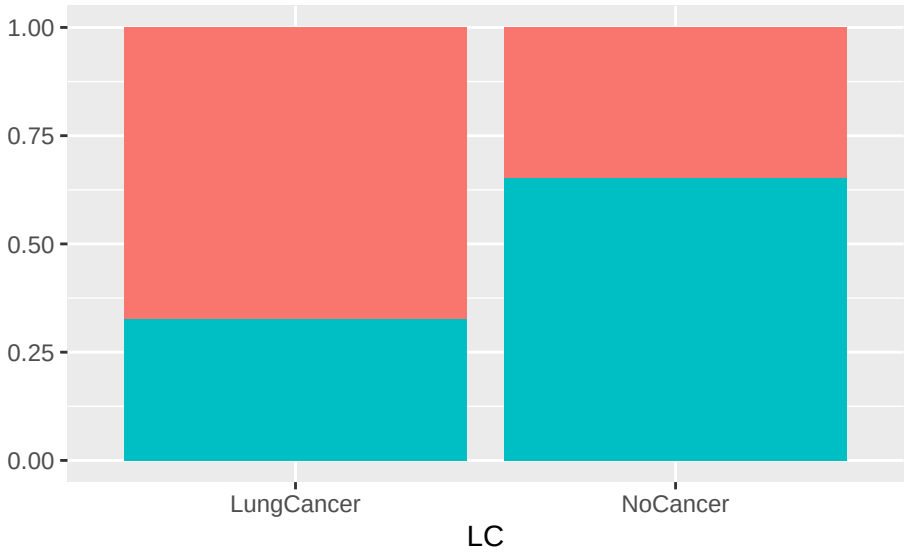
Exploratory Data Analysis

- Primary Research Question: Is there an association between birdkeeping and an increased risk of lung cancer?

LC	binary	Whether subject has lung cancer (the response)
FM	binary	Sex of subject (Female or Male)
SS	binary	Socioeconomic status (High or Low)
BK	binary	Indicator for birdkeeping (Bird or NoBird)
AG	integer	Age of subject (years)
YR	integer	Years of smoking prior to diagnosis or examination
CD	integer	Average rate of smoking (cigarettes per day)

Subject	LC	FM	SS	BK	AG	YR	CD
1	LungCancer	Male	Low	Bird	37	19	12
2	LungCancer	Male	Low	Bird	41	22	15
3	LungCancer	Male	High	NoBird	43	19	15
4	LungCancer	Male	Low	Bird	46	24	15
5	LungCancer	Male	Low	Bird	49	31	20

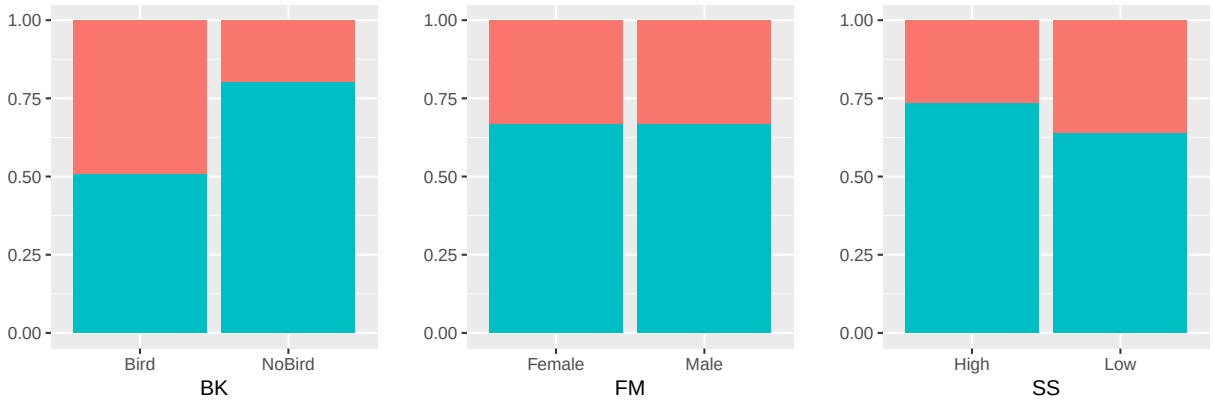
	LungCancer	NoCancer	Total
Bird	33	34	67
NoBird	16	64	80
Total	49	98	147



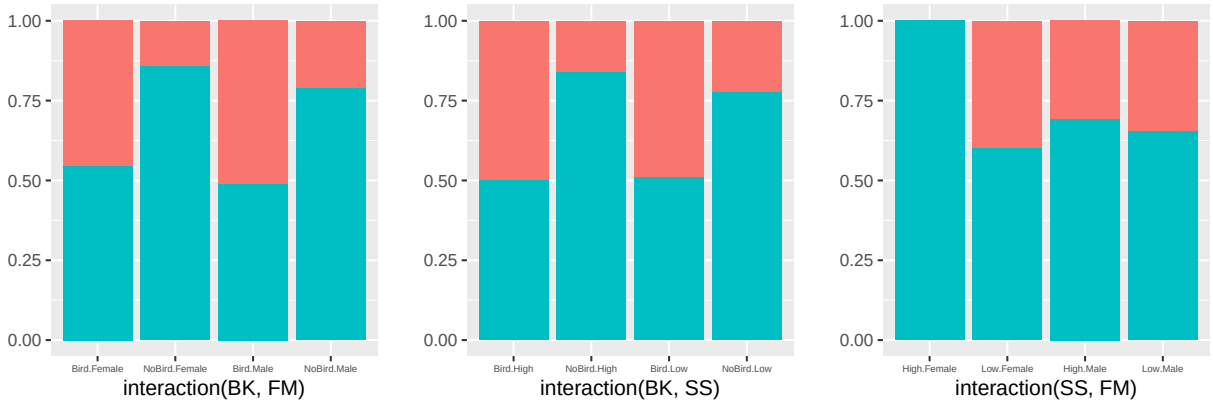
$$\widehat{\text{OR}} = \hat{\psi} = \frac{(33)(64)}{(16)(34)} = 3.882353$$

- So there is the suggestion of an association, but we need to take other potentially important explanatory variables into account.

Proportion of Lung Cancer (top) versus No Cancer (bottom) for Binary Explanatory Variables
(BK = birdkeeping, FM = sex, SS = socioeconomic status)



Proportion of Lung Cancer (top) versus No Cancer (bottom) for Combinations of Binary Explanatory Variables (BK = birdkeeping, FM = sex, SS = socioeconomic status)



Model Specification

- We will fit logistic regression models to the data using R
- The full main effects model is:

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

where

$$\begin{aligned}
\pi_i &= \mathbb{P}(\text{subject } i \text{ has lung cancer}) = \texttt{LC} \\
x_{i1} &= \mathbb{I}\{\text{Birdkeeper}\} = \texttt{BK} \\
x_{i2} &= \mathbb{I}\{\text{Male}\} = \texttt{FM} \\
x_{i3} &= \mathbb{I}\{\text{Low SES}\} = \texttt{SS} \\
x_{i4} &= \text{Age of subject (years)} = \texttt{YR} \\
x_{i5} &= \text{Years of smoking} = \texttt{AG} \\
x_{i6} &= \text{Cigarettes per day} = \texttt{CD}
\end{aligned}$$

Estimation and Model Checking

Model Building Plan

- First, we will consider models that do not include the birdkeeping BK $= x_{i1}$ explanatory variable.
  - i.e., look for associations between lung cancer and other explanatory variables.
  - Find the best fitting model without birdkeeping.
- Then find the best model that includes birdkeeping.
- The model fitting process is iterative and can be somewhat subjective.
- Unclear whether Age and Sex should be considered due to possible matching in the design of the case control study.

myGlm1: Main Effects Model (no BK)

```
myGlm1 <- glm(LC ~ FM + SS + AG + YR + CD, family = binomial)
summary(myGlm1)


Call:
glm(formula = LC ~ FM + SS + AG + YR + CD, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3910  -0.9718  -0.5519   1.1733   2.5020

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.37895    1.67206   0.227  0.82070
FMMale      -0.74923    0.50501  -1.484  0.13792
SSLow        0.07303    0.43893   0.166  0.86785
AG          -0.05799    0.03432  -1.690  0.09112 .
YR           0.07955    0.02636   3.018  0.00255 **
CD           0.01978    0.02422   0.817  0.41421
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 165.87  on 141  degrees of freedom
AIC: 177.87

Number of Fisher Scoring iterations: 5
```

Example of a Wald Test for a Single Parameter

Is years of smoking associated with lung cancer?

$$H_0\text{: } \beta_4 = 0 \text{ versus } H_A\text{: } \beta_4 \neq 0$$

Wald-based test statistic: ($t \sim \mathcal{N}(0,1)$ under H_0):

$$t = \frac{\hat{\beta}_4 - 0}{\mathsf{se}(\hat{\beta}_4)} = \frac{0.07955}{0.02636} = 3.018$$

Now find the p-value by comparing to $Z \sim \mathcal{N}(0,1)$:

$$p = 2\,\mathbb{P}(Z > |t|) = 2\,\mathbb{P}(Z > 3.018) = 0.0026$$

```
2 * (1 - pnorm(3.018))

[1] 0.002544489
```

Therefore, reject the null hypothesis that smoking is not associated with lung cancer (after adjustment for sex, socioeconomic status, age, and cigarettes per day).

myGlm2: Drop SS

```
myGlm2 <- update(myGlm1, ~. - SS)
summary(myGlm2)


Call:
glm(formula = LC ~ FM + AG + YR + CD, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4134  -0.9744  -0.5430   1.1749   2.5123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.46101    1.59688   0.289  0.77282
FMMale      -0.76832    0.49178  -1.562  0.11821
AG          -0.05858    0.03415  -1.715  0.08628 .
YR           0.08027    0.02603   3.083  0.00205 **
CD           0.01959    0.02420   0.810  0.41820
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 165.90  on 142  degrees of freedom
AIC: 175.9

Number of Fisher Scoring iterations: 5
```

myGlm3: Drop CD

```
myGlm3 <- update(myGlm2, ~. - CD)
summary(myGlm3)


Call:
glm(formula = LC ~ FM + AG + YR, family = binomial)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2597  -0.9794  -0.5462   1.1718   2.4894

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.82886    1.52662   0.543 0.587172
FMMale      -0.73638    0.48914  -1.505 0.132210
AG          -0.06363    0.03359  -1.894 0.058195 .
YR           0.08776    0.02452   3.579 0.000344 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 166.55  on 143  degrees of freedom
AIC: 174.55

Number of Fisher Scoring iterations: 5
```

myGlm4: Drop FM

```
myGlm4 <- update(myGlm3, ~. - FM)
summary(myGlm4)


Call:
glm(formula = LC ~ AG + YR, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2933  -0.9869  -0.5682   1.2448   2.5943

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.67653    1.49597   0.452 0.651100
AG          -0.06568    0.03291  -1.996 0.045976 *
YR           0.07815    0.02321   3.368 0.000758 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 168.83  on 144  degrees of freedom
AIC: 174.83

Number of Fisher Scoring iterations: 5
```

myGlm5: Add BK (Birdkeeping)

```
BK <- factor(BK, levels = c("NoBird", "Bird"))  # Make 'no bird' the ref level
myGlm5 <- update(myGlm4, ~. + BK)  # Now add bird keeping
summary(myGlm5)


Call:
glm(formula = LC ~ AG + YR + BK, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5466  -0.8649  -0.4911   0.9763   2.2584

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.03359    1.66069  -0.622 0.533686
AG          -0.04610    0.03430  -1.344 0.178952
YR           0.07485    0.02296   3.261 0.001111 **
BKBird       1.37656    0.40073   3.435 0.000592 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 156.22  on 143  degrees of freedom
AIC: 164.22

Number of Fisher Scoring iterations: 5
```

myGlm6: Add YR:BK and AG:YR Interactions

```
myGlm6 <- update(myGlm5, ~. + BK:YR + AG:YR)  # Try interaction terms
summary(myGlm6)


Call:
glm(formula = LC ~ AG + YR + BK + YR:BK + AG:YR, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6689  -0.8118  -0.4656   0.9643   2.2142

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.332602   5.154442  -1.229    0.219
AG           0.047055   0.086231   0.546    0.585
YR           0.294877   0.197828   1.491    0.136
BKBird       1.101153   1.291672   0.853    0.394
YR:BKBird    0.008546   0.037603   0.227    0.820
AG:YR       -0.003768   0.003215  -1.172    0.241
```

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

    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 154.60  on 141  degrees of freedom
AIC: 166.6

Number of Fisher Scoring iterations: 5
```

Example of a Deviance Test for Nested Models

H_0: Model 5 is adequate compared to model 6 versus H_A: Model 5 is not adequate. H_0: $\beta_{14} = \beta_{45} = 0$ versus H_A: $\beta_{14} \neq 0$ or $\beta_{45} \neq 0$.

Deviance/LR test statistic ($\Delta D \sim \chi^2(2)$ under H_0):

$$\Delta D = D_0 - D_{\text{A}} = D_5 - D_6 = 156.22 - 154.60 = 1.62$$

Now find the p-value by comparing to $\chi^2(2)$:

$$p = \mathbb{P}\big(\chi^2(2) > 1.62\big) = 0.45$$

```r
1 - pchisq(1.62, 2)
```

```
[1] 0.4448581
```

Therefore we do not reject the null hypothesis that model 5 is adequate. We conclude that the interactions are not necessary.

Summary of Deviance Tests

```r
anova(myGlm5, myGlm6)  # Test interaction terms jointly
```

```
Analysis of Deviance Table

Model 1: LC ~ AG + YR + BK
Model 2: LC ~ AG + YR + BK + YR:BK + AG:YR
  Resid. Df Resid. Dev Df Deviance
1       143     156.22
2       141     154.60  2   1.6163
```

```r
1 - pchisq(1.6163, 2)
```

```
[1] 0.4456818
```

- Do not reject the null hypothesis that `myGlm5` (no interactions) is adequate compared to `myGlm6` (interactions)

```r
anova(myGlm4, myGlm5)  # Test for bird keeping effect
```

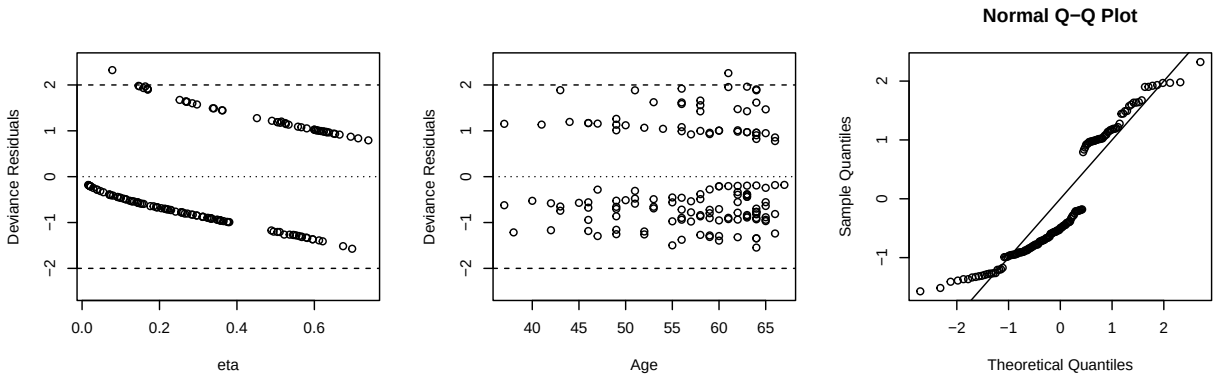
```
Analysis of Deviance Table

Model 1: LC ~ AG + YR
Model 2: LC ~ AG + YR + BK
  Resid. Df Resid. Dev Df Deviance
1       144     168.83
2       143     156.22  1   12.612
```

```
1 - pchisq(12.612, 1)
```

```
[1] 0.0003832782
```

- Reject the null hypothesis that `myGlm4` (no BK) is adequate compared to `myGlm5`.

myGLM5: Deviance Residuals



Final Model: myGLM5b: BK + AG + YR

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_4 x_{i4} + \beta_5 x_{i5}$$

```
# put explanatory variables in expected order
myGlm5b <- glm(LC ~ BK + AG + YR, family = binomial)
summary(myGlm5b)$coefficients
```

```
               Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -1.03359488 1.66069096 -0.6223885 0.5336864701
BKBird       1.37655906 0.40072983  3.4351300 0.0005922696
AG          -0.04609820 0.03429953 -1.3439892 0.1789518774
YR           0.07485289 0.02295533  3.2608062 0.0011109596
```

```
summary(myGlm5b)$cov.unscaled
```

```
            (Intercept)        BKBird            AG            YR
(Intercept)  2.75789447 -0.2110100613 -0.0520288121  0.0106701054
BKBird      -0.21101006  0.1605843947  0.0019968775  0.0002914915
AG          -0.05202881  0.0019968775  0.0011764579 -0.0004893738
YR           0.01067011  0.0002914915 -0.0004893738  0.0005269473
```

Inference and Prediction

Find and estimate and 95 % confidence interval of the Odds Ratio of lung cancer in birdkeepers versus non-birdkeepers.

Estimate:

$$\widehat{\text{OR}} = \exp\{\hat{\beta}_1\} = \exp\{1.3766\} = 3.96$$

95 % Confidence Interval:

$$\begin{aligned}\exp\{\hat{\beta}_1 \pm 1.96\,\text{se}(\hat{\beta}_1)\} &= \exp\{1.3766 \pm 1.96(0.4007)\} \\ &= \big(\exp\{0.5912\}, \exp\{2.1620\}\big) \\ &= (1.81, 8.69)\end{aligned}$$

In this sample, what is the probability that a 50-year-old, non-smoking, non-birdkeeper has lung cancer?

Estimate:

$$\begin{aligned}\hat{\pi}_i &= \text{expit}(\hat{\beta}_0 + 50\hat{\beta}_4) \\ &= \text{expit}\big(-1.0336 + 50(-0.0461)\big) \\ &= \text{expit}(-3.3385) \\ &= 0.03427 \text{ or } 3.43\,\%\end{aligned}$$

95 % Confidence Interval:

$$\begin{aligned}\text{Var}(\hat{\beta}_0 + 50\hat{\beta}_4) &= \text{Var}(\hat{\beta}_0) + 50^2\,\text{Var}(\hat{\beta}_4) + 2(50)\,\text{Cov}(\hat{\beta}_0, \hat{\beta}_4) \\ &= 2.579 + 50^2(0.001176) + 100(-0.05203) \\ &= 0.4949\end{aligned}$$

$$\text{expit}(-3.3385 \pm 1.96\sqrt{0.4949}) = (0.008845, 0.1237) \text{ or } (0.88\,\%, 12.37\,\%)$$

```
# Inference and Prediction
exp(myGlm5b$coefficients)  # Odds Ratios

(Intercept)      BKBird          AG          YR
  0.3557259   3.9612477   0.9549482   1.0777256

# 95% CI, OR for lung cancer, birdkeepers vs
# non-birdkeepers, controlling for age and years of smoking
exp(myGlm5b$coef[2] + c(-1, 1) * qnorm(0.975) * sqrt(summary(myGlm5b)$cov.unscaled[2,
  2]))

[1] 1.806052 8.688281

# 95% CI, OR for lung cancer, one year increase in smoking,
# controlling for age and birdkeeping status
exp(myGlm5b$coef[4] + c(-1, 1) * qnorm(0.975) * sqrt(summary(myGlm5b)$cov.unscaled[4,
  4]))

[1] 1.030312 1.127322

expit <- function(x) {
  exp(x)/(1 + exp(x))
}
x <- as.matrix(c(1, 0, 50, 0), ncol = 1)  # 50-year-old non-smoker, non-birdkeeper
expit(t(x) %*% myGlm5b$coefficients)
```

```
           [,1]
[1,] 0.03427361

v <- summary(myGlm5b)$cov.unscaled
t(x) %*% v %*% x   # Var(beta_0 + 50 beta_4)

         [,1]
[1,] 0.496158

# 95% CI, predicted probability of lung cancer for
# 50-year-old non-smoker birdkeeper
expit(t(x) %*% myGlm5b$coefficients + c(-1, 1) * qnorm(0.975) *
  sqrt(t(x) %*% v %*% x))

[1] 0.008844516 0.123690591
```

Inference

- Controlling for age and years of smoking, the odds ratio of getting lung cancer for birdkeepers vs non-birdkeepers is 3.96 $(1.81, 8.69)$.
- Controlling for age and birdkeeping, the odds ratio of getting lung cancer for each additional year of smoking is 1.08 $(1.03, 1.13)$.

Prediction

- In this study, the probability that a 50-year-old, non-smoking, non-birdkeeper has developed lung cancer is $3.43\,\%$ $(0.88\,\%, 12.37\,\%)$.
- Does this estimate extend to the general population?

Topic 3a: Introduction to Poisson GLMs

1. Setting up a Poisson GLM for Counts.
   - Review of the Poisson distribution as a member of the exponential family.
   - Specification of a Poisson GLM (i.e., Log Linear Regression Model).
   - Derivation of Poisson deviance and deviance residuals.
2. Regression for Poisson Processes.
   - Definition of a Poisson Process.
   - Log Linear Regression Model for a Time Homogeneous Poisson Process.
   - Introduction of the offset term.

The Poisson Distribution

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

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

- The Poisson is a member of the exponential family with:

$$\theta = \log(\mu), \qquad \phi = 1, \qquad b(\theta) = e^{\theta} = \mu$$
$$a(\phi) = 1, \qquad c(y;\phi) = -\log(y!)$$

- With mean and variance:

$$\mathbb{E}[Y] = b'(\theta) = e^{\theta} = \mu$$
$$\mathsf{Var}(Y) = b''a(\phi) = e^{\theta} = \mu$$

- And Canonical link:

$$\theta = \eta = g(\mu) \implies g(\mu) = \log(\mu)$$

Poisson Likelihood

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

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

- Response vector: $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top$.
- Mean vector: $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^\top$.
- The likelihood and log-likelihood are:

$$L(\boldsymbol{\mu}) = \prod_{i=1}^{n} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

$$\ell(\boldsymbol{\mu}, \boldsymbol{y}) = \sum_{i=1}^{n} \big(y_i \log(\mu_i) - \mu_i - \log(y_i!)\big)$$

Log Linear Regression

- Explanatory variables: $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip-1})^\top$, $i = 1, \ldots, n$.
- Regression parameters: $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p-1})^\top$.
- Using the Canonical link (i.e., log link):

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

- The use of the log link gives the term log linear regression.
- We can obtain the log-likelihood in terms of $\boldsymbol{\beta}$ by substitution:

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

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 $\hat{\boldsymbol{\beta}}$ via Fisher Scoring.
- Use `glm()` function in R with `family=poisson(link=log)`.

Inference for β from log linear regression: Wald Tests

$$H_0\text{: } \beta_k = \beta_{k0} \text{ versus } H_A\text{: } \beta_k \neq \beta_{k0}$$

- The general Wald Result for scalar β_k is:

$$(\hat{\beta}_k - \beta_{k0})^2 \big(I^{kk}(\hat{\boldsymbol{\beta}}) \big)^{-1} \sim \chi^2_{(1)}$$

equivalently $\dfrac{\hat{\beta}_k - \beta_{k0}}{\mathsf{se}(\hat{\beta}_k)} \sim \mathcal{N}(0,1)$ where $\mathsf{se}(\hat{\beta}_k) = \sqrt{I^{kk}(\hat{\boldsymbol{\beta}})}$.

- And we can find the p-value of the test using:

$$p = 2\,\mathbb{P}\left(U > \frac{|\hat{\beta}_k| - \beta_{k0}}{\mathsf{se}(\hat{\beta}_k)} \right) \qquad \text{where } U \sim \mathcal{N}(0,1)$$

- The `summary()` output gives the test statistics and p-values for testing H_0: $\beta_k = 0$ vs H_A: $\beta_k \neq 0$.

Poisson Deviance/Likelihood Ratio Tests

- Let $\tilde{\mu}_i$ be the MLE under the saturated model (i.e., $\tilde{\mu}_i = y_i$).
- Let $\hat{\mu}_i$ be the MLE under a p-dimensional constrained model.
- Recall the Likelihood Ratio or Deviance Statistic has the form:

$$D = -2 \log\left(\frac{\mathcal{L}(\hat{\mu})}{\mathcal{L}(\tilde{\mu})} \right) = 2\big(\ell(\tilde{\mu}) - \ell(\hat{\mu}) \big) \sim \chi^2_{(n-p)}$$

asymptotically under the assumption that the constrained model is appropriate.

- For the Poisson we have:

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

- Note that the Deviance Statistic has the same form as in the Binomial case:

$$D = 2\sum O_i \log\left(\frac{O_i}{E_i}\right)$$

provided an intercept is included in the model so that $\sum(y_i - \hat{\mu}_i) = 0$.

Poisson Deviance/Likelihood Ratio Tests

- Use the Deviance to test nested models:
  - H_0: the null model with p parameters is adequate versus

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{1p-1}$$

  - H_A: the alternative model with q parameters ($p < q$)

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{1p-1} + \cdots + \beta_{q-1} x_{1q-1}$$

  - With test statistic:

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

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

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

Deviance Residuals

- We can write the Deviance as a sum:

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

- The Deviance Residuals are given by:

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

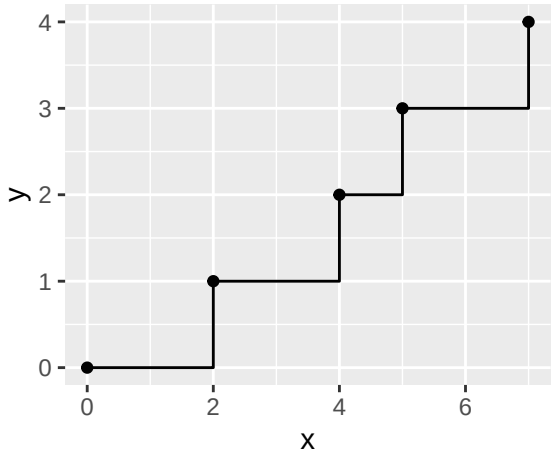
and are approximately $\mathcal{N}(0,1)$ if H_0 holds.

Regression for Poisson Processes

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 $(2, 4, 5, 7)$:



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 distribution of $N(t)$ the number of events occurring over $(0, t]$ is given by:

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

Regression for Poisson Processes $N(t)$

- $N(t)$ is a special kind of Poisson random variable with:

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

- Use the log link to do regression:

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

- $\lambda =$ Rate parameter.
- $t =$ Length of observation (data).
- Since λ is constant (not a function of t) we call this a time homogeneous poisson process.

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

$$\begin{aligned}\log\big(\mu_i(t_i)\big) &= \log(\lambda_i) + \log(t_i) \\ &= \boldsymbol{x}_i^\top \boldsymbol{\beta} + \log(t_i)\end{aligned}$$

- The term $\log(t_i)$ is called an offset term.
- It *explains* some variation in the event counts N_i across subjects due to differing lengths of observation t_i.

WEEK 8
25th to 29th October

Topic 3b: Ship Damage Example

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:

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

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.

```
# 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 $\beta_4 \neq 0$.

```
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))`

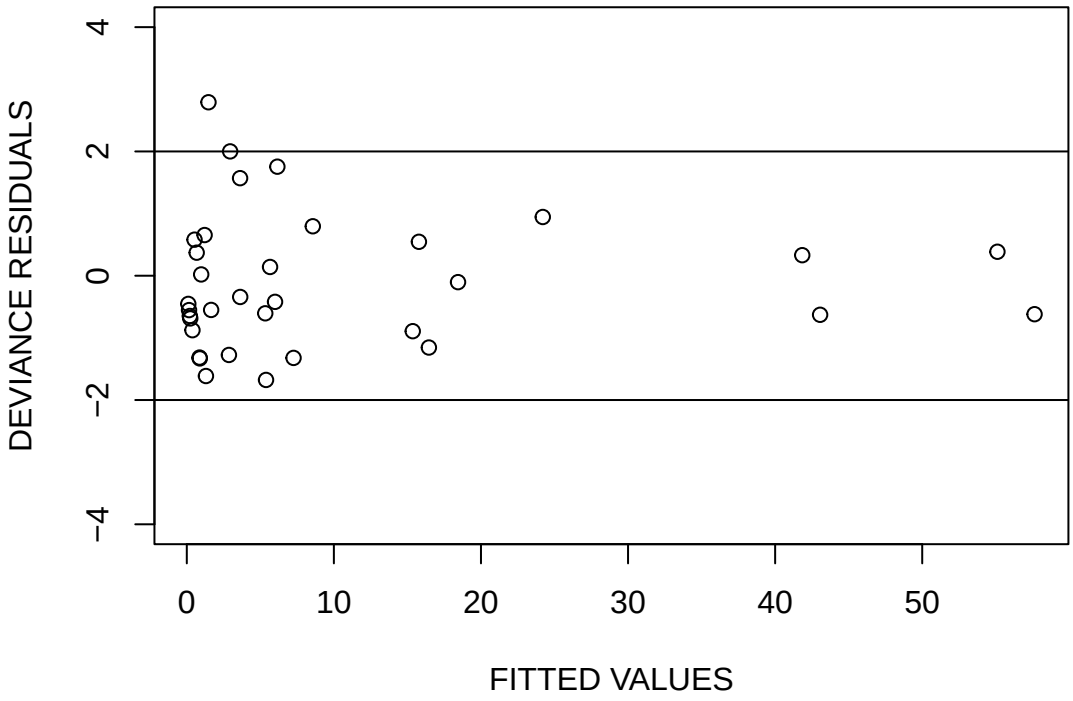
- Use this model to test:
  - H_0: Operation year is unimportant (i.e., $\beta_8 = 0$).
  - H_A: $\beta_8 \neq 0$.

```
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))`

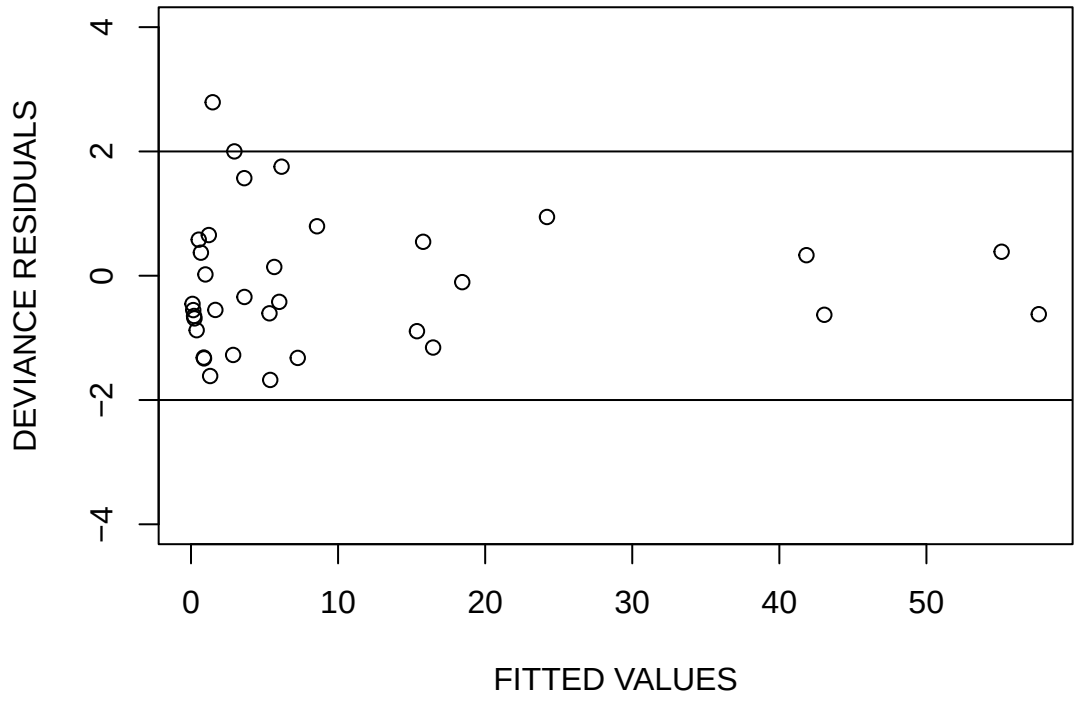
- Use this model to test:
  - H_0: the `cyrf*oyrf` interaction is unimportant (Model 1).
  - H_A: (model 6).

```
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 $\exp\{\beta_4 - \beta_1\}$:

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

- *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 3c: Log Linear Models

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:

$$\begin{aligned} f(y) &= \binom{m}{y}\pi^y(1-\pi)^{m-y} \\ f(y) = \quad &= \frac{(m)(m-1)\cdots(m-y)(m-y-1)\cdots(1)}{(m-y)!y!}\left(\frac{\mu}{m}\right)^y\left(1-\frac{\mu}{m}\right)^{m-y} \\ &= \underbrace{\frac{(m)(m-1)\cdots(m-(y-1))}{(m)(m)\cdots(m)}}_{\to 1 \text{ as } m\to\infty}\frac{\mu^y}{y!}\underbrace{\left(1-\frac{\mu}{m}\right)^m}_{\to e^{-\mu}}\underbrace{\left(1-\frac{\mu}{m}\right)^{-y}}_{\to 1}. \end{aligned}$$

- 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.

Example: Poisson Approximation to the Binomial

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

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

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

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

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

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

- 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.

- 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:
$$\lambda(t) = b_1\,\mathbb{I}\{0 < t < t_1\} + b_2\,\mathbb{I}\{t_1 \le t < t_2\} + \cdots.$$
  - Piecewise linear:
$$\lambda(t) = (m_1 t + b_1)\,\mathbb{I}\{0 < t < t_1\} + (m_2 t + b_2)\,\mathbb{I}\{t_1 \le t < t_2\} + \cdots.$$
  - 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

```
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`)

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

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

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

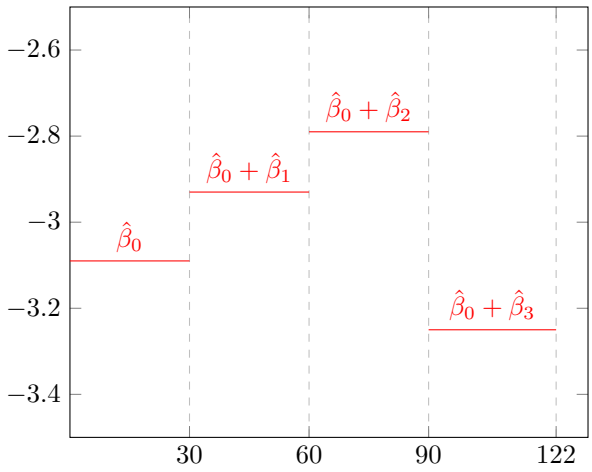
- `interval` is a categorical variable at 4 levels:

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

- 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 $\log(\lambda(t))$ 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:

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

- 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`)

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

- $\beta_0 =$ 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:

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

- 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 3d: 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.

		Factor W							
		1	2	3	$\cdots$	j	$\cdots$	J	
Factor V	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}	$y_{3\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	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}	$y_{I\bullet}$
		$y_{\bullet 1}$	$y_{\bullet 2}$	$y_{\bullet 3}$	$\cdots$	$y_{\bullet j}$	$\cdots$	$y_{\bullet J}$	$y_{\bullet\bullet}$

- 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:

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

- Recall the standard Multinomial distribution:

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

  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

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

- $\boldsymbol{\pi} = (\pi_{11},\ldots,\pi_{IJ})^\top$ 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 χ^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 $Y_{i\bullet} = y_{i\bullet}$, $i = 1, \ldots, I$

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

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
Age	<45	78	78	44	200
	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.

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

- 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 $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{IJ})^\top$ 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 $i = 1, \ldots, I$

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

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

- 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 χ^2_4 distribution:

$$p = \mathbb{P}\left(\chi^2_4 > 21.25615\right) < 0.001.$$

- 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 3e: Log Linear Models for Two-way Tables

Likelihood Based Analysis of 2-way Contingency Tables

		Factor W						
		1	2	3	$\cdots$	j	$\cdots$	J	
Factor V	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}	$y_{3\bullet}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	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:

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

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

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

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

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

- 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:

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

- 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:

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

  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:

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

- 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:

$$H_0\colon\ \log(\mu_{ij}) = u + u_i^V + u_j^W$$

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

- 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:

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

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

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

- 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

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

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

- 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 χ^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

$$p = \mathbb{P}\big(\chi^2_2 > 98.305\big) < 0.001$$

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:

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

$$p = \mathbb{P}\left(\chi_4^2 > 25.19226\right) < 0.001.$$

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 3f: 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
```

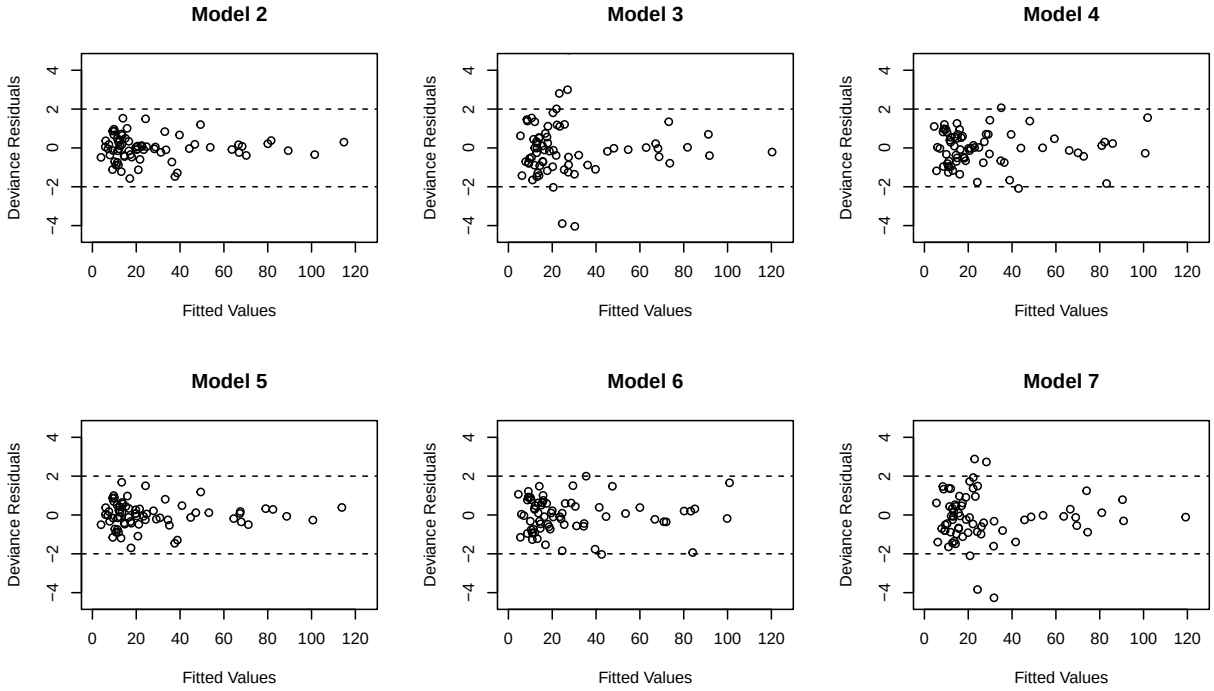
```
[1] 30

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.

Topic 3g: Log Linear Model Applications Wrap-Up

Application 2: Seatbelt Use and Fatality of Accidents

We now consider the special case of a $2 \times 2 \times 2$ table.

Florida Department of Highway Safety and Motor vehicles (Bishop *et al* 1975)

Seatbelt (V)	Ejected (W)	Injury (Z) Non-fatal	Fatal
Used	Yes	1105	14
	No	411111	483
Not Used	Yes	4624	497
	No	157342	1008

- Rewriting the data table in general notation, we have:

Seatbelt (V)	Ejected (W)	Injury (Z) Non-fatal ($k=1$)	Fatal ($k=2$)
Used ($i=2$)	Yes ($j=2$)	y_{221}	y_{222}
	No ($j=1$)	y_{211}	y_{212}
Not Used ($i=1$)	Yes ($j=2$)	y_{121}	y_{122}
	No ($j=1$)	y_{111}	y_{112}

- Where $Y_{ijk} \sim \text{POI}(\mu_{ijk})$, $i = 1, 2$, $j = 1, 2$, $k = 1, 2$.
- The saturated model is:

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

- Try to identify simpler models which still fit the data well and are easy to interpret.
- Note that in this table only $y_{\bullet\bullet\bullet}$ is fixed and so only the intercept needs to be included by design.

R Dataset

Accident Data

```
  s e i      y
1 1 1 1 157342
2 1 1 2   1008
3 1 2 1   4624
4 1 2 2    497
5 2 1 1 411111
6 2 1 2    483
7 2 2 1   1105
8 2 2 2     14
```

R Code

```
acc.dat$s <- factor(acc.dat$s)
acc.dat$e <- factor(acc.dat$e)
acc.dat$i <- factor(acc.dat$i)
# Model 1: (VWZ) Saturated
model1 <- glm(y ~ s * e * i, family = poisson, data = acc.dat)
```

```
summary(model1)
# Model 2 (VW,VZ,WZ) Homogeneous Association (test vs Model
# 1)
model2 <- glm(y ~ s * e + s * i + e * i, family = poisson, data = acc.dat)
summary(model2)
1 - pchisq(model2$deviance - model1$deviance, model2$df.residual -
  model1$df.residual)
acc.dat$fv <- model2$fitted.values
acc.dat
# Model 3: (VW, VZ) Conditional Independence Model (test vs
# Model 2)
model3 <- glm(y ~ s * e + s * i, family = poisson, data = acc.dat)
model3$df.residual
model3$deviance
1 - pchisq(model3$deviance - model2$deviance, model3$df.residual -
  model2$df.residual)
# Model 4: (VW, WZ) Conditional Independence Model (test vs
# Model 2)
model4 <- glm(y ~ s * e + e * i, family = poisson, data = acc.dat)
model4$df.residual
model4$deviance
1 - pchisq(model4$deviance - model2$deviance, model4$df.residual -
  model2$df.residual)
# Model 5: (VZ, WZ) Conditional Independence Model (test vs
# Model 2)
model5 <- glm(y ~ s * i + e * i, family = poisson, data = acc.dat)
model5$df.residual
model5$deviance
1 - pchisq(model5$deviance - model2$deviance, model5$df.residual -
  model2$df.residual)
```

R Output: Model 1 (VWZ)

```
summary(model1)


Call:
glm(formula = y ~ s * e * i, family = poisson, data = acc.dat)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0

Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) 11.966177   0.002521 4746.547   <2e-16 ***
s2           0.960441   0.002964  323.985   <2e-16 ***
e2          -3.527162   0.014920 -236.398   <2e-16 ***
i2          -5.050454   0.031598 -159.836   <2e-16 ***
s2:e2       -2.391856   0.033616  -71.153   <2e-16 ***
s2:i2       -1.696148   0.055419  -30.606   <2e-16 ***
e2:i2        2.820028   0.056805   49.644   <2e-16 ***
s2:e2:i2    -0.441970   0.278627   -1.586    0.113
```

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

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1.6249e+06  on 7  degrees of freedom
Residual deviance: 7.4852e-11  on 0  degrees of freedom
AIC: 92.999

Number of Fisher Scoring iterations: 3
```

R Output: Model 2 (VW, VZ, WZ)

```
summary(model2)


Call:
glm(formula = y ~ s * e + s * i + e * i, family = poisson, data = acc.dat)

Deviance Residuals:
       1         2         3         4         5         6         7         8
 0.01731  -0.21583  -0.10095   0.30951  -0.01071   0.31400   0.20704  -1.59987

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.966133   0.002521 4746.70   <2e-16 ***
s2           0.960502   0.002964  324.03   <2e-16 ***
e2          -3.525634   0.014879 -236.95   <2e-16 ***
i2          -5.043620   0.031202 -161.65   <2e-16 ***
s2:e2       -2.399636   0.033340  -71.97   <2e-16 ***
s2:i2       -1.717321   0.054015  -31.79   <2e-16 ***
e2:i2        2.797795   0.055256   50.63   <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: 1.6249e+06  on 7  degrees of freedom
Residual deviance: 2.8540e+00  on 1  degrees of freedom
AIC: 93.853

Number of Fisher Scoring iterations: 3

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

[1] 0.09114565

acc.dat$fv <- model2$fitted.values
acc.dat

  s e i      y           fv
1 1 1 1 157342 157335.13193
```

```
2 1 1 2   1008   1014.86807
3 1 2 1   4624   4630.86807
4 1 2 2    497    490.13193
5 2 1 1 411111 411117.86807
6 2 1 2    483    476.13193
7 2 2 1   1105   1098.13193
8 2 2 2     14     20.86807
```

- H_0: $u_{222}^{VWZ} = 0$ versus H_A: $u_{222}^{VWZ} \neq 0$.

$$p = \mathbb{P}\left(\chi_1^2 > 2.854\right) = 0.09.$$

- Do not reject the null hypothesis that the fit of model 2 is adequate, as compared to model 1 (saturated).

R Output: Models 3 (VW, VZ)**, 4** (VW, WZ)**, 5** (VZ, WZ)

```
# Model 3: (VW, VZ) Conditional Independence Model (test vs
# Model 2)
model3 <- glm(y ~ s * e + s * i, family = poisson, data = acc.dat)
model3$df.residual
```

```
[1] 2
```

```
model3$deviance
```

```
[1] 1680.412
```

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

```
[1] 0
```

```
# Model 4: (VW, WZ) Conditional Independence Model (test vs
# Model 2)
model4 <- glm(y ~ s * e + e * i, family = poisson, data = acc.dat)
model4$df.residual
```

```
[1] 2
```

```
model4$deviance
```

```
[1] 1144.636
```

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

```
[1] 0
```

```
# Model 5: (VZ, WZ) Conditional Independence Model (test vs
# Model 2)
model5 <- glm(y ~ s * i + e * i, family = poisson, data = acc.dat)
model5$df.residual
```

```
[1] 2
```

```
model5$deviance
```

```
[1] 7133.978
```

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

```
[1] 0
```

- Reject the null hypotheses that the fit of models 3, 4, 5 are adequate, as compared to model 2.

Summary of Fitted Models

The following analysis of deviance table summarizes our findings.

Model	Form	Residual Deviance	Residual d.f.	p-value
1	(VWZ)	0	0	NA
2	(VW, VZ, WZ)	2.85	1	0.09 (vs 1)
3	(VW, VZ)	1680.41	2	< 0.001 (vs 2)
4	(VW, WZ)	1144.64	2	< 0.001 (vs 2)
5	(VZ, WZ)	7133.98	2	< 0.001 (vs 2)

- Conclude that Model 2 (VW, VZ, WZ) is most appropriate.
- Homogeneous Association: All variables are associated in a pairwise fashion, but this degree of association does not depend on the level of the third variable.

Interpretation of Model 2 (VW, VZ, WZ)

- Consider the following table of fitted and observed values from this model:

		Injury (Z)	
Seatbelt (V)	Ejected (W)	Non-fatal ($k = 1$)	Fatal ($k = 2$)
Used ($i = 2$)	Yes ($j = 2$)	$\hat{\mu}_{221} = 1098.1, y_{221} = 1105$	$\hat{\mu}_{222} = 20.9, y_{222} = 14$
	No ($j = 1$)	$\hat{\mu}_{211} = 411117.9, y_{211} = 411111$	$\hat{\mu}_{212} = 476.1, y_{212} = 483$
Not Used ($i = 1$)	Yes ($j = 2$)	$\hat{\mu}_{121} = 4630.9, y_{121} = 4624$	$\hat{\mu}_{122} = 490.1, y_{122} = 497$
	No ($j = 1$)	$\hat{\mu}_{111} = 157335.1, y_{111} = 157342$	$\hat{\mu}_{112} = 1014.9, y_{112} = 1008$

- And the following row percentages (across levels of VW):

		Injury (Z)	
Seatbelt (V)	Ejected (W)	Non-fatal	Fatal
Used	Yes	98.7	1.3
	No	99.9	0.1
Not Used	Yes	90.4	9.6
	No	99.4	0.6

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

- In previous Poisson log linear models (ships & rats examples) the regression coefficients had log Relative Rate interpretation
- Parameters for log linear models of contingency tables will have a log Odds Ratio interpretation!
- For $2 \times 2 \times K$ tables we can define:
  - Conditional Odds Ratio:
  $$\psi_{(k)}^{VW} = \frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}.$$
  - Marginal Odds Ratio:
  $$\psi^{VW} = \frac{\pi_{11\bullet}\pi_{22\bullet}}{\pi_{12\bullet}\pi_{21\bullet}}.$$

Conditional Odds Ratio $(2 \times 2 \times K)$

- The Odds Ratio of response $(V = 2)$ in subjects with $W = 2$ versus $W = 1$ at (conditional on being) level $Z = k$.

$$\begin{aligned}\psi_{(k)}^{VW} &= \frac{\mathbb{P}(V = 2 \mid W = 2, Z = k)/\,\mathbb{P}(V = 1 \mid W = 2, Z = k)}{\mathbb{P}(V = 2 \mid W = 1, Z = k)/\,\mathbb{P}(V = 1 \mid W = 1, Z = k)} \\ &= \frac{\pi_{22k}/\pi_{12k}}{\pi_{21k}/\pi_{11k}} \\ &= \frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}} \\ &= \frac{\pi_{22k}/\pi_{21k}}{\pi_{12k}/\pi_{11k}}.\end{aligned}$$

- This is also the Odds Ratio of response $(W = 2)$ in subjects with $V = 2$ versus $V = 1$ at (conditional on being) level $Z = k$.

Interpretation of Model 2 (VW, VZ, WZ)

Q: Find the (conditional OR) for a $\overbrace{\text{fatal}}^{Z=2}$ accident for those $\overbrace{\text{ejected}}^{W=2}$ versus $\overbrace{\text{not ejected}}^{W=1}$ among passengers who $\underbrace{\text{did not use}}_{V=1}$ their seatbelt.

$$\text{cOR} = \psi_{(1)}^{ZW} = \frac{\pi_{122}/\pi_{121}}{\pi_{112}/\pi_{111}}.$$

- Find the odds of a fatal accident $(Z = 2)$ in ejected $(W = 2)$, no seatbelt $(V = 1)$.

$$\frac{\mathbb{P}(Z = 2 \mid W = 2, V = 1)}{\mathbb{P}(Z = 1 \mid W = 2, V = 1)} = \frac{\mathbb{P}(Z = 2, W = 2, V = 1)}{\mathbb{P}(Z = 1, W = 2, V = 1)} = \frac{\pi_{122}}{\pi_{121}} = \frac{\mu_{122}/\mu_{\bullet\bullet\bullet}}{\mu_{121}/\mu_{\bullet\bullet\bullet}} = \frac{\mu_{122}}{\mu_{121}}.$$

V	W	Z	$\log(\mu_{ijk})$
1	2	2	$u + u_2^W + u_2^Z + u_{22}^{WZ}$
1	2	1	$u + u_2^W$
$\log(\mu_{122}/\mu_{121})$			$= u_2^Z + u_{22}^{WZ}$

- Find the odds of a fatal accident $(Z = 2)$ in not ejected $(W = 1)$, no seatbelt $(V = 1)$.

$$\frac{\mathbb{P}(Z = 2 \mid W = 1, V = 1)}{\mathbb{P}(Z = 1 \mid W = 1, V = 1)} = \frac{\mathbb{P}(Z = 2, W = 1, V = 1)}{\mathbb{P}(Z = 1, W = 1, V = 1)} = \frac{\pi_{112}}{\pi_{111}} = \frac{\mu_{112}/\mu_{\bullet\bullet\bullet}}{\mu_{111}/\mu_{\bullet\bullet\bullet}} = \frac{\mu_{112}}{\mu_{111}}.$$

V	W	Z	$\log(\mu_{ijk})$
1	1	2	$u + u_2^W + u_2^Z$
1	1	1	u
$\log(\mu_{122}/\mu_{121})$			$= u_2^Z$

- The expression for the (log) conditional OR is:

$$\log(\text{cOR}) = \log\left(\frac{\pi_{122}/\pi_{121}}{\pi_{112}/\pi_{111}}\right) = \log\left(\frac{\mu_{122}/\mu_{121}}{\mu_{112}/\mu_{111}}\right) = (u_2^Z + u_{22}^{WZ}) - u_2^Z = u_{22}^{WZ}.$$

- The estimate of the conditional OR is:

$$\widehat{\text{cOR}} = \exp\{\hat{u}_{22}^{WZ}\} = \exp\{2.80\} = 16.4.$$

Q: How would this change if the 3-way interaction term were included in the model?

- No change in the cOR for $V = 1$ (no seatbelt).
- Check that for $V = 2$ (seatbelt worn) the cOR becomes:

$$\psi_{(2)}^{ZW} = \frac{\pi_{222}/\pi_{221}}{\pi_{212}/\pi_{211}} = \exp\{u_{22}^{WZ} + u_{222}^{VWZ}\}.$$

V	W	Z	$\log(\mu_{ijk})$
2	2	2	$u + u_2^V + u_2^W + u_2^Z + u_{22}^{VW} + u_{22}^{VZ} + u_{22}^{WZ} + u_{222}^{VWZ}$
2	2	1	$u + u_2^V + u_2^W + u_{22}^{VW}$
$\log(\mu_{122}/\mu_{121})$			$= u_2^Z + u_{22}^{VZ} + u_{22}^{WZ} + u_{22}^{VWZ}$

V	W	Z	$\log(\mu_{ijk})$
2	1	2	$u + u_2^V + u_2^Z + u_{22}^{VZ}$
2	1	1	$u + u_2^V$
$\log(\mu_{122}/\mu_{121})$			$= u_2^Z + u_{22}^{VZ}$

$$(u_2^Z + u_{22}^{VZ} + u_{22}^{WZ} + u_{22}^{VWZ}) - (u_2^Z + u_{22}^{VZ}) = u_{22}^{WZ} + u_{222}^{VWZ}.$$

- We can construct several conditional odds ratios for this data set for the saturated model M1 and evaluate under M2:

Outcome	Comparison	At	Form (M1)	Value (M2)
$Z = 2$	$W = 2$ vs. $W = 1$	$V = 1$	$\exp\{u_{22}^{WZ}\}$	$\exp\{2.80\} = 16.4$
$Z = 2$	$W = 2$ vs. $W = 1$	$V = 2$	$\exp\{u_{22}^{WZ} + u_{222}^{VWZ}\}$	$\exp\{2.80 + 0\} = 16.4$
$Z = 2$	$V = 2$ vs. $V = 1$	$W = 1$	$\exp\{u_{22}^{VZ}\}$	$\exp\{-1.72\} = 0.18$
$Z = 2$	$V = 2$ vs. $V = 1$	$W = 2$	$\exp\{u_{22}^{VZ} + u_{222}^{VWZ}\}$	$\exp\{-1.72 + 0\} = 0.18$
$W = 2$	$V = 2$ vs. $V = 1$	$Z = 1$	$\exp\{u_{22}^{VW}\}$	$\exp\{-2.40\} = 0.09$
$W = 2$	$V = 2$ vs. $V = 1$	$Z = 2$	$\exp\{u_{22}^{VW} + u_{222}^{VWZ}\}$	$\exp\{-2.40 + 0\} = 0.09$

- Homogeneous Association: All variables are associated in a pairwise fashion, but this degree of association does not depend on the level of the third variable.
- That is, the conditional odds ratios between two factors are identical at all levels of the third factor.
- These odds ratios make sense since they suggest:
  - The relative odds of fatality among those ejected compared to those not ejected is 16.4,
  - The relative odds of fatality among those using a seatbelt compared to those who do not use a seatbelt is 0.18, and

  - The relative odds of ejection for those using a seatbelt compared to those who do not use a seatbelt is 0.09.

- The fact that we could not reduce the model further means these terms are all significant.
- The odds ratios relating to fatality could have been obtained from a logistic model, so it is natural to ask: What we have gained here?
- We are able to examine the relationship between all variables including V and W.

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

- Recall the best fitting model was Model 5 (GP, DP).
- Conditional Independence: The responders educational level (D) is conditionally independent of his/her gender (G), given his/her party affiliation (P).

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

- The regression parameters will have various log Odds Ratio interpretations.

"Odds Ratio" Definitions in 2-way Tables ($I \times J$)

- For a general $(I \times J)$ table, *many* types of "OR" can be defined.
  - Nominal Odds Ratios are formed by comparing back to a reference category (e.g $V = 1$, $W = 1$ in a 2-way table):

$$\psi_{ij}^{N\ VW} = \frac{\frac{\mathbb{P}(V=i|W=j)}{\mathbb{P}(V=1|W=j)}}{\frac{\mathbb{P}(V=i|W=1)}{\mathbb{P}(V=1|W=1)}} = \frac{\pi_{ij}\pi_{11}}{\pi_{1i}\pi_{1j}}.$$

  - Local Odds Ratios are formed by comparing 2 successive rows (i and $i+1$) and columns (j and $j+1$) of an $I \times J$ table:

$$\psi_{ij}^{L\ VW} = \frac{\frac{\mathbb{P}(V=i+1|W=j+1)}{\mathbb{P}(V=i|W=j+1)}}{\frac{\mathbb{P}(V=i+1|W=j)}{\mathbb{P}(V=i|W=j)}} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i+1,j}\pi_{i,j+1}}.$$

Odds Ratios for 3-way Tables ($I \times J \times K$)

- For a 3-way table consider conditional and marginal versions of the nominal or local odds ratios.
  - Conditional Local OR condition on the level of a third variable:

$$\psi^{L\ VW}_{ij(k)} = \frac{\pi_{i+1,j+1,k}/\pi_{i,j+1,k}}{\pi_{i+1,j,k}/\pi_{ijk}} = \frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i+1,j,k}\pi_{i,j+1,k}}.$$

  - Marginal Local OR ignore the level of a third variable:

$$\psi^{M\ VW}_{ij\bullet} = \frac{\pi_{i+1,j+1,\bullet}/\pi_{i,j+1,\bullet}}{\pi_{i+1,j,\bullet}/\pi_{ij\bullet}} = \frac{\pi_{ij\bullet}\pi_{i+1,j+1,\bullet}}{\pi_{i+1,j,\bullet}\pi_{i,j+1,\bullet}}.$$

- Most relevant when factor variables have a meaningful order.
- These aren't true odds ratios. Actually ratios of relative probabilities.

General Social Survey: Interpretation

$$\log(\mu_{ijk}) = u + u^G_i + u^D_j + u^P_k + u^{GP}_{ik} + u^{DP}_{jk}.$$

Q: Find an expression for the Conditional Local OR for being a "not strong Democrat" ($P=2$) versus a "strong Democrat" ($P=1$) comparing those with "a High School Degree" ($D=2$) to those with "less than High School" ($D=1$), among "males" ($G=1$).

$$\psi^{L\ DP}_{(i)jk} = \frac{\pi_{ijk}\pi_{i,j+1,k+1}}{\pi_{i,j+1,k}\pi_{i,j,k+1}} \quad \text{(general expression)}$$

$$\psi^{L\ DP}_{(1)11} = \frac{\pi_{111}\pi_{122}}{\pi_{121}\pi_{112}}.$$

$$\log(\psi^{L\ DP}_{(1)11}) = \log(\mu_{122}/\mu_{121}) - \log(\mu_{112}/\mu_{111}) = (u^P_2 + u^{DP}_{22}) - u^P_2 = u^{DP}_{22}.$$

G	D	P	$\log(\mu_{ijk})$
1	2	2	$u + u^D_2 + u^P_2 + u^{DP}_{22}$
1	2	1	$u + u^D_2$
$\log(\mu_{122}/\mu_{121})$			$= u^P_2 + u^{DP}_{22}$
1	1	2	$u + u^P_2$
1	1	1	u
$\log(\mu_{112}/\mu_{111})$			$= u^P_2$

$$\hat{\psi}^{L\ DP}_{(1)11} = \exp\{\hat{u}^{DP}_{22}\} = \exp\{0.32560\} = 1.38.$$

R Output: Model 5 (GP, DP)

```
Call:
glm(formula = Y ~ G * P + D * P, family = poisson, data = party)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6951  -0.3522   0.0008   0.3338   1.6843

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.18762    0.14133  22.554  < 2e-16 ***
Gfemale       0.47000    0.10408   4.516 6.31e-06 ***
P2           -0.17913    0.21421  -0.836 0.403026
P3           -0.37758    0.23105  -1.634 0.102218
P4            0.43821    0.18857   2.324 0.020134 *
P5           -0.74581    0.27684  -2.694 0.007059 **
P6           -0.96398    0.27169  -3.548 0.000388 ***
P7           -0.75026    0.25665  -2.923 0.003464 **
DHSc          1.07722    0.14587   7.385 1.53e-13 ***
DJunCol      -0.67740    0.21708  -3.121 0.001805 **
DBachelor    -0.11778    0.18366  -0.641 0.521316
DGraduate    -0.15415    0.18545  -0.831 0.405845
Gfemale:P2   -0.26994    0.15179  -1.778 0.075332 .
Gfemale:P3   -0.42419    0.16158  -2.625 0.008658 **
Gfemale:P4   -0.19499    0.15327  -1.272 0.203302
Gfemale:P5   -0.89609    0.19147  -4.680 2.87e-06 ***
Gfemale:P6   -0.31790    0.15528  -2.047 0.040631 *
Gfemale:P7   -0.30044    0.17572  -1.710 0.087303 .
P2:DHSc       0.32560    0.22128   1.471 0.141170
P3:DHSc       0.27922    0.24138   1.157 0.247374
P4:DHSc      -0.49327    0.19795  -2.492 0.012704 *
P5:DHSc       0.33505    0.29450   1.138 0.255252
P6:DHSc       0.91748    0.27943   3.283 0.001026 **
P7:DHSc       0.28887    0.26736   1.080 0.279943
P2:DJunCol    0.27193    0.32043   0.849 0.396082
P3:DJunCol    0.09548    0.35940   0.266 0.790501
P4:DJunCol   -0.69747    0.32260  -2.162 0.030618 *
P5:DJunCol    0.72869    0.38698   1.883 0.059698 .
P6:DJunCol    1.17817    0.35697   3.301 0.000965 ***
P7:DJunCol    0.02347    0.40503   0.058 0.953786
P2:DBachelor  0.09531    0.28050   0.340 0.734016
P3:DBachelor  0.63447    0.28405   2.234 0.025506 *
P4:DBachelor -0.91414    0.27836  -3.284 0.001023 **
P5:DBachelor  0.72869    0.33902   2.149 0.031601 *
P6:DBachelor  1.45278    0.31127   4.667 3.05e-06 ***
P7:DBachelor  0.66011    0.31143   2.120 0.034037 *
P2:DGraduate -0.28522    0.30182  -0.945 0.344669
P3:DGraduate -0.37648    0.33735  -1.116 0.264425
P4:DGraduate -1.05366    0.29043  -3.628 0.000286 ***
P5:DGraduate -0.48770    0.43246  -1.128 0.259431
P6:DGraduate  0.45426    0.34847   1.304 0.192375
P7:DGraduate  0.02632    0.34619   0.076 0.939403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

    Null deviance: 1257.316  on 69  degrees of freedom
Residual deviance:   29.323  on 28  degrees of freedom
AIC: 453.83

Number of Fisher Scoring iterations: 4
```

General Social Survey: Interpretation

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

Q: Find an expression for the Marginal Nominal OR of being a $\overbrace{\text{strong Republican}}^{P=7}$ versus a $\underbrace{\text{“strong Democrat”}}_{P=1}$ for $\underbrace{\text{“females”}}_{G=2}$ vs $\underbrace{\text{“males”}}_{G=1}$.

$$\psi_{i\bullet k}^{M\,GP} = \frac{\pi_{i\bullet k}\pi_{1\bullet 1}}{\pi_{i\bullet 1}\pi_{1\bullet k}} \quad \text{(general expression)}$$
$$\psi_{2\bullet 7}^{M\,GP} = \frac{\mu_{2\bullet 7}\mu_{1\bullet 1}}{\mu_{2\bullet 1}\mu_{1\bullet 7}}.$$

$$\log(\psi_{2\bullet 7}^{M\,GP}) = \log(\mu_{2\bullet 7}/\mu_{2\bullet 1}) - \log(\mu_{1\bullet 7}/\mu_{1\bullet 1}) = u_{27}^{GP}$$

G	D	P	$\log(\mu_{ijk})$
2	j	7	$u + u_2^G + u_j^D + u_7^P + u_{27}^{GP} + u_{7j}^{DP}$
2	j	1	$u + u_2^G + u_j^D$
$\log(\mu_{2\bullet 7}/\mu_{2\bullet 1})$			$= u_7^P + u_{27}^{GP} + u_{7j}^{DP}$
1	j	7	$u + u_j^D + u_7^P + u_{7j}^{DP}$
1	j	1	$u + u_j^D$
$\log(\mu_{1\bullet 7}/\mu_{1\bullet 1})$			$= u_7^P + u_{7j}^{DP}$

Q: Find a general expression for the Marginal Nominal OR for party affiliation and gender.

$$\psi_{i\bullet k}^{M\,GP} = \exp\{u_{ik}^{GP}\},\ i = 2,\ k = 2, \ldots, 7$$

Q: Find a general expression for the Marginal Nominal OR for party affiliation and highest degree earned.

$$\begin{aligned}\psi_{\bullet jk}^{M\,DP} &= \frac{\pi_{\bullet jk}\pi_{\bullet,j+1,k+1}}{\pi_{\bullet,j+1,k}\pi_{\bullet,j,k+1}} \\ &= \frac{\mu_{\bullet jk}\mu_{\bullet,j+1,k+1}}{\mu_{\bullet,j+1,k}\mu_{\bullet,j,k+1}} \\ &= \exp\{u_{jk}^{DP} + u_{j+1,k+1}^{DP} - u_{j+1,k}^{DP} - u_{j,k+1}^{DP}\}.\end{aligned}$$

$$\begin{aligned}\log(\psi^{M\,DP}_{\bullet jk}) &= \log(\mu_{\bullet jk}/\mu_{\bullet,j+1,k}) - \log(\mu_{\bullet,j,k+1}/\mu_{\bullet,j+1,k+1}) \\ &= (u^D_j + u^{DP}_{jk} - u^D_{j+1} - u^{DP}_{j+1,k}) - (u^D_j + u^{DP}_{j,k+1} - u^D_{j+1} - u^{DP}_{j+1,k+1}) \\ &= u^D_j + u^{DP}_{jk} - u^D_{j+1} - u^{DP}_{j+1,k} - u^D_j - u^{DP}_{j,k+1} + u^D_{j+1} + u^{DP}_{j+1,k+1} \\ &= u^{DP}_{jk} - u^{DP}_{j+1,k} - u^{DP}_{j,k+1} + u^{DP}_{j+1,k+1} \\ &= u^{DP}_{jk} + u^{DP}_{j+1,k+1} - u^{DP}_{j+1,k} - u^{DP}_{j,k+1}.\end{aligned}$$

Therefore,

$$\psi^{M\,DP}_{\bullet jk} = \exp\{u^{DP}_{jk} + u^{DP}_{j+1,k+1} - u^{DP}_{j+1,k} - u^{DP}_{j,k+1}\}.$$

G	D	P	$\log(\mu_{ijk})$
i	j	k	$u + u^G_i + u^D_j + u^P_k + u^{GP}_{ik} + u^{DP}_{jk}$
i	$j+1$	k	$u + u^G_i + u^D_{j+1} + u^P_k + u^{GP}_{ik} + u^{DP}_{j+1,k}$
$\log(\mu_{\bullet jk}/\mu_{\bullet,j+1,k})$			$= u^D_j + u^{DP}_{jk} - u^D_{j+1} - u^{DP}_{j+1,k}$
i	j	$k+1$	$u + u^G_i + u^D_j + u^P_{k+1} + u^{GP}_{i,k+1} + u^{DP}_{j,k+1}$
i	$j+1$	$k+1$	$u + u^G_i + u^D_{j+1} + u^P_{k+1} + u^{GP}_{i,k+1} + u^{DP}_{j+1,k+1}$
$\log(\mu_{\bullet,j,k+1}/\mu_{\bullet,j+1,k+1})$			$= u^D_j + u^{DP}_{j,k+1} - u^D_{j+1} - u^{DP}_{j+1,k+1}$

Topic 4a: Introduction to Overdispersion

Chapter 4: Introduction to Overdispersion

- Recall the Exponential Family:

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

- Canonical parameter: θ.
- Dispersion parameter: ϕ.
- Mean: $\mathbb{E}[Y] = b'(\theta) = \mu$.
- Variance: $\mathsf{Var}(Y) = b''(\theta)a(\phi)$.
- Up to now, we always had either $\phi = 1$ (Binomial & Poisson) or ϕ known (Normal with σ^2 known).

Introduction to Overdispersion

- Frequently we will have a poor fit of a GLM to the data because the variance of our model is too restrictive.
- Recall for the Poisson: $\mathbb{E}[Y] = \mathsf{Var}(Y) = \mu$.
- But what if we observe count data where $\mathsf{Var}(Y) > \mathbb{E}[Y]$?
  - Here, we say the data is overdispersed.
- Recall for the Binomial: $\mathbb{E}[Y] = n\pi$ and $\mathsf{Var}(Y) = n\pi(1-\pi)$.
- We will cover 2 methods for dealing with overdispersion:
  1. Ad hoc: Introduce and estimate a dispersion parameter ϕ for a distribution that doesn't naturally have one.
  2. Mixed Model: Introduce a new random variable which acts as a dispersion factor.

Why Adjust for Overdispersion?

- If the data is overdispersed, then generally $\mathsf{Var}(\hat{\beta}_j)$ will be underestimated.
- We adjust for overdispersion to:
  - "Correct" the fitted standard errors.
  - Increase the width of confidence intervals to reflect variation in the data.
  - Reduce the risk false positive findings for covariate effects (i.e., when we reject H_0: $\beta_j = 0$ because $\mathsf{se}(\hat{\beta}_j)$ is too small).

Ad Hoc Method for Poisson

- Consider one observation from a Poisson distribution:

$$f(y;\theta,\phi) = \frac{\mu^y e^{-\mu}}{y!} = \exp\{y\log(\mu) - \mu - \log(y!)\}.$$

$$\begin{array}{lll} \theta = \log(\mu) & b(\theta) = e^\theta & \mathbb{E}[Y] = b'(\theta) = e^\theta = \mu. \\ \phi = 1 & a(\phi) = 1 & \mathsf{Var}(Y) = b''(\theta)a(\phi) = e^\theta = \mu. \end{array}$$

- We want to allow for data where $\mathsf{Var}(Y) > \mathbb{E}[Y]$.
- Introduce a dispersion parameter $a(\phi) = \phi > 0$.
- Let $\mathsf{Var}(Y) = \mu\phi$ to allow for *extra Poisson variation*.
- This does not actually correspond to an actual probability model.
- How do we estimate ϕ?

Ad Hoc Method for any GLM

- Consider one observation from the exponential family:

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

  where we assume $a_i(\phi) = \phi/w_i$.
- We can then write the log-likelihood of a random sample as:

$$\begin{aligned} \ell(\theta;\phi) &= \sum \ell_i(\theta_i;y_i,\phi) \\ &= \sum w_i \frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i;\phi). \end{aligned}$$

- Consider a LR/Deviance test of
  - H_0: p-dim model constrained is adequate ($\hat{\theta}_i$).
  - H_A: q-dim model is adequate ($\tilde{\theta}_i$), $n \geq q > p$.
- The Deviance can be written as:

$$\begin{aligned} D^\star &= 2\big(\ell(\tilde{\theta},\phi) - \ell(\hat{\theta},\phi)\big) \\ &= 2\left(\sum w_i \frac{y_i\tilde{\theta}_i - b(\tilde{\theta}_i)}{\phi} - \sum w_i \frac{y_i\hat{\theta}_i - b(\hat{\theta}_i)}{\phi}\right) \\ &= \frac{D}{\phi}, \end{aligned}$$

  where D is the deviance from a LRT when $\phi = 1$, that is, the distribution with $a_i(\phi) = 1/w_i$ (easy to get from R).

$$D^\star = \frac{D}{\phi}.$$

- Recall H_0: unsaturated p-dim model is adequate vs q-dim super model.
- Scaled Deviance: $D^\star \sim \chi^2_{q-p}$ under H_0, which implies Deviance: $D \sim \phi\chi^2_{q-p}$ under H_0.
- Note: $q = n$ if the alternative model is the saturated model.
- Fact: $\mathbb{E}[\chi^2_m] = m$ and therefore $\mathbb{E}[\chi^2_{n-p}] = n - p$.
- Fit an unsaturated p-dim model with a GLM with $\phi = 1$ to estimate D.
- Check if $D \sim \chi^2_{n-p}$ by comparing to $\mathbb{E}[D] = n - p$.
  - If $D \gg n - p$, then this indicates overdispersion exists (need to estimate ϕ).
  - If $D \approx n - p$, then $\phi \approx 1$, and there's no overdispersion.
- How do we estimate ϕ?
- Fit an unsaturated p-dim model with a GLM with $\phi = 1$.
- Method of Moments estimator:

$$\mathbb{E}[D] = \phi(n - p) \implies \hat{\phi} = \frac{D}{n - p}.$$

- How do we use $\hat{\phi}$ to adjust standard errors?
- Unadjusted covariance matrix (from GLM with $\phi = 1$):

$$\mathsf{Cov}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}\boldsymbol{\mathcal{W}}\boldsymbol{X}^\top)^{-1} = \mathcal{I}^{-1}.$$

- Adjusted covariance matrix:

$$\mathsf{Cov}_{\mathsf{adj}}(\hat{\boldsymbol{\beta}}) \simeq \hat{\phi}(\boldsymbol{X}\boldsymbol{\mathcal{W}}\boldsymbol{X}^\top)^{-1} = \hat{\phi}\mathcal{I}^{-1}.$$

- Adjusted standard errors:

$$\mathsf{se}_{\mathsf{adj}}(\hat{\beta}_j) = \sqrt{\hat{\phi}}\,\mathsf{se}(\hat{\beta}_j).$$

Summary: Ad Hoc Method

1. Fit the usual GLM to the data and find the best fitting model.
2. Check for evidence of overdispersion (i.e., $D \gg n - p$).
3. If overdispersion is present, estimate

$$\hat{\phi} = \frac{D}{n - p}.$$

4. Adjusted covariance matrix and standard error estimates

$$\mathsf{Cov}_{\mathsf{adj}}(\hat{\boldsymbol{\beta}}) = \hat{\phi}\,\mathsf{Cov}(\hat{\boldsymbol{\beta}}) = \hat{\phi}\mathcal{I}^{-1}, \qquad \mathsf{se}_{\mathsf{adj}}(\hat{\beta}_j) = \sqrt{\hat{\phi}}\,\mathsf{se}(\hat{\beta}_j).$$

   - This does not change the estimates $\hat{\beta}_j$ from the GLM.
   - May change the significance of the estimates though.
   - With $\phi > 1$, confidence intervals will increase in width.

Application: Analysis of an Epilepsy Trial

- Clinical trial was conducted involving 59 patients with epilepsy.
- Patients were randomized to one of two treatments, a standard therapy or a new drug designed to reduce the number of epileptic attacks experienced.
- The primary response is the number of attacks experienced during the first two weeks after randomization.
- The data are given on the next slide where:
  - Y_k = number attacks in k^{th} period after randomization.
  - `treat` is the treatment indicator variable with `treat=1` for the experimental treatment and `treat=0` otherwise.
  - `prior` records the number of epileptic attacks experience for the month prior to entry into the study.
  - `age` is a patient age at randomization in years.

R Data for Univariate Analysis

- First consider analyses based on the data from the first two-week period after randomization, that is, Y_{i1}, $i = 1, 2, \ldots, n$.
- Poisson Model, that is, $Y_{i1} \sim \text{POI}(\mu_i)$, with $\log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$.
- Explanatory variables:

$$
\begin{aligned}
x_{i1} &= \mathbb{I}\{\texttt{treat=1}\}, \\
x_{i2} &= \mathbb{I}\{\texttt{prior}\}, \\
x_{i3} &= \mathbb{I}\{\texttt{age}\}.
\end{aligned}
$$

- We are primarily interested in the treatment effect β_1.

Data from Epilepsy Trial

We show the first 5 rows.

```
  treat prior age yi1 treatf treatft
1     0    11  31   5      0       0
2     0    11  30   3      0       0
3     0     6  25   2      0       0
4     0     8  36   4      0       0
5     0    66  22   7      0       0
```

R Code and Output: Poisson Model

```
poisson1 <- glm(yi1 ~ treatft + prior + age, family = poisson,
  data = epi.dat)
summary(poisson1)


Call:
glm(formula = yi1 ~ treatft + prior + age, family = poisson,
    data = epi.dat)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1032  -1.3062  -0.5186   0.2927   5.1109

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.2049069  0.2269314   0.903   0.3666
treatft1    -0.2046787  0.0895007  -2.287   0.0222 *
prior        0.0253958  0.0009733  26.092  < 2e-16 ***
age          0.0324881  0.0063375   5.126 2.95e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 746.44  on 58  degrees of freedom
Residual deviance: 197.61  on 55  degrees of freedom
AIC: 402.11

Number of Fisher Scoring iterations: 5
```

Results of Fitted Poisson Model

$$\log(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

Model	Parameter	Estimate	$\widehat{\text{se}}$
poisson1	β_1	-0.2047	0.0895

$$\widehat{\text{RR}} = \exp\{\hat{\beta}_1\} = \exp\{-0.2047\} = 0.815.$$

- The relative rate of seizures in the treatment group versus the control group over the first two weeks of the study controlling for prior seizure count and age.

Ad Hoc Method

- $D \gg n - p$, so we need to account for overdispersion in this model:

$$\hat{\phi} = \frac{D}{n-p} = \frac{197.61}{59-4} = 3.593.$$

- Use $\hat{\phi}$ to adjust the standard error estimates:

$$\widehat{\text{se}}_{\text{adj}}(\hat{\beta}_1) = \sqrt{\hat{\phi}}\widehat{\text{se}}(\hat{\beta}_1) = \sqrt{3.593}(0.0895) = 0.1696.$$

Model	Parameter	Estimate	$\widehat{se}$	$\widehat{se}_{\text{adj}}$
poisson1	β_1	-0.2047	0.0895	0.1696
	ϕ	3.593		

- In the Poisson model the treatment effect was statistically significant ($p = 0.0222$).

- Is β_1 statistically significant after account for the overdispersion?
- Test H_0: $\beta_1 = 0$ versus H_A: $\beta_1 \neq 0$ using a Wald test:

$$t = \frac{|\hat{\beta}_1 - 0|}{\widehat{\mathsf{se}}_{\mathsf{adj}}(\hat{\beta}_1)} = \frac{|-0.2047|}{0.1696} = 1.21.$$

$$p = \mathbb{P}(|Z| > 1.21) = 0.23.$$

- After adjustment for overdispersion, we do not reject the null hypothesis of no treatment effect.

Mixed Poisson Model

- To form a Mixed Model we introduce a new random variable u_i which acts as a dispersion factor. Assume:

$$\mathbb{E}[u_i] = 1 \qquad \text{and} \qquad \mathsf{Var}(u_i) = \phi.$$

- For a Poisson model let: $Y_i \mid u_i \sim \text{POI}(u_i\lambda)$, so that

$$f(y_i \mid u_i; \lambda) = \frac{(u_i\lambda)^{y_i} e^{-u_i\lambda}}{y_i!}, \; y = 0, 1, 2, \ldots.$$

- This is called a mixed Poisson model.
- The $u_i > 0$ factor deflates ($u_i < 1$) or inflates ($u_i > 1$) the mean response for the i^{th} subject relative to λ.
- Recall we assume $\mathbb{E}[u_i] = 1$ and $\mathsf{Var}(u_i) = \phi$.
- Find the unconditional mean and variance of Y_i:

$$\begin{aligned}
\mathbb{E}[Y_i] &= \mathbb{E}\big[\mathbb{E}[Y_i \mid u_i]\big] \\
&= \mathbb{E}[u_i\lambda] \\
&= \lambda\,\mathbb{E}[u_i] \\
&= \lambda = \text{(the population mean response)}.
\end{aligned}$$

$$\begin{aligned}
\mathsf{Var}(Y_i) &= \mathsf{Var}\big(\mathbb{E}[Y_i \mid u_i]\big) + \mathbb{E}\big[\mathsf{Var}(Y_i \mid u_i)\big] \\
&= \mathsf{Var}(u_i\lambda) + \mathbb{E}[u_i\lambda] \\
&= \lambda^2\,\mathsf{Var}(u_i) + \lambda\,\mathbb{E}[u_i] \\
&= \lambda^2\phi + \lambda \\
&= \lambda(1 + \lambda\phi).
\end{aligned}$$

- The variance is inflated by a factor of $1 + \lambda\phi$.
- Now we need to pick a distribution for $u_i > 0$.
- Assume u_i has a Gamma Distribution:

$$g(u_i; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} u_i^{\alpha-1} e^{-u_i/\beta},$$

  with $\mathbb{E}[u_i] = \alpha\beta$ and $\mathsf{Var}(u_i) = \alpha\beta^2$.
- With $\mathbb{E}[u_i] = 1$ and $\mathsf{Var}(u_i) = \phi$, this implies $\alpha = 1/\phi$ and $\beta = \phi$.
- Mixed Poisson model with Gamma distribution $\implies$ Negative Binomial distribution.

- We will derive the marginal (unconditional) likelihood.
- We've assumed that u_i has a Gamma Distribution:

$$g(u_i; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} u_i^{\alpha-1} e^{-u_i/\beta}.$$

- Fact: the probability mass function's integrate to one,

$$1 = \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} u^{\alpha-1} e^{-u/\beta} \, \mathrm{d}u.$$

- Which implies:

$$\Gamma(\alpha)\beta^\alpha = \int_0^\infty u^{\alpha-1} e^{-u/\beta} \, \mathrm{d}u.$$

- Therefore,

$$\begin{aligned}
p(y_i; \lambda, \phi) &= \int_0^\infty f(y_i \mid u_i, \lambda) g(u_i; \phi) \, \mathrm{d}u_i \\
&= \int_0^\infty \frac{(u_i\lambda)^{y_i} e^{-u_i\lambda}}{y_i!} \frac{u_i^{\alpha-1} e^{-u_i/\beta}}{\Gamma(\alpha)\beta^\alpha} \, \mathrm{d}u_i \\
&= \frac{\lambda^{y_i}}{y_i!\Gamma(\alpha)\beta^\alpha} \int_0^\infty u_i^{y_i+\alpha-1} e^{-u_i(\lambda+1/\beta)} \, \mathrm{d}u_i \\
&= \frac{\lambda^{y_i}}{y_i!\Gamma(\alpha)\beta^\alpha} \Gamma(y_i+\alpha) \left(\frac{\beta}{1+\beta\lambda}\right)^{y_i+\alpha} \\
&= \frac{\Gamma(y_i+\alpha)}{y_i!\Gamma(\alpha)} \left(\frac{\lambda\beta}{1+\lambda\beta}\right)^{y_i} \left(\frac{1}{1+\lambda\beta}\right)^{\alpha} \\
&= \frac{\Gamma(y_i+\phi^{-1})}{y_{i!}\Gamma(\phi^{-1})} \left(\frac{\lambda\phi}{1+\lambda\phi}\right)^{y_i} \left(\frac{1}{1+\lambda\phi}\right)^{\phi^{-1}}.
\end{aligned}$$

Negative Binomial Distribution

- The pmf of the Negative Binomial can be written as:

$$\mathbb{P}(X = x) = \frac{\Gamma(a+x)}{\Gamma(a)\Gamma(x+1)} \left(\frac{b}{1+b}\right)^x \left(\frac{1}{1+b}\right)^a,$$

  with $\mathbb{E}[X] = ab$, and $\mathsf{Var}(X) = ab(1+b)$.
- Here we have $Y_i \sim \mathsf{NB}(a = 1/\phi, b = \lambda\phi)$, where

$$\begin{aligned}
\mathbb{E}[Y_i] &= ab = \frac{1}{\phi}(\lambda\phi) = \lambda. \\
\mathsf{Var}(Y_i) &= ab(1+b) = \frac{1}{\phi}(\lambda\phi)(1+\lambda\phi) = \lambda(1+\lambda\phi).
\end{aligned}$$

- Therefore, we've shown that Poisson model mixed with Gamma distribution $\implies$ Negative Binomial distribution.

Negative Binomial Model

- Now consider including covariates in the model.
- Assume we are using a log link:
$$\log(\lambda_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$
- The likelihood for a sample of size n with $Y_1, \ldots, Y_n$ and $\boldsymbol{x}_i$ a $p \times 1$ vector of explanatory variables is:
$$L(\boldsymbol{\beta}, \phi) = \prod_{i=1}^{n} \left(\frac{\Gamma(y_i + \phi^{-1})}{y_i! \Gamma(\phi^{-1})} \left(\frac{\phi e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + \phi e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \right)^{y_i} \left(\frac{1}{1 + \phi e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \right)^{\phi^{-1}} \right).$$
- Not a member the exponential family (unless ϕ known).
- Use iterative maximization in R: `glm.nb()` from `MASS` library.
  - Maximize $\ell(\boldsymbol{\beta}, \hat{\phi}^{(r)})$ at the current estimate $\hat{\phi}^{(r)} \implies \hat{\boldsymbol{\beta}}^{(r)}$ (IRWLS).
  - Maximize $\ell(\boldsymbol{\beta}^{(r)}, \phi)$ with respect to $\phi \implies \hat{\phi}^{(r+1)}$.

WEEK 12
1129 to 3rd December

Topic 4b: Poisson Overdispersion

Methods for Handling Overdispersion

1. Ad hoc Method:
   1. Fit the usual GLM to the data and find the best fitting model.
   2. Check for evidence of overdispersion (i.e., $D \gg n - p$).
   3. If overdispersion is present, estimate
   $$\hat{\phi} = \frac{D}{n - p}.$$
   4. Adjusted covariance matrix and standard error estimates
   $$\mathsf{Cov}_{\mathsf{adj}}(\hat{\boldsymbol{\beta}}) \simeq \hat{\phi}\, \mathsf{Cov}(\hat{\boldsymbol{\beta}}) = \hat{\phi} \mathcal{I}^{-1}, \qquad \mathsf{se}_{\mathsf{adj}}(\hat{\beta}_j) = \sqrt{\hat{\phi}}\, \mathsf{se}(\hat{\beta}_j).$$
2. Mixed Model Method:
   - For a Poisson model, let $Y_i \mid u_i \sim \text{POI}(u_i \lambda)$.
   - Introduce dispersion factor $u_i \sim \text{GAM}(\alpha = 1/\phi, \beta = \phi)$.
   - Then, $Y_i \sim \text{NB}(a = 1/\phi, b = \lambda\phi)$ with $\mathbb{E}[Y_i] = \lambda_i$ and $\mathsf{Var}(Y_i) = \lambda_i(1 + \lambda_i \phi)$.

Application: Analysis of an Epilepsy Trial

- Clinical trial was conducted involving 59 patients with epilepsy.
- Patients were randomized to one of two treatments, a standard therapy or a new drug designed to reduce the number of epileptic attacks experienced.
- The primary response is the number of attacks experienced during the first two weeks after randomization.
- The data are given on the next slide where:
  - Y_k = number attacks in k^{th} period after randomization.
  - `treat` is the treatment indicator variable with `treat=1` for the experimental treatment and `treat=0` otherwise.
  - `prior` records the number of epileptic attacks experience for the month prior to entry into the study.
  - `age` is a patient age at randomization in years.

Univariate Analyses

- First consider analyses based on the data from the first two week period after randomization, that is, Y_{i1}, $i = 1, 2, \ldots, n$.
- Poisson Model, that is, $Y_i \sim \text{POI}(\mu_i)$:
$$\log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$
- Negative Binomial Model, that is, $Y_{i1} \mid u_i \sim \text{POI}(u_i \lambda_i)$ and $u_i \sim \text{GAM}(1/\phi, \phi)$. Regression:
$$\begin{aligned} \mu_i &= \mathbb{E}[Y_{i1} \mid u_i] \\ \mu_i &= u_i \lambda_i \\ \log(\mu_i) &= \log(u_i) + \log(\lambda_i) \\ \log(\mu_i) &= \alpha_i + \boldsymbol{x}_i^\top \boldsymbol{\beta}. \end{aligned}$$
- $\alpha_i =$ log gamma random variable called a "random effect", $\mathbb{E}[\alpha_i] = 0$.

R Program

```
epi.dat <- read.table("epi.dat", header = T)
epi.dat$treatf <- factor(epi.dat$treat)
attach(epi.dat)
library(MASS)
# contrast findings from Poisson and negative binomial
# regression
poisson1 <- glm(yi1 ~ treatf + prior + age, family = poisson,
  data = epi.dat)
summary(poisson1)
epi.dat$rdeviance1 <- residuals.glm(poisson1, type = "deviance")
epi.dat$fitted.values1 <- poisson1$fitted.values
negbin2 <- glm.nb(yi1 ~ treatf + prior + age, link = log, init.theta = 1,
  trace = T, data = epi.dat)
summary(negbin2)
epi.dat$rdeviance2 <- residuals.glm(negbin2, type = "deviance")
epi.dat$fitted.values2 <- negbin2$fitted.values
epi.dat
# Constructing deviance residual plots
plot(log(epi.dat$fitted.values1), epi.dat$rdeviance1, ylim = c(-5,
  5), xlab = "LOG FITTED VALUES", ylab = "DEVIANCE RESIDUALS",
  main = "POISSON MODEL")
abline(h = -2, lty = 2)
abline(h = 2, lty = 2)
plot(log(epi.dat$fitted.values2), epi.dat$rdeviance2, ylim = c(-5,
  5), xlab = "LOG FITTED VALUES", ylab = "DEVIANCE RESIDUALS",
  main = "NEG BIN MODEL")
abline(h = -2, lty = 2)
abline(h = 2, lty = 2)
# Fitting some additional negative binomial models
negbin3 <- glm.nb(yi1 ~ treatf, link = log, init.theta = 1, trace = T)
summary(negbin3)
```

R Output: Poisson Model

```
Call:
glm(formula = yi1 ~ treatf + prior + age, family = poisson, data = epi.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1032  -1.3062  -0.5186   0.2927   5.1109

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.2049069  0.2269314   0.903   0.3666
treatf1     -0.2046787  0.0895007  -2.287   0.0222 *
prior        0.0253958  0.0009733  26.092  < 2e-16 ***
age          0.0324881  0.0063375   5.126 2.95e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 746.44  on 58  degrees of freedom
Residual deviance: 197.61  on 55  degrees of freedom
AIC: 402.11

Number of Fisher Scoring iterations: 5
```

Results of Fitted Poisson Model

- Estimated relative rate of seizures in treatment versus control:

$$\widehat{\text{RR}} = \exp\{\hat{\beta}_1\} = \exp\{-0.2047\} = 0.815.$$

- Ad hoc estimate of dispersion factor and adjusted standard error for $\hat{\beta}_1$:

$$\hat{\phi} = \frac{D}{n-p} = \frac{197.61}{59-4} = 3.593.$$

$$\widehat{\text{se}}_{\text{adj}}(\hat{\beta}_1) = \sqrt{\hat{\phi}}\widehat{\text{se}}(\hat{\beta}_1)\sqrt{3.593}(0.0895) = 0.1696.$$

- Adjusted Wald-based hypothesis test of H_0: $\beta_1 = 0$ versus H_A: $\beta_1 \neq 0$:

$$p = \mathbb{P}\left(|Z| > \frac{|-0.2047|}{0.1696}\right) = 0.23.$$

- Adjusted 95% confidence interval for the relative rate:

$$\exp\{\hat{\beta}_1 \pm z_{0.975}\,\widehat{\text{se}}_{\text{adj}}(\hat{\beta}_1)\} = \exp\{-0.2047 \pm 1.96(0.1696)\} = (0.58, 1.14).$$

R Output: Negative Binomial Model

```
Call:
glm.nb(formula = yi1 ~ treatf + prior + age, data = epi.dat,
    trace = T, init.theta = 3.078568736, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4126  -0.7686  -0.2725   0.2993   2.7605

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.282960   0.423819   0.668   0.5044
treatf1     -0.330196   0.185378  -1.781   0.0749 .
prior        0.028080   0.003171   8.855   <2e-16 ***
age          0.028195   0.012390   2.276   0.0229 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(3.0786) family taken to be 1)

    Null deviance: 172.151  on 58  degrees of freedom
Residual deviance:  63.831  on 55  degrees of freedom
AIC: 333.51

Number of Fisher Scoring iterations: 1


              Theta:  3.079
          Std. Err.:  0.875

 2 x log-likelihood:  -323.513
```

Results of Fitted Models

Model	Parameter	Estimate	$\widehat{\text{se}}$	$\widehat{\text{se}}_{\text{adj}}$
poisson1	β_1	-0.2047	0.0895	0.1696
	ϕ	3.593		
negbin2	β_1	-0.3302	0.1854	
	θ	3.079	0.875	

- For R Negative Binomial Model, $\theta = \phi^{-1}$, so

$$\text{Var}(Y_i) = \lambda_i(1 + \lambda_i\phi) = \lambda_i(1 + \lambda_i/\theta).$$

- Note: ϕ from Poisson is not the same as ϕ in Negative Binomial.
  - Poisson: $\widehat{\text{se}}_{\text{adj}}(\hat{\beta}_j) = \sqrt{\hat{\phi}}\widehat{\text{se}}(\hat{\beta}_j)$.
  - Negative Binomial: $\phi = \theta^{-1}$ already incorporated into standard error estimates.
- $\widehat{\text{se}}$ Negative Binomial is larger than naive Poisson $\widehat{\text{se}}$ because the Negative Binomial model accounts for the overdispersion.
- $\widehat{\text{se}}_{\text{adj}}$ from the Poisson is comparable to $\widehat{\text{se}}$ from Negative Binomial.

Results of Fitted Negative Binomial Model

- Estimated relative rate of seizures in treatment versus control:

$$\widehat{\text{RR}} = \exp\{\hat{\beta}_1\} = \exp\{-0.3302\} = 0.719.$$

- Wald-based hypothesis test of H_0: $\beta_1 = 0$ versus H_A: $\beta_1 \neq 0$:

$$p = \mathbb{P}\left(|Z| > \frac{|-0.3302|}{0.1854}\right) = 0.0749.$$

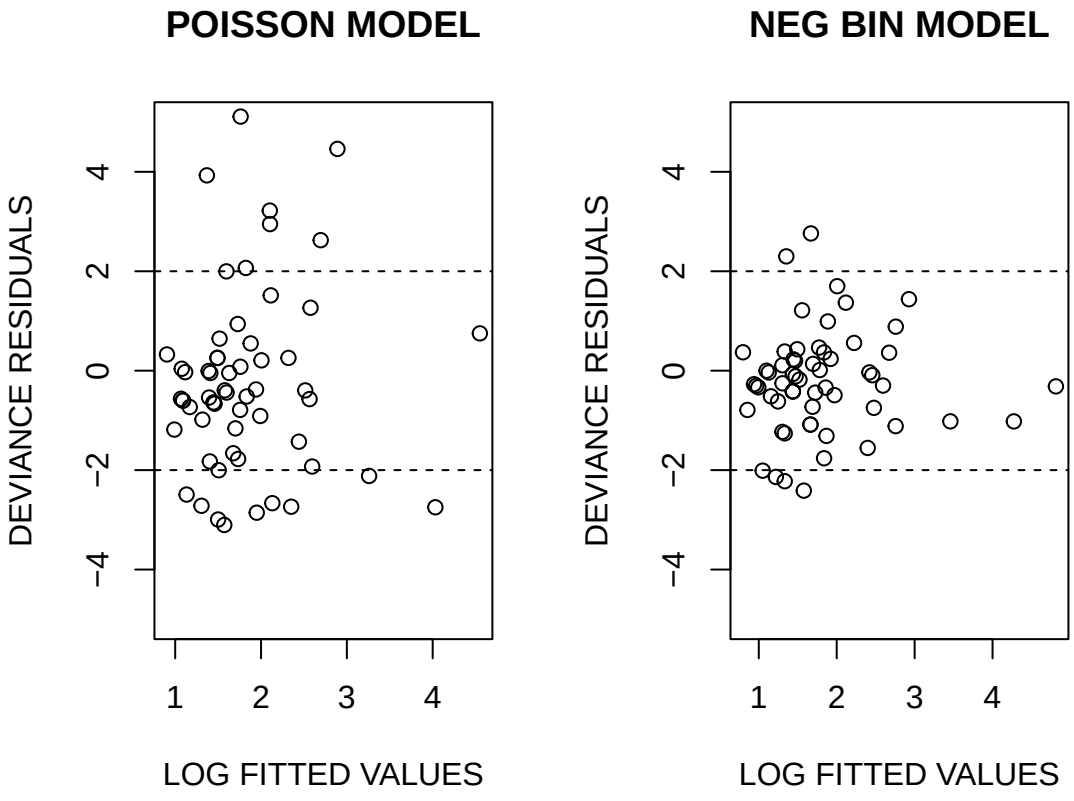
- 95% confidence interval for the relative rate:

$$\exp\{\hat{\beta}_1 \pm z_{0.975}\, \widehat{\text{se}}_{\text{adj}}(\hat{\beta}_1)\} = \exp\{-0.3302 \pm 1.96(0.1854)\} = (0.50, 1.03).$$

Poisson and Negative Binomial Models — Remarks

- The estimates and standard errors are different in the Poisson and Negative Binomial models.
- The estimates are different in part because the observations are weighted differently for the Poisson and Negative Binomial estimating equations.
- The standard errors are larger with the Negative Binomial model because it accounts for more variability in the data (which is needed here).
- Notice the `treat` variable is only statistically significant in the Poisson model.

Residual Plots for Poisson and Negative Binomial Models



R Output: Alternative Negative Binomial Model

```
Call:
glm.nb(formula = yi1 ~ treatf, trace = T, init.theta = 0.8738380555,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0736  -1.0095  -0.5943   0.1446   3.7245

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.23614    0.21139  10.578   <2e-16 ***
treatf1     -0.08663    0.29217  -0.297    0.767
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.8738) family taken to be 1)

    Null deviance: 66.075  on 58  degrees of freedom
Residual deviance: 65.987  on 57  degrees of freedom
AIC: 388.35

Number of Fisher Scoring iterations: 1


              Theta:  0.874
          Std. Err.:  0.166

 2 x log-likelihood:  -382.355
```

Results of Fitted Models

Model	Parameter	Estimate	$\widehat{\text{se}}$	$\widehat{\text{se}}_{\text{adj}}$
poisson1	β_1	-0.2047	0.0895	0.1696
	ϕ	3.593		
negbin2	β_1	-0.3302	0.1854	
	θ	3.079	0.875	
negbin3	β_1	-0.0866	0.2922	
	θ	0.874	0.166	

- Recall for the Negative Binomial that $\text{Var}(Y_i) = \lambda_i(1 + \lambda_i\phi) = \lambda_i(1 + \lambda_i/\theta)$.
- Compare the θ estimates from the two Negative Binomial Models:

$$\hat{\phi}_2 = \theta_2^{-1} = (3.079)^{-1} = 0.325.$$

$$\hat{\phi}_3 = \theta_3^{-1} = (0.874)^{-1} = 1.144,$$

Univariate Analyses — Final Remarks

- $\hat{\phi}_3$ is much larger than $\hat{\phi}_2$ because the `negbin3` model excludes the `prior` count and `age` variables.

- This makes sense since they explain much of the variability between the subjects for the rate of events.
- For Negative Binomial: $\phi = 0$ would imply no overdispersion in the data.
- To get $\mathsf{Var}(\phi)$ use the δ-method ($\mathsf{Var}(\theta^{-1}) \neq 1/\mathsf{Var}(\theta)$).
- Note that Overdispersion can be caused by a number of factors including:
  - Missing important explanatory variables.
  - Excess variation that can not be explained by Poisson model.
  - Non-independent observations (e.g., clustered data).

Adaptation to Clustered Count Data

- Up to now we have always assumed responses Y_i are iid.
- Now consider the following data structure:

$$\begin{array}{ccc} y_{11} & \cdots & y_{1n_1} \\ y_{21} & \cdots & y_{2n_2} \\ \vdots & \ddots & \vdots \\ y_{K1} & \cdots & y_{Kn_K}, \end{array}$$

  where $i = 1, \ldots, K$ are clusters, and $j = 1, \ldots, n_i$ are observations per cluster.
- $Y_{ij} =$ response for observation j of cluster i.
- Expect observations within the same cluster to be correlated.
- For example, cluster = families, litters, schools, etc.
- Assume $Y_{ij} \mid u_i \sim \text{POI}(u_i\lambda)$ independently.
  - Observation from same cluster are independent given u_i.
  - Observation from different clusters are independent.
- Assume $\mathbb{E}[u_i] = 1$ and $\mathsf{Var}(u_i) = \phi$.
- Then, $\mathbb{E}[Y_{ij}] = \lambda$ and $\mathsf{Var}(Y_{ij}) = \lambda(1 + \lambda\phi)$ as before.
- Correlation within clusters?

$$\begin{aligned} \mathsf{Cov}(Y_{ij}, Y_{ik}) &= \mathsf{Cov}\big(\mathbb{E}[Y_{ij} \mid u_i], \mathbb{E}[Y_{ik} \mid u_i]\big) + \mathbb{E}\big[\mathsf{Cov}(Y_{ij}, Y_{ik} \mid u_i)\big] \\ &= \mathsf{Cov}(u_i\lambda, u_i\lambda) + \mathbb{E}[0] \\ &= \lambda^2 \mathsf{Var}(u_i) \\ &= \lambda^2\phi. \end{aligned}$$

$$\mathsf{Corr}(Y_{ij}, Y_{ik}) = \frac{\mathsf{Cov}(Y_{ij}, Y_{ik})}{\sqrt{\mathsf{Var}(Y_{ij})\,\mathsf{Var}(Y_{ik})}} = \frac{\lambda^2\phi}{\lambda(1 + \lambda\phi)} = \frac{\lambda\phi}{1 + \lambda\phi}.$$

- We have a model which accommodates a correlation of responses within clusters.

Application: Joint Analyses of an Epilepsy Trial

- We now consider analyses based on the full 8 weeks of follow-up data.
- Four responses per subject: $Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}$.
- Consider the total seizure count: $Y_{i\bullet} = \sum_{j=1}^{4} Y_{ij}$.
- Assume $Y_{ij} \mid u_i \sim \text{POI}(u_i \lambda_i)$ independently.
- This implies $Y_{i\bullet} \mid u_i \sim \text{POI}(4u_i\lambda_i)$.
- Can show that $Y_{i\bullet}$ has an (almost) Negative Binomial distribution (Problem 4.3).

Problem 4.3

$$\begin{aligned}\mathbb{P}(Y_{i1}, \ldots, Y_{i4}) &= \int_0^\infty \prod_{j=1}^{4} \underbrace{p(y_{ij} \mid u_i\lambda_i)}_{\text{POI}(u_i\lambda_i)} \underbrace{f(u_i;\phi)}_{\text{GAM}(\alpha,\beta)} \, \mathrm{d}u_i \\ &\vdots \\ &\propto \frac{\Gamma(y_{i\bullet}+\alpha)}{\Gamma(\alpha)\prod y_{ij}!} \left(\frac{4\lambda_i\beta}{1+4\lambda_i\beta}\right)^{y_{i\bullet}} \left(\frac{1}{1+4\lambda_i\beta}\right)^{\alpha}.\end{aligned}$$

- This is proportional to a Negative Binomial distribution.
- $Y_{i\bullet}$ is sufficient for the joint distribution of $Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}$.
  - Factorization Theorem: If you can write $p(x) = h(x)g\big(\theta, T(x)\big)$, then $T(x)$ is a sufficient statistic.

Epilepsy Trial Joint Analysis Data

We show the first 5 rows.

```
  id treat prior age yidot treatf
1  1     0    11  31    14      0
2  2     0    11  30    14      0
3  3     0     6  25    11      0
4  4     0     8  36    13      0
5  5     0    66  22    55      0
```

- `yidot` is the sum of the seizure counts for each of the four counts obtained every two weeks.
- Explanatory variables:

$$\begin{aligned}x_{i1} &= \mathbb{I}\{\texttt{treat=1}\}, \\ x_{i2} &= \mathbb{I}\{\texttt{prior}\}, \\ x_{i3} &= \mathbb{I}\{\texttt{age}\}.\end{aligned}$$

- We are primarily interested in the treatment effect β_1.

Negative Binomial Model

$$\begin{aligned}\mu_i &= \mathbb{E}[Y_{i\bullet} \mid u_i] \\ \mu_i &= 4u_i\lambda_i \\ \log(\mu_i) &= \log(u_i) + \log(4) + \log(\lambda_i) \\ \log(\mu_i) &= \alpha_i + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.\end{aligned}$$

- α_i is an unobservable random effect, $\mathbb{E}[\alpha_i] = 0$.
- We could include $\log(4)$ as an offset or let it be absorbed into the intercept term:

$$\beta_0 = \beta_0^\star + \log(4).$$

R Program

```r
epi8.dat <- read.table("epi8.dat", header = T)
epi8.dat$treatf <- factor(epi.dat$treat)
epi.dat
# fitting the negative binomial model for clustered count
# data
joint <- glm.nb(yidot ~ treatf + prior + age, link = log, init.theta = 1,
  trace = T, data = epi8.dat)
summary(joint)
```

R Output: Joint Negative Binomial Model

```
Call:
glm.nb(formula = yidot ~ treatf + prior + age, data = epi8.dat,
    trace = T, init.theta = 3.35735873, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3405  -0.7920  -0.1943   0.2992   2.6623

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.060124   0.347696   5.925 3.12e-09 ***
treatf1     -0.212428   0.153399  -1.385    0.166
prior        0.027540   0.002811   9.796  < 2e-16 ***
age          0.012689   0.010326   1.229    0.219
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(3.3574) family taken to be 1)

    Null deviance: 181.136  on 58  degrees of freedom
Residual deviance:  63.697  on 55  degrees of freedom
AIC: 476.77

Number of Fisher Scoring iterations: 1
```

```
            Theta:  3.357
        Std. Err.:  0.709

2 x log-likelihood:  -466.767
```

Joint Negative Binomial Model

Model	Parameter	Estimate	$\widehat{\text{se}}$	$\widehat{\text{se}}_{\text{adj}}$
joint	β_1	-0.2124	0.1534	
	θ	3.357	0.709	
	ϕ	0.2979		

Q1: Based on the joint model give an estimate and 95% confidence interval for the relative rate (over 8 weeks) of epileptic attacks for treated versus control subjects.

- Estimated relative rate of seizures in treatment versus control:

$$\widehat{\text{RR}} = \exp\{\hat{\beta}_1\} = \exp\{-0.2124\} = 0.809.$$

- 95% confidence interval for the relative rate:

$$\exp\{\hat{\beta}_1 \pm z_{0.975}\widehat{\text{se}}(\hat{\beta}_1)\} = \exp\{-0.2124 \pm 1.96(0.1534)\} = (0.60, 1.09).$$

Q2: Based on the joint model estimate the correlation between the first and third responses (Y_{i1} & Y_{i3}) for an untreated subject with a prior seizure count of 11, age 31.

- First, we need an estimate of $\hat{\lambda}_i$ for this subject:

$$\begin{aligned}
\log(4) + \log(\hat{\lambda}_i) &= \hat{\beta}_0 + \hat{\beta}_1(0) + \hat{\beta}_2(11) + \hat{\beta}_3(31) \\
\hat{\lambda}_i &= \exp\{2.0601 + 0.0275(11) + 0.0127(31)\}/4 \\
&= \exp\{2.756\}/4 \\
&= 3.936.
\end{aligned}$$

- Now, find the Correlation:

$$\hat{\rho} = \frac{\hat{\lambda}_i\hat{\phi}}{1 + \hat{\lambda}_i\hat{\phi}} = \frac{3.936(1/3.357)}{1 + 3.936(1/3.357)} = 0.54.$$

- Moderate positive correlation between seizure counts within the same subject across time periods.

Topic 4c: Binomial Overdispersion

Origin of Overdispersion for Binomial Responses

- Recall two methods of dealing with Poisson overdispersion:
  - Ad hoc method,

  - Mixed model.

- These can also be used to deal with overdispersion with Binomial data.
- For example, extra binomial variation often arises due to unaccounted for clustering in the population.
- When sampling from populations with clustering present the assumptions necessary for the binomial distribution are violated (i.e., independent and identically distributed binary outcomes).
- Examples of clusters: families, classes, neighbourhoods, litters, repeated measures on individuals.

Clustered Binomial Data

- Suppose a pop consists of a number of clusters each of size k.
- Suppose m individuals are sampled from m/k clusters:

$$\begin{aligned} Y_{ij} &= 1 \text{ or } 0 && j^{\text{th}} \text{ response in the } i^{\text{th}} \text{ cluster} \\ Y_{i\bullet} &= \sum_{j=1}^{k} Y_{ij} && \text{total responses in the } i^{\text{th}} \text{ cluster} \\ Y_{\bullet\bullet} &= \sum_{i=1}^{m/k} \sum_{j=1}^{k} Y_{ij} && \text{grand total} \end{aligned}$$

- Overdispersion is induced by assuming clusters have different response probabilities:

$$\begin{aligned} Y_{ij} \mid \pi_i &\sim \text{BIN}(1, \pi_i) && \text{independent observations } j = 1, \ldots, k \\ Y_{i\bullet} \mid \pi_i &\sim \text{BIN}(k, \pi_i) && \text{independent clusters } i = 1, \ldots, m/k \end{aligned}$$

- Consider a setting where:

$$\mathbb{E}[\pi_i] = \pi, \qquad \mathsf{Var}(\pi_i) = \rho\pi(1-\pi), \qquad 0 < \rho < 1.$$

- π_i is analogous to u_i in Poisson setting.
- With $\mathbb{E}[\pi_i] = \pi$ and $\mathsf{Var}(\pi_i) = \rho\pi(1-\pi)$ examine the effect at three levels: individual, cluster and grand/overall total.

1. Individual Level — Clustered Binomial Data

$$\mathbb{E}[Y_{ij}] = \mathbb{E}\big[\mathbb{E}[Y_{ij} \mid \pi_i]\big] = \mathbb{E}[\pi_i] = \pi.$$

$$\begin{aligned} \mathsf{Var}(Y_{ij}) &= \mathbb{E}\big[\mathsf{Var}(Y_{ij} \mid \pi_i)\big] + \mathsf{Var}\big(\mathbb{E}[Y_{ij} \mid \pi_i]\big) \\ &= \mathbb{E}\big[\pi_i(1-\pi_i)\big] + \mathsf{Var}(\pi_i) \\ &= \mathbb{E}[\pi_i] - \mathbb{E}[\pi_i^2] + \mathbb{E}[\pi_i^2] - \mathbb{E}[\pi_i]^2 \\ &= \pi - \pi^2 \\ &= \pi(1-\pi), \end{aligned}$$

as expected for a Bernoulli random variable.

$$\begin{aligned}\mathsf{Cov}(Y_{ij}, Y_{ik}) &= \mathbb{E}\big[\mathsf{Cov}(Y_{ij}, Y_{ik} \mid \pi_i)\big] + \mathsf{Cov}\big(\mathbb{E}[Y_{ij} \mid \pi_i], \mathbb{E}[Y_{ik} \mid \pi_i]\big) \\ &= \mathbb{E}[0] + \mathsf{Cov}(\pi_i, \pi_i) \\ &= \mathsf{Var}(\pi_i) \\ &= \rho\pi(1-\pi).\end{aligned}$$

$$\begin{aligned}\mathsf{Corr}(Y_{ij}, Y_{ik}) &= \frac{\mathsf{Cov}(Y_{ij}, Y_{ik})}{\sqrt{\mathsf{Var}(Y_{ij})\,\mathsf{Var}(Y_{ik})}} \\ &= \frac{\rho\pi(1-\pi)}{\pi(1-\pi)} \\ &= \rho.\end{aligned}$$

- $\rho =$ intraclass correlation coefficient (accounts for correlation within observations from the same cluster).

2. Cluster Level — Clustered Binomial Data

$$\mathbb{E}[Y_{i\bullet}] = \mathbb{E}\big[\mathbb{E}[Y_{i\bullet} \mid \pi_i]\big] = \mathbb{E}[k\pi_i] = k\pi.$$

$$\begin{aligned}\mathsf{Var}(Y_{i\bullet}) &= \mathbb{E}\big[\mathsf{Var}(Y_{i\bullet} \mid \pi_i)\big] + \mathsf{Var}\big(\mathbb{E}[Y_{i\bullet} \mid \pi_i]\big) \\ &= \mathbb{E}\big[k\pi_i(1-\pi_i)\big] + \mathsf{Var}(k\pi_i) \\ &= k\,\mathbb{E}[\pi_i] - k\,\mathbb{E}[\pi_i^2] + k^2\,\mathsf{Var}(\pi_i) \\ &= k\,\mathbb{E}[\pi_i] - k\big(\mathsf{Var}(\pi_i) + \mathbb{E}[\pi_i]^2\big) + k^2\,\mathsf{Var}(\pi_i) \\ &= k(k-1)\rho\pi(1-\pi) + k\pi - k\pi^2 \\ &= k(k-1)\rho\pi(1-\pi) + k\pi(1-\pi) \\ &= k\pi(1-\pi)\big((k-1)\rho + 1\big).\end{aligned}$$

- $k\pi(1-\pi)$ is the standard variance for $Y_{i\bullet} \sim \text{BIN}(k, \pi)$.
- Dispersion parameter $\sigma^2 = \big((k-1)\rho + 1\big)$ accounts for overdispersion at the clustered level.

3. Grand Total Level — Clustered Binomial Data

$$Y = Y_{\bullet\bullet} = \sum_{i=1}^{m/k} \sum_{j=1}^{k} Y_{ij} = \sum_{i=1}^{m/k} Y_{i\bullet}.$$

$$\begin{aligned}\mathbb{E}[Y_{\bullet\bullet}] &= \sum \mathbb{E}[Y_{i\bullet}] \\ &= \left(\frac{m}{k}\right) k\pi \\ &= m\pi.\end{aligned}$$

$$\begin{aligned}\mathsf{Var}(Y_{\bullet\bullet}) &= \sum \mathsf{Var}(Y_{i\bullet}) \\ &= \left(\frac{m}{k}\right) k\pi(1-\pi)\sigma^2 \\ &= m\pi(1-\pi)\sigma^2.\end{aligned}$$

- Dispersion parameter $\sigma^2 = \big((k-1)\rho + 1\big)$ also accounts for overdispersion at the grand total level. It depends on:
  - Cluster size k, and the
  - Intraclass correlation coefficient ρ.

Methods for Adjusting for Overdispersion

1. Ad Hoc Method.
   - Try to use
   $$\hat{\sigma}^2 = \hat{\phi} = \frac{D}{n-p},$$
   when $\mathsf{Var}(Y_{\bullet\bullet}) \gg m\pi(1-\pi)$.
   - Recall $\sigma^2 = \big((k-1)\rho + 1\big)$.
   - Scaling the variances by $\hat{\sigma}^2$ will be inefficient when clusters are of unequal size.
   - Instead, we prefer to use a mixture/random effects model to account for variation within the clusters.
2. Binomial Mixture Model.
   - Recall we assumed $\mathbb{E}[\pi_i] = \pi$ and $\mathsf{Var}(\pi_i) = \rho\pi(1-\pi)$.
   - Note $0 < \pi_i < 1$ which restricts our choice of distributions.
   - Let $\pi_i \sim \text{Beta}(\gamma_1, \gamma_2)$, where $\gamma_1, \gamma_2 > 0$ with pdf
   $$g(\pi_i; \gamma_1, \gamma_2) = \frac{\Gamma(\gamma_1+\gamma_2)}{\Gamma(\gamma_1)\Gamma(\gamma_2)} \pi_i^{\gamma_1 - 1}(1-\pi_i)^{\gamma_2 - 1}.$$
   - Mean and Variance of the Beta are derived in the course notes:
   $$\mathbb{E}[\pi_i] = \frac{\gamma_1}{\gamma_1+\gamma_2}, \qquad \mathsf{Var}(\pi_i) = \frac{\gamma_1\gamma_2}{(\gamma_1+\gamma_2)^2(1+\gamma_1+\gamma_2)}.$$
   - So we select (γ_1, γ_2) such that:
   $$\pi = \frac{\gamma_1}{\gamma_1+\gamma_2}, \qquad \rho = \frac{1}{1+\gamma_1+\gamma_2}.$$
   - Derive the marginal distribution of the cluster counts $Y_{i\bullet}$.
   - Relax the assumption that all clusters are equal sized, that is, let $Y_{i\bullet} \mid \pi_i \sim \text{BIN}(k_i, \pi_i)$.
   - Beta function: $B(\gamma_1, \gamma_2) = \frac{\Gamma(\gamma_1)\Gamma(\gamma_2)}{\Gamma(\gamma_1+\gamma_2)}$.
   $$\begin{aligned}
   \mathbb{P}(Y_{i\bullet} = y_{i\bullet}) &= \int_0^1 \mathbb{P}(Y_{i\bullet} = y_{i\bullet} \mid \pi_i) g(\pi_i; \gamma_1, \gamma_2)\, \mathrm{d}\pi_i \\
   &= \int_0^1 \binom{k_i}{y_{i\bullet}} \pi_i^{y_{i\bullet}} (1-\pi_i)^{k_i - y_{i\bullet}} \frac{1}{B(\gamma_1,\gamma_2)} \pi_i^{\gamma_1-1}(1-\pi_i)^{\gamma_2-1}\, \mathrm{d}\pi_i \\
   &= \binom{k_i}{y_{i\bullet}} \frac{1}{B(\gamma_1,\gamma_2)} \int_0^1 \pi_i^{y_{i\bullet}+\gamma_1-1}(1-\pi_i)^{k_i - y_{i\bullet}+\gamma_2-1}\, \mathrm{d}\pi_i \\
   &= \binom{k_i}{y_{i\bullet}} \frac{1}{B(\gamma_1,\gamma_2)} B(y_{i\bullet}+\gamma_1, k_i - y_{i\bullet} + \gamma_2).
   \end{aligned}$$
   - This is called the Beta-Binomial Distribution.

- It can be shown that for the Beta-Binomial:

$$\begin{aligned}\mathbb{E}[Y_{i\bullet}] &= k_i\left(\frac{\gamma_1}{\gamma_1+\gamma_2}\right)\\ &= k_i\pi,\end{aligned}$$

$$\begin{aligned}\mathsf{Var}(Y_{i\bullet}) &= k_i\left(\frac{\gamma_1}{\gamma_1+\gamma_2}\right)\left(\frac{\gamma_2}{\gamma_1+\gamma_2}\right)\left(\frac{k_i+\gamma_1+\gamma_2}{1+\gamma_1+\gamma_2}\right)\\ &= k_i\pi(1-\pi)\big(1+(k_i-1)\rho\big),\end{aligned}$$

  where

$$\pi = \frac{\gamma_1}{\gamma_1+\gamma_2}, \qquad \rho = \frac{1}{1+\gamma_1+\gamma_2}.$$

- See notes for various derivations including

$$\mathsf{Corr}(Y_{ij}, Y_{ik}) = \rho.$$

- Reduces to Binomial variance $\mathsf{Var}(Y_{i\bullet}) = k_i\pi(1-\pi)$ when
  - $\rho = 0$ (no correlation within observations from the same cluster), or
  - $k_i = 1$ (clusters of size 1).
- Could test H_0: $\rho = 0$ to test for overdispersion.
- The Binomial is nested within the Beta-Binomial so could do so using a Deviance/LR Test.
- It can be difficult to get the MLE's from the Beta-Binomial.
  - The gamma function $\Gamma(\,\cdot\,)$ is non-linear.
  - R: `glm.binom.disp()` function in `library(dispmod)`.
  - Iterative algorithm for estimating ρ and β.

Application — Pacific Cod Hatching Data

Hatching Data for Pacific Cod Eggs

- To learn about the importance of salinity, temperature, and oxygen concentration on the probability of hatching for eggs from Pacific cod fish, the following experiment was conducted.
  - Salinity (measured in *ppt*), temperature (measured in Celsius), and oxygen concentration (measured in *ppm*), were varied over ranges of practical relevance.
  - A known number of eggs were then placed in each of four tanks controlled at each specified settings for these factors.
  - The eggs were then observed to either hatch, or not hatch.
  - The total number of eggs hatching for each tank under each set of conditions was recorded.
- This gave four binomial samples for each configuration.
- The eggs in the same tank can not be considered independent.
- See Problem 4.1 of course notes.

R Code

```
cod <- read.table("cod.dat", header = T)
attach(cod)
# Fit a logistic regression model with all 2-way
# interactions
binom1 <- glm(cbind(hatch, total - hatch) ~ salin * temp + temp *
  O2 + salin * O2, family = binomial)
summary(binom1)
# Fit a beta binomial model to account for overdispersion
library(dispmod)
betabinom1 <- glm.binomial.disp(binom1)
summary(betabinom1)
1 - pchisq(binom1$deviance - betabinom1$deviance, 1)
binom2 <- glm(cbind(hatch, total - hatch) ~ temp * O2 + salin *
  O2, family = binomial)
betabinom2 <- glm.binomial.disp(binom2)
summary(betabinom2)
betabinom2$dispersion
# Constructing deviance residual plots
par(mfrow = c(1, 3))
fv1 <- binom1$fitted.values
rd1 <- residuals.glm(binom1, "deviance")
fv2 <- betabinom1$fitted.values
rd2 <- residuals.glm(betabinom1, "deviance")
fv3 <- betabinom2$fitted.values
rd3 <- residuals.glm(betabinom2, "deviance")
plot(fv1, rd1, xlab = "Fitted Values", ylab = "Deviance Residuals",
  main = "Binomial Model", ylim = c(-18, 15))
abline(h = -2)
abline(h = 2)
plot(fv2, rd2, xlab = "Fitted Values", ylab = "Deviance Residuals",
  main = "Beta-Binomial Model", ylim = c(-3, 3))
abline(h = -2)
abline(h = 2)
plot(fv3, rd3, xlab = "Fitted Values", ylab = "Deviance Residuals",
  main = "Beta-Binomial2 Model", ylim = c(-3, 3))
abline(h = -2)
abline(h = 2)
```

R Output — Dataset

```
print(cod[1:28, ], row.names = F)    print(cod[29:56, ], row.names = F)

 salin temp  O2 hatch total           salin temp   O2 hatch total
    14  2.7 3.6   224   283           26.00  9.3 8.60    74   293
    14  2.7 3.6   160   235           26.00  9.3 8.60    68   181
    14  2.7 3.6   180   245           26.00  9.3 8.60   152   307
    14  2.7 3.6   182   320           26.00  9.3 8.60    45   167
    14  2.7 8.6   231   325           20.00  6.0 6.10   221   259
    14  2.7 8.6   171   207           20.00  6.0 6.10   238   277
    14  2.7 8.6   237   283           20.00  6.0 6.10   224   296
    14  2.7 8.6   178   270           20.00  6.0 6.10   281   333
    14  9.3 3.6   159   240           12.71  6.0 6.10   222   268
    14  9.3 3.6   234   349           12.71  6.0 6.10   197   289
    14  9.3 3.6   163   229           12.71  6.0 6.10   279   341
    14  9.3 3.6   295   385           12.71  6.0 6.10   294   350
    14  9.3 8.6   186   314           27.29  6.0 6.10    46   230
    14  9.3 8.6    97   298           27.29  6.0 6.10   243   370
    14  9.3 8.6   214   297           27.29  6.0 6.10    62   214
    14  9.3 8.6    74   244           27.29  6.0 6.10   138   265
    26  2.7 3.6     5   217           20.00  2.0 6.10    20   230
    26  2.7 3.6     2   243           20.00  2.0 6.10    11   175
    26  2.7 3.6     5   316           20.00  2.0 6.10    10   233
    26  2.7 3.6     3   224           20.00  2.0 6.10     7   236
    26  2.7 8.6   143   292           20.00 10.0 6.10   130   389
    26  2.7 8.6   159   301           20.00 10.0 6.10   119   226
    26  2.7 8.6   186   316           20.00 10.0 6.10    98   247
    26  2.7 8.6   138   264           20.00 10.0 6.10   122   292
    26  9.3 3.6    19   262           20.00  6.0 3.08   187   293
    26  9.3 3.6    36   277           20.00  6.0 3.08   168   258
    26  9.3 3.6    18   263           20.00  6.0 3.08   214   271
    26  9.3 3.6    44   290           20.00  6.0 3.08   179   220
```

R Output — Logistic Regression Model

```
Call:
glm(formula = cbind(hatch, total - hatch) ~ salin * temp + temp *
    O2 + salin * O2, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-17.929   -5.575   -1.081    5.144   14.775

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.339730   0.279098  19.132  < 2e-16 ***
salin       -0.388016   0.013975 -27.764  < 2e-16 ***
temp         0.083883   0.030418   2.758 0.005821 **
O2          -0.127545   0.035243  -3.619 0.000296 ***
salin:temp   0.009308   0.001277   7.288 3.15e-13 ***
temp:O2     -0.043843   0.003054 -14.355  < 2e-16 ***
```

```
salin:O2     0.028164   0.001603  17.571  < 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: 8209.9  on 75  degrees of freedom
Residual deviance: 4308.2  on 69  degrees of freedom
AIC: 4746

Number of Fisher Scoring iterations: 5
```

Logistic Regression Model

- All 2-way interaction terms are statistically significant (before ad hoc adjustment).
- The 3-way interaction term is not statistically significant (model not shown).
- Ad Hoc Method:

$$\hat{\phi} = \frac{D}{n-p} = \frac{4308.2}{69} = 62.44.$$

- Examine the significance the `salin:temp` interaction term:

$$\text{se}_{\text{adj}}(\hat{\beta}_4) = \sqrt{\hat{\phi}}\,\text{se}(\hat{\beta}_4) = \sqrt{62.44}(0.001277) = 0.01009072.$$

H_0: $\beta_4 = 0$ versus H_A: $\beta_4 \neq 0$:

$$p = 2\,\mathbb{P}\left(Z > \frac{|\hat{\beta}_4 - 0|}{\text{se}_{\text{adj}}(\hat{\beta}_4)}\right) = 2\,\mathbb{P}(Z > 0.9224) = 0.356.$$

- The `salin:temp` interaction term is no longer statistically significant.

R Output — Beta Binomial Regression Model

```
Call:
glm(formula = cbind(hatch, total - hatch) ~ salin * temp + temp *
    O2 + salin * O2, family = binomial, weights = disp.weights)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.40644  -0.68571   0.01133   0.64835   1.83682

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.645924   2.050995   2.753  0.00591 **
salin       -0.421677   0.105007  -4.016 5.93e-05 ***
temp         0.152677   0.228123   0.669  0.50332
O2          -0.223837   0.264152  -0.847  0.39678
salin:temp   0.008644   0.009475   0.912  0.36162
temp:O2     -0.049173   0.022930  -2.144  0.03200 *
salin:O2     0.034040   0.012323   2.762  0.00574 **
```

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

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 141.76  on 75  degrees of freedom
Residual deviance:  75.03  on 69  degrees of freedom
AIC: 96.17

Number of Fisher Scoring iterations: 5
```

Beta Binomial Regression Model

- The `salin:temp` interaction term is not statistically significant.
- Note β_j's from this model still have $\log(\text{OR})$ interpretations.
- The dispersion parameter is $\hat{\rho} = 0.198$ (correlation coefficient).
- Test for overdispersion using a Deviance Test: H_0: $\rho = 0$ (Binomial) versus H_A: $\rho \neq 0$ (Beta-Binomial):

$$\Delta D = D_0 - D_A = 4302.8 - 75.03 = 4227.77.$$

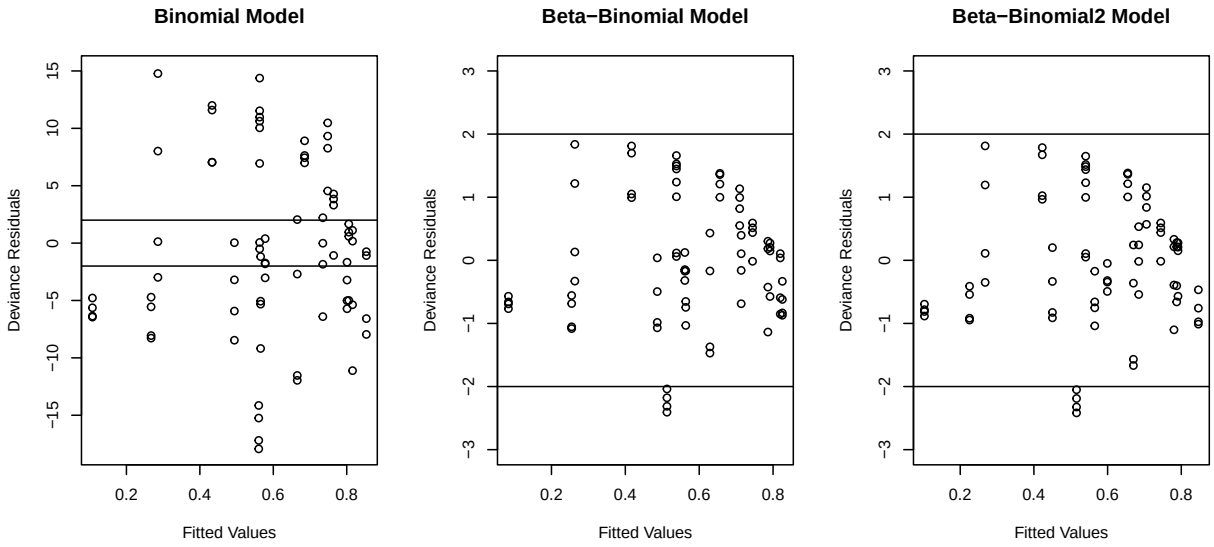
$$p = \mathbb{P}(\chi^2_1 > 4227.77) < 0.001.$$

  Therefore, we reject the null hypothesis of no overdispersion.

R Output — Beta Binomial 2 Regression Model

```
Call:
glm(formula = cbind(hatch, total - hatch) ~ temp * O2 + salin *
    O2, family = binomial, weights = disp.weights)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.41745  -0.70916  -0.01631   0.65327   1.81255

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.69472    1.74939   2.684  0.00728 **
temp         0.30839    0.15378   2.005  0.04492 *
O2          -0.24170    0.26184  -0.923  0.35595
salin       -0.36705    0.08472  -4.333 1.47e-05 ***
temp:O2     -0.04647    0.02272  -2.046  0.04081 *
O2:salin     0.03388    0.01231   2.752  0.00593 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 141.681  on 75  degrees of freedom
Residual deviance:  75.824  on 70  degrees of freedom
AIC: 94.961

Number of Fisher Scoring iterations: 5
```

Residual Plots for Binomial and Beta-Binomial Models



Binomial Overdispersion Wrap-Up

- We would select `betabinom2` model as our final model.
- Interpretation of β's is as in logistic regression ($\log(\text{OR})$).
- Here, we had no problems fitting the Beta Binomial models.
- May not always be the case.
- Chapter 5: introduction to Quasi Likelihood (not covered).
  - Relaxes parametric assumptions (Binomial, Poisson, Beta Binomial, Negative Binomial, Exponential, Gamma, etc).
  - Can be used in settings with overdispersion.