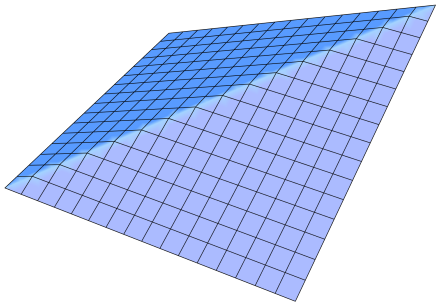

STAT 431: Generalized Linear Models

Prof. Cecilia Cotton • Winter 2016 • University of Waterloo
TR 11:30AM–12:50PM in STC 0060

Transcribed, LaTeXed, and edited by Rob Zimmerman

Note: These are personal lecture notes, not official course materials. They may contain errors (which by default you should attribute to me, Rob Zimmerman, rather than the course instructor). The permanent home of these — and all of my other lecture notes — is https://rob-zimmerman.github.io/coursenotes.

Contents

Lecture 1 — Tuesday, January 05, 2016

1. Generalized Linear Model Foundations and Exponential Families

Review of Linear Regression

The standard model fitting process begins with exploratory data analysis, then specifies a probability model for the response and a regression equation linking the response to the explanatory variables. After estimating the model parameters, we check model fit and then use the fitted model for inference.

In multiple linear regression, for each subject $i = 1, \ldots, n$, let Y_i be the response and let

$$\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$$

be the vector of explanatory variables. The usual model assumes independent Gaussian responses

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \qquad \text{and} \qquad \mathbb{E}[Y_i] = \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

Equivalently,

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

In matrix form this is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{and} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

Here

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

Before fitting the model, it is useful to inspect the data graphically. Figure 1 shows the Dobson data: birthweight plotted against gestational age, with points distinguished by sex and with

separate smooth curves added in the second plot. The plots suggest a positive relationship between gestational age and birthweight, along with a possible sex difference.

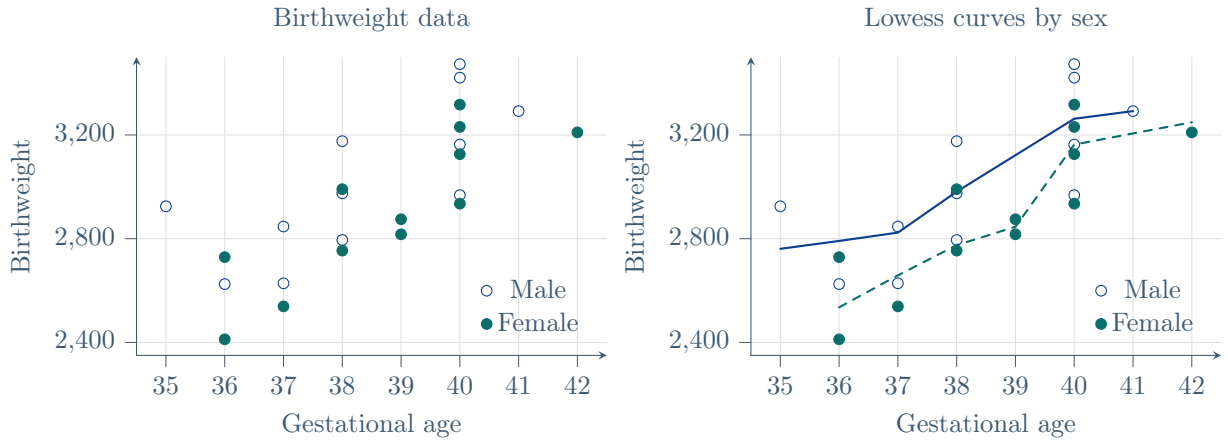

FIGURE 1

The **least squares criterion** is

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2 .$$

The **least squares estimates** are obtained by setting the partial derivatives of $S(\boldsymbol{\beta})$ equal to zero. These derivatives are

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right),$$
$$\frac{\partial S}{\partial \beta_j} = -2 \sum_{i=1}^{n} x_{ij} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right), \quad j = 1, \ldots, p.$$

The same estimate is obtained from maximum likelihood when σ^2 is treated as fixed. The density of Y_i is

$$f(y_i; \mu_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\right].$$

Therefore, the log-likelihood for $\boldsymbol{\beta}$ is

$$\ell(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n} \left[-\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(y_i - \mu_i)^2\right]$$

$$= \sum_{i=1}^{n} \left[-\frac{1}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\right)^2 \right].$$

Maximizing this log-likelihood in $\boldsymbol{\beta}$ is equivalent to minimizing the sum of squared residuals, since the only term depending on $\boldsymbol{\beta}$ is the negative multiple of that sum. When $\boldsymbol{X}$ has full column rank, the least squares estimate is

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

For the normal linear model, this is also the **maximum likelihood estimate** of $\boldsymbol{\beta}$ when σ^2 is treated as fixed. The **fitted values** and **residuals** are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \qquad \text{and} \qquad \hat{r}_i = y_i - \hat{y}_i.$$

If $n > p + 1$, an unbiased estimate of the error variance is

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

and

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

For the Dobson birthweight data below, Y_i is the birthweight in grams of baby i, x_{i1} is the sex indicator with $x_{i1} = 0$ for female and $x_{i1} = 1$ for male, and x_{i2} is gestational age in weeks. The multiple linear regression model is

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

The data and fitted model can be reproduced in R as follows:

```
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)))

plot(age, birthw, pch = 1 + 18 * as.numeric(sex == "F"),
     ylab = "Birthweight", main = "Dobson's Birth Weight 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))

m1 <- lm(birthw ~ sex + age)
summary(m1)

plot(age, birthw, pch = 1 + 18 * as.numeric(sex == "F"),
     ylab = "Birth weight", main = "Fitted Regression Lines (m1)")
abline(m1$coeff[1] + m1$coeff[2], m1$coeff[3])
abline(m1$coeff[1], m1$coeff[3], lty = 2)
```

The fitted coefficient estimates are as follows:

Coefficient	Estimate	Standard error	p-value
β_0	-1773.32	794.59	0.0367
β_1 (male)	163.04	72.81	0.0361
β_2 (gestational age)	120.89	20.46	7.28×10^{-6}

The residual standard error is 177.1 on 21 degrees of freedom, the multiple R^2 value is 0.64, the adjusted R^2 value is 0.6057, and the overall F-statistic is 18.67 on 2 and 21 degrees of freedom, with p-value 2.194×10^{-5}.

In this model, β_0 is the expected birthweight of a female baby at gestational age zero. This intercept is part of the fitted line, but it has no practical interpretation because gestational age zero is outside the range of the data. The coefficient β_1 is the expected change in birthweight for male babies compared with female babies at a fixed gestational age, and β_2 is the expected change in birthweight associated with a one-week increase in gestational age after adjusting for sex. After adjusting for gestational age, male babies are estimated to be 163.04 grams heavier than female babies. Using $t_{21,0.975} = 2.079614$, a 95% confidence interval for this estimate is

$$163.04 \pm 2.079614(72.81) = (11.62, 314.46).$$

The corresponding two-sided p-value is

$$2\left(1 - \mathbb{P}(T_{21} \leq 2.239)\right) = 0.0361,$$

so we reject the null hypothesis that $\beta_1 = 0$.

Figure 2 shows the fitted regression lines for the main effects model. The two fitted lines have the same slope in gestational age and differ only by the sex coefficient.

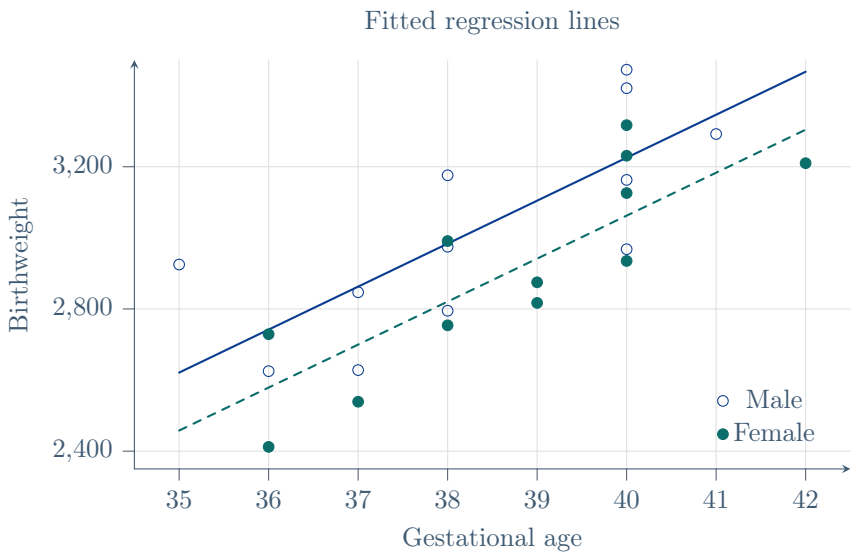

FIGURE 2

Model checking for a linear regression fit is usually based on **standardized residual plots**. If the model provides a good fit to the data, then the standardized residuals

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

should look approximately like independent $\mathcal{N}(0,1)$ random variables. Here h_{ii} is the (i,i) entry of the **hat matrix**

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top.$$

Common diagnostic plots include standardized residuals against fitted values, standardized residuals against explanatory variables, normal quantile plots of the standardized residuals, and added-variable plots. Figure 3 shows two such diagnostics for the birthweight model.

Under the usual Gaussian linear model assumptions, the fitted regression parameters are exactly normally distributed:

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\right).$$

If v_{jj} is the jth diagonal entry of $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, then

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

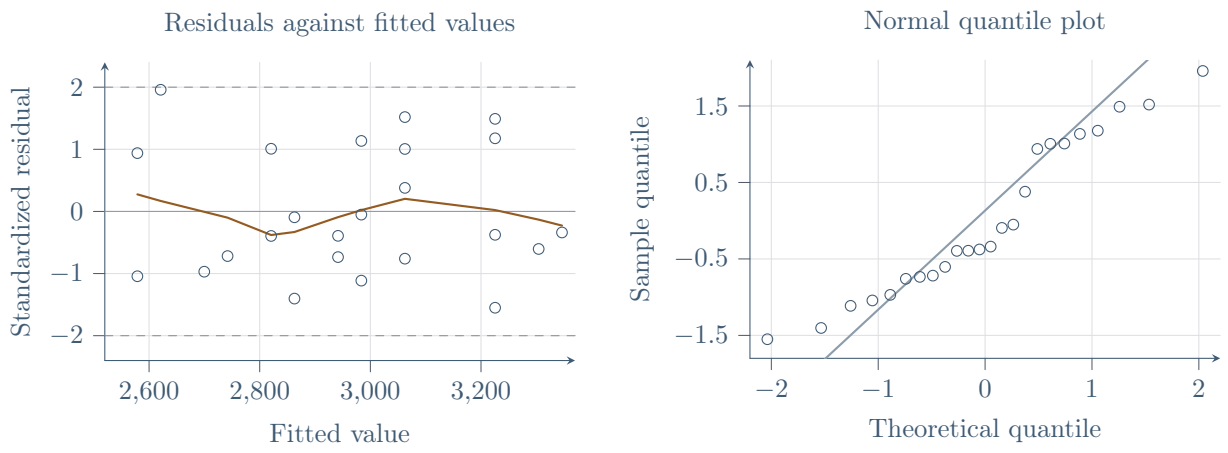

FIGURE 3

If σ^2 were known, a 95% confidence interval for β_j would be

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

Since σ^2 is usually unknown, we replace it with $\hat{\sigma}^2$ and use the estimated standard error

$$\widehat{\text{standard error}}(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 v_{jj}}.$$

A $100(1-\alpha)\%$ confidence interval is therefore

$$\hat{\beta}_j \pm t_{n-p-1,1-\alpha/2}\widehat{\text{standard error}}(\hat{\beta}_j).$$

To test

$$H_0 : \beta_j = \beta_j^* \qquad \text{against} \qquad H_a : \beta_j \neq \beta_j^*,$$

we use the test statistic

$$T = \frac{\hat{\beta}_j - \beta_j^*}{\widehat{\text{standard error}}(\hat{\beta}_j)} = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\hat{\sigma}^2 v_{jj}}}.$$

Under H_0, this statistic has a t_{n-p-1} distribution, so the two-sided level-α test rejects H_0 when

$$|T| > t_{n-p-1,1-\alpha/2}.$$

Linear regression is useful, but it is not appropriate when the response cannot reasonably be treated as normally distributed, such as with binary data or count data. It also does not naturally handle settings where the variance of Y depends on its mean. **Generalized linear models** extend

the linear regression framework to handle these cases, with ordinary Gaussian linear regression as a special case.

The birthweight example also motivates **interaction terms**. To ask whether the rate of increase of birthweight with gestational age is the same for boys and girls, consider

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}.$$

Fitting this interaction model gives the following coefficient estimates.

Coefficient	Estimate	Standard error	p-value
β_0	-2141.67	1163.60	0.080574
β_1 (male)	872.99	1611.33	0.593952
β_2 (gestational age)	130.40	30.00	0.000313
β_3 (male by gestational age)	-18.42	41.76	0.663893

When x_{i1} is binary, the coefficient β_1 changes the intercept between the two groups, while β_3 changes the slope in x_{i2}. Figure 4 shows the case in which $\beta_1 > 0$ and $\beta_3 < 0$.

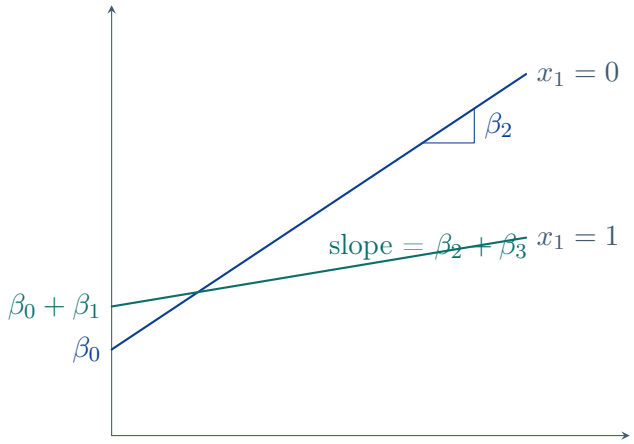

FIGURE 4

For a fixed value of x_1, increasing x_2 by one unit changes the mean response by the corresponding slope. Thus, when $x_1 = 0$, the change is β_2, while when $x_1 = 1$, the change is $\beta_2 + \beta_3$.

(F)

x_1	x_2	$\mathbb{E}[Y_i]$
0	x_2+1	$\beta_0+\beta_2(x_2+1)$
0	x_2	$\beta_0+\beta_2 x_2$
		β_2
1	x_2+1	$\beta_0+\beta_1+\beta_2(x_2+1)+\beta_3(x_2+1)$
1	x_2	$\beta_0+\beta_1+\beta_2 x_2+\beta_3 x_2$
		$\beta_2+\beta_3$

Lecture 2 — Thursday, January 07, 2016

Likelihood Methods for Scalar Parameters

Let $Y_1,\ldots,Y_n$ be independent and identically distributed with density or mass function $f(y_i;\theta)$. The **likelihood function** is

$$L(\theta;\boldsymbol{y}) = c\prod_{i=1}^{n} f(y_i;\theta),$$

where c is any positive constant that does not depend on θ. The maximum likelihood estimate is

$$\hat{\theta} = \underset{\theta}{argmax}\, L(\theta;\boldsymbol{y}).$$

Equivalently, since the logarithm is monotone, we may maximize the **log-likelihood**

$$\ell(\theta;\boldsymbol{y}) = \log L(\theta;\boldsymbol{y}) = \log c + \sum_{i=1}^{n}\log f(y_i;\theta).$$

For example, suppose that $Y_1,\ldots,Y_n$ are independent and identically distributed Poisson random variables with unknown parameter θ. Then

$$f(y_i;\theta) = \frac{\theta^{y_i}e^{-\theta}}{y_i!}, \quad y_i = 0,1,2,\ldots.$$

The likelihood is

$$L(\theta) = \prod_{i=1}^{n}\frac{\theta^{y_i}e^{-\theta}}{y_i!} = \frac{\theta^{\sum_{i=1}^n y_i}e^{-n\theta}}{\prod_{i=1}^n y_i!} \propto \theta^{\sum_{i=1}^n y_i}e^{-n\theta}.$$

Consequently,

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

The **score function** is

$$s(\theta) = \frac{\mathrm{d}\ell}{\mathrm{d}\theta} = \frac{\sum_{i=1}^{n} y_i}{\theta} - n.$$

Solving $s(\theta) = 0$ gives

$$\hat{\theta} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}.$$

The **observed information** is

$$I(\theta) = -\frac{\mathrm{d}^2\ell}{\mathrm{d}\theta^2} = -s'(\theta) = \frac{\sum_{i=1}^{n} y_i}{\theta^2}.$$

If $\sum_i y_i > 0$, this is positive for every $\theta > 0$, so the critical point is the maximum likelihood estimate. If every count is zero, the likelihood is instead maximized at the boundary value $\hat{\theta} = 0$. When $\sum_i y_i > 0$, the **log relative likelihood** is

$$r(\theta) = \ell(\theta) - \ell(\hat{\theta}) = \left(\sum_{i=1}^{n} y_i\right) \log\left(\theta/\hat{\theta}\right) - n(\theta - \hat{\theta}).$$

At $\theta = \hat{\theta}$, this gives

$$r(\hat{\theta}) = \left(\sum_{i=1}^{n} y_i\right) \log 1 - n(0) = 0.$$

For the tropical cyclone data, the response is the number of tropical cyclones in northeastern Australia during each of the 13 successive seasons from 1956–1957 through 1968–1969. The observed counts are

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

If Y_i is the number of cyclones in season i, then a Poisson(θ) model gives

$$\hat{\theta} = \bar{y} = \frac{72}{13} = 5.538.$$

Figure 5 shows the log relative likelihood for this Poisson model. Its maximum occurs at $\hat{\theta} = \bar{y} = 5.538$, and the curve is used below for likelihood-based confidence intervals.

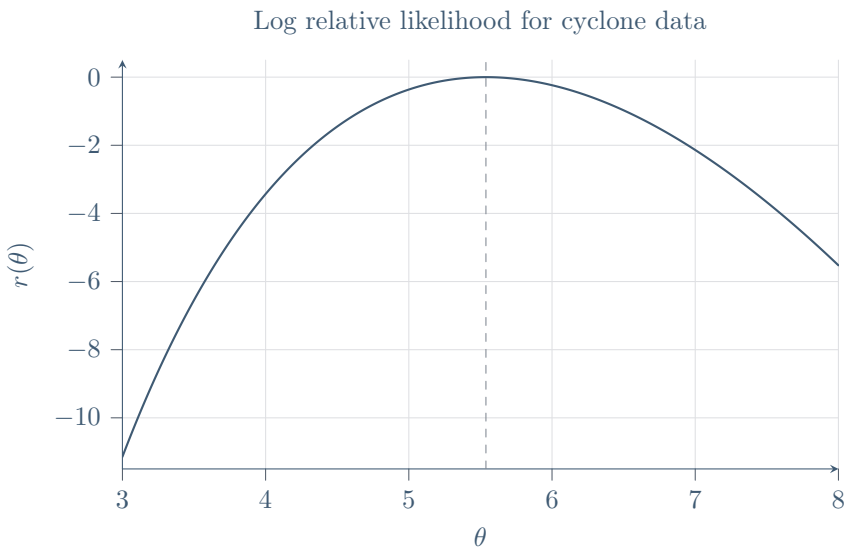

FIGURE 5

Newton–Raphson and Wald Inference

To solve $s(\theta) = 0$ iteratively, Taylor expand the score function about a current estimate θ_0:

$$\begin{aligned} s(\theta) &= s(\theta_0) + s'(\theta_0)(\theta - \theta_0) + \dots \\ s(\theta) &\simeq s(\theta_0) - I(\theta_0)(\theta - \theta_0) \end{aligned}$$

Let $\hat{\theta}$ be the maximum likelihood estimate, so $s(\hat{\theta}) = 0$.

$$\begin{aligned} s(\hat{\theta}) &\simeq s(\theta_0) - I(\theta_0)(\hat{\theta} - \theta_0) \\ I(\theta_0)(\hat{\theta} - \theta_0) &\simeq s(\theta_0) \\ (\hat{\theta} - \theta_0) &\simeq I^{-1}(\theta_0)s(\theta_0) \\ \hat{\theta} &\simeq \theta_0 + I^{-1}(\theta_0)s(\theta_0) \end{aligned}$$

The Newton–Raphson update is therefore defined by

$$\theta_1 = \theta_0 + I^{-1}(\theta_0)s(\theta_0),$$

where θ_1 is the next estimate and θ_0 is the previous estimate.

Example 2.1 For the tropical cyclone data, take $\theta_0 = 5$. Then

$$s(\theta_0) = \frac{72}{5} - 13 = 1.4$$
$$I(\theta_0) = \frac{72}{25} = 2.88$$
$$\theta_1 = \theta_0 + I^{-1}(\theta_0)s(\theta_0) = 5 + \frac{1.4}{2.88} = 5.486111$$

Recall that if $Z \sim \mathcal{N}(0,1)$, then $Z^2 \sim \chi^2_1$. The approximate **Wald confidence set** for θ is

$$\left\{\theta : (\hat{\theta} - \theta)^2 I(\hat{\theta}) \leq \chi^2_1(1-\alpha)\right\}.$$

Figure 6 shows the relationship between the one-sided chi-square cutoff and the corresponding two-sided normal cutoff.

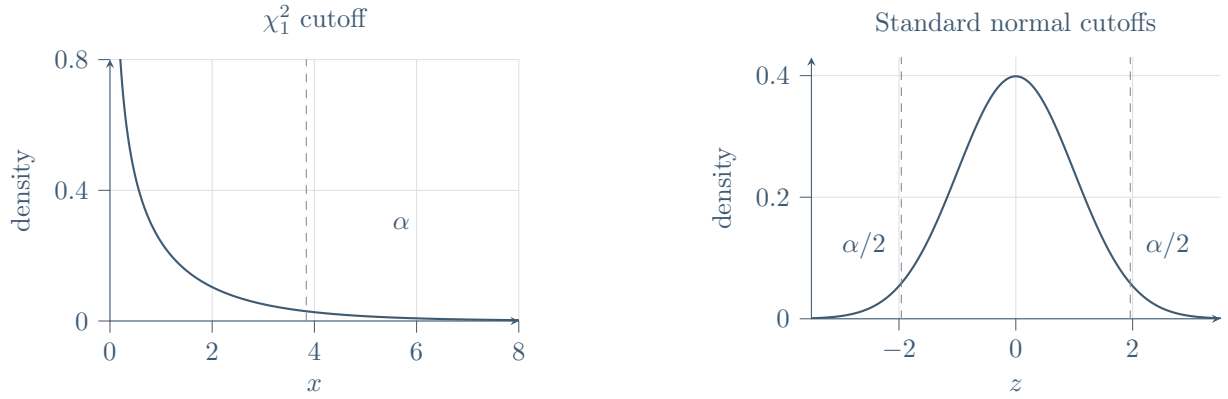

FIGURE 6

$$\left|(\hat{\theta} - \theta)I(\hat{\theta})^{1/2}\right| < z_{1-\alpha/2}$$
$$\left|\hat{\theta} - \theta\right| < z_{1-\alpha/2} I(\hat{\theta})^{-1/2}$$
$$\text{confidence interval:} \quad \hat{\theta} \pm z_{1-\alpha/2} I(\hat{\theta})^{-1/2}.$$

Figure 7 represents the absolute distance $\left|\hat{\theta} - \theta\right|$ that appears in the Wald confidence interval.

Here

$$L_W = \hat{\theta} - z_{1-\alpha/2} I(\hat{\theta})^{-1/2} \qquad \text{and} \qquad U_W = \hat{\theta} + z_{1-\alpha/2} I(\hat{\theta})^{-1/2}.$$

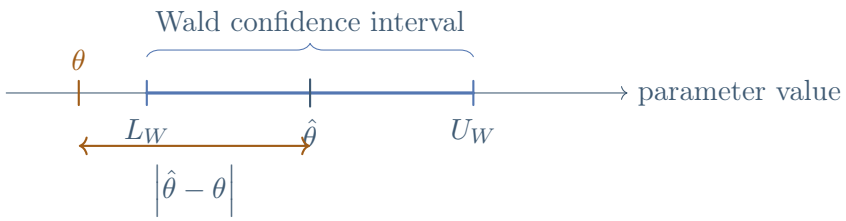

FIGURE 7

The **likelihood interval** and the **Wald interval** can be compared on the log relative likelihood scale, as shown in Figure 8.

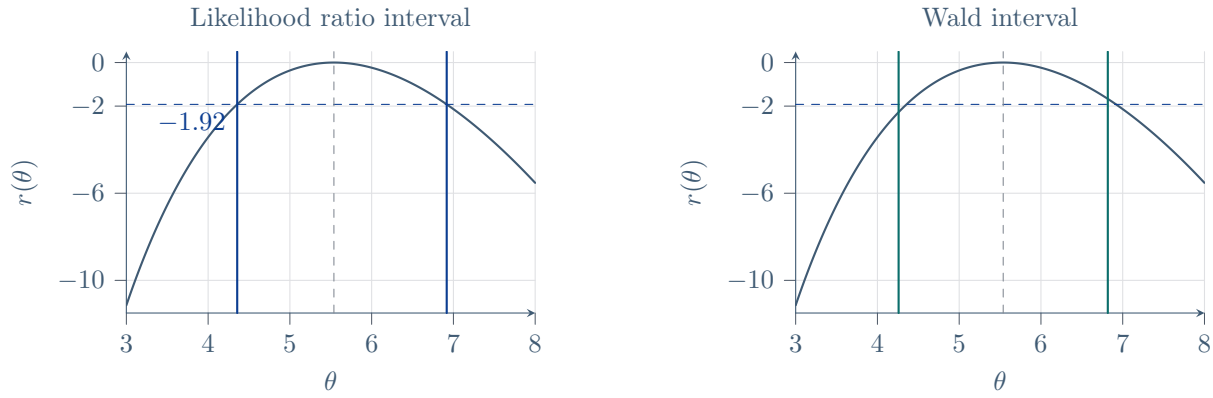

FIGURE 8

For the tropical cyclone data, the Wald interval is $(4.26, 6.82)$. The corresponding likelihood interval is read from the log relative likelihood cutoff

$$-\frac{1}{2}\chi^2_1(0.95) = -1.92 \qquad \text{and} \qquad \chi^2_1(0.95) = 3.84,$$

which gives the approximate interval $(4.36, 6.92)$.

The same likelihood and Wald methods apply to binomial examples, where the scalar parameter is a success probability rather than a Poisson mean.

Likelihood Methods for Vector Parameters

For a vector parameter, the likelihood ideas are the same, but the score becomes a vector and the information becomes a matrix. In linear regression, for instance,

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

where μ_i is the scalar mean and $\boldsymbol{\beta}$ is the parameter vector. The estimate $\hat{\boldsymbol{\beta}}$ is therefore the maximum likelihood estimate of a vector.

For a two-dimensional parameter $\boldsymbol{\theta} = (\alpha, \beta)$, such as a gamma distribution parameterized by α and β, the score vector is

$$\boldsymbol{s}(\alpha, \beta) = \begin{bmatrix} \dfrac{\partial \ell}{\partial \alpha} \\ \dfrac{\partial \ell}{\partial \beta} \end{bmatrix}_{2\times 1}.$$

The maximum likelihood estimate solves

$$\boldsymbol{s}(\alpha, \beta) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The corresponding information matrix is

$$\boldsymbol{I}(\alpha, \beta) = \begin{bmatrix} -\dfrac{\partial^2 \ell}{\partial \alpha^2} & -\dfrac{\partial^2 \ell}{\partial \alpha\, \partial \beta} \\ -\dfrac{\partial^2 \ell}{\partial \alpha\, \partial \beta} & -\dfrac{\partial^2 \ell}{\partial \beta^2} \end{bmatrix}.$$

Figure 9 depicts the same cutoff idea applied through a **profile likelihood** curve for one component of a vector parameter.

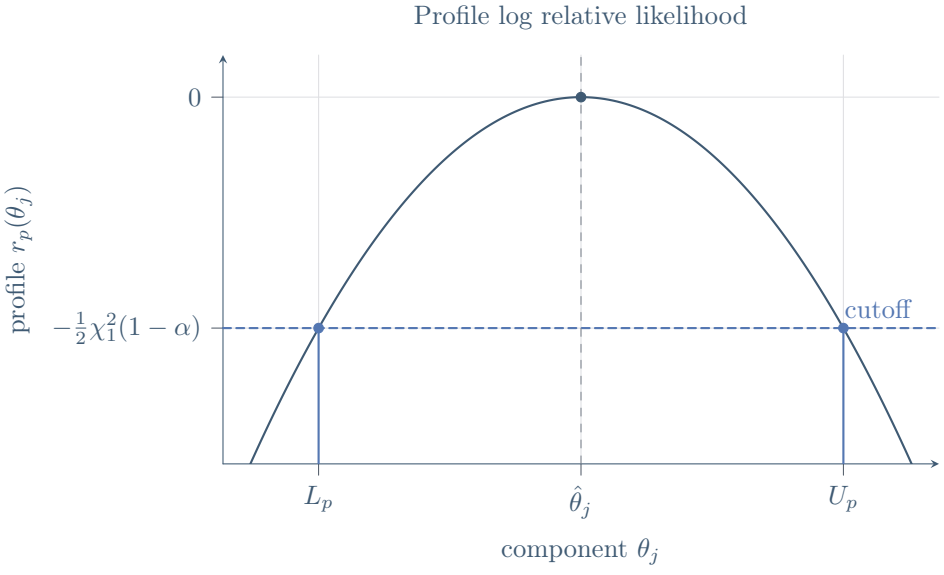

FIGURE 9

Lecture 3 — Tuesday, January 12, 2016

Exponential Families

In ordinary linear regression, we usually assume that the observations $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$ are independent and model them as

$$Y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \qquad \text{so that} \qquad \mathbb{E}[Y_i] = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

This motivates the generalized linear model framework, which begins with the same regression idea but allows the response distribution to be nonnormal.

Definition 3.1 A random variable Y, with density or mass function $f(y; \theta, \phi)$, is said to belong to the **exponential family** if

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

where a, b, and c are specified functions. The parameter θ is the **canonical parameter**, and ϕ is the **scale** or **dispersion parameter**; depending on the model, ϕ may be known or treated as a nuisance parameter and estimated separately.

Example 3.2 For $Y \sim \text{Poisson}(\lambda)$,

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

Thus,

$$\theta = \log \lambda \qquad b(\theta) = \lambda = \exp(\theta)$$
$$\phi = 1 \qquad a(\phi) = 1$$
$$c(y) = -\log y!$$

So the Poisson distribution is a member of the exponential family.

Under the usual regularity conditions, including permission to differentiate under the integral or

sum, we will use the following likelihood identities:

1. For a density or mass function f,

$$\mathbb{E}\left[g(X)\right] = \int g(x)f(x)\,\mathrm{d}x \qquad \text{or} \qquad \mathbb{E}\left[g(X)\right] = \sum g(x)f(x),$$

   respectively.

2. The score has mean zero:

$$\mathbb{E}\left[s(\theta)\right] = 0.$$

3. The second moment of the score equals the expected information:

$$\mathbb{E}\left[I(\theta)\right] = \mathbb{E}\left[s(\theta)^2\right] = \mathcal{I}(\theta).$$

4. Since $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, (2) and (3) imply

$$\mathrm{Var}\left(s(\theta)\right) = \mathbb{E}\left[I(\theta)\right].$$

Applying these identities to the exponential family gives the mean and variance in terms of b. First,

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

Thus,

$$\mathbb{E}\left[Y\right] = b'(\theta).$$

For the variance,

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

Thus,

$$\mathrm{Var}(Y) = a(\phi)b''(\theta).$$

Link Functions

The **link function** relates the mean $\mu = \mathbb{E}[Y]$ to the linear predictor $\eta = \boldsymbol{x}^\top \boldsymbol{\beta}$:

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

For linear regression, the **identity link** $g(\mu) = \mu$ gives

$$g(\mu) = \mathbb{E}[Y] = \eta = \boldsymbol{x}^\top \boldsymbol{\beta} \qquad \text{and} \qquad \mathbb{E}[Y] = \boldsymbol{x}^\top \boldsymbol{\beta}.$$

Another common link is the **log link**,

$$\log \mu = \boldsymbol{x}^\top \boldsymbol{\beta}.$$

When Y belongs to the exponential family, the **canonical link** is the link for which the canonical parameter equals the linear predictor:

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

Equivalently,

$$\mu = g^{-1}(\boldsymbol{x}^\top \boldsymbol{\beta}).$$

Example 3.3 For $Y \sim \text{Poisson}(\lambda)$, verify that the Poisson distribution belongs to the generalized linear model exponential family and identify the key components.

1. Exponential family template:

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

   Poisson probability mass function:

$$f(y; \lambda) = \exp\{y \log \lambda - \lambda - \log(y!)\}, \quad y = 0, 1, 2, \ldots$$

2. Parameter identification:

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

   By (1) and (2), the Poisson distribution is a member of the exponential family.

3. Mean, variance function, and variance:

$$\mu = \mathbb{E}[Y] = b'(\theta) = e^{\theta} \qquad \text{and} \qquad V(\mu) = b''(\theta) = e^{\theta} = \mu.$$

Therefore,

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

4. Canonical link: set $\eta = \theta = g(\mu)$. Since $\theta = \log \mu$, we get

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

5. Generalized linear model specification:

$$\log \mu_i = \eta_i = \boldsymbol{x}_i^{\top}\boldsymbol{\beta} \qquad \text{and} \qquad \mu_i = \exp\left(\boldsymbol{x}_i^{\top}\boldsymbol{\beta}\right).$$

Note that the exponential family form used here is a special case of the **general exponential family**

$$f(y;\alpha) = \exp\left(\sum_{i=1}^{k} w_i(\alpha)t_i(y) + B(\alpha) + h(y)\right).$$

Indeed, when ϕ is known,

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

This matches the general form with $k = 1$, $w_1(\alpha) = \theta/a(\phi)$, $t_1(y) = y$, $B(\alpha) = -b(\theta)/a(\phi)$, and $h(y) = c(y;\phi)$. Hence every distribution in the STAT 431 exponential family, with ϕ known, is also a member of the general exponential family.

For independent responses $Y_1, \ldots, Y_n$ with observation-specific canonical parameters $\theta_1, \ldots, \theta_n$, write $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^{\top}$. Then

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i;\theta_i,\phi) = \sum_{i=1}^{n} \ell_i(\theta_i).$$

The score components are

$$\frac{\partial \ell}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)}$$

for $i = 1, \ldots, n$. Because each likelihood contribution depends only on its own canonical parameter,

the observed information matrix in $\boldsymbol{\theta}$ is diagonal, with

$$I_{ii}(\boldsymbol{\theta}) = \frac{b''(\theta_i)}{a(\phi)} \qquad \text{and} \qquad I_{ij}(\boldsymbol{\theta}) = 0 \quad (i \neq j).$$

Lecture 4 — Thursday, January 14, 2016

Regression

A generalized linear model consists of the following three components:

1. The random component: the distribution of each independent response Y_i comes from a parametric family in the exponential family.
2. The systematic component: the linear predictor is $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$.
3. The link function: the mean $\mu_i = \mathbb{E}[Y_i]$ satisfies

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

Equivalently,

$$\mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \qquad \text{and} \qquad \mathbb{E}\left[Y_i\right] = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}).$$

Example 4.1 For the normal distribution with identity link, the generalized linear model has

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

Example 4.2 For the Poisson distribution with log link, the generalized linear model has

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

To estimate $\boldsymbol{\beta}$, we solve the score equation $\boldsymbol{S}(\boldsymbol{\beta}) = \mathbf{0}$ and then use the information matrix for inference on $\hat{\boldsymbol{\beta}}$. For a generalized linear model, the jth element of the score vector is

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \mu_i) w_i \frac{\partial \eta_i}{\partial \mu_i} x_{ij}, \text{ where } w_i^{-1} = \mathrm{Var}(Y_i) \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2$$

and where the derivative $\partial\eta_i/\partial\mu_i$ is evaluated at the covariate vector $\boldsymbol{x}_i$. In matrix form,

$$\underset{p\times 1}{\boldsymbol{S}(\boldsymbol{\beta})} = \underset{(p\times n)}{\boldsymbol{X}^\top}\underset{(n\times n)}{\boldsymbol{W}}\left[\underset{(n\times 1)}{(\boldsymbol{y}-\boldsymbol{\mu})} * \underset{(n\times 1)}{\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}}\right],$$

where $*$ denotes componentwise multiplication, and

$$\boldsymbol{S}(\boldsymbol{\beta}) = \begin{bmatrix} \dfrac{\partial\ell}{\partial\beta_0} \\ \vdots \\ \dfrac{\partial\ell}{\partial\beta_{p-1}} \end{bmatrix},$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix},$$

$$\boldsymbol{W} = \begin{bmatrix} w_1 & & 0 & \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_n \end{bmatrix},$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}.$$

Example 4.3 Consider the score vector for $\boldsymbol{\beta}$ in a Poisson generalized linear model with canonical link. We have $Y_i \sim \text{Poisson}(\lambda_i)$, $\lambda_i = e^{\theta_i} = \mu_i$, and $\eta_i = \boldsymbol{x}_i^\top\boldsymbol{\beta} = \log\mu_i$. Hence

$$\lambda_i = \exp\Big(\boldsymbol{x}_i^\top\boldsymbol{\beta}\Big).$$

1. Direct likelihood method. For one observation y,

$$\ell_i(\lambda) = y\log\lambda - \lambda - \log(y!).$$

Reparameterizing in terms of β gives

$$\ell_i(\boldsymbol{\beta}) = y\boldsymbol{x}^\top\boldsymbol{\beta} - \exp\Big(\boldsymbol{x}^\top\boldsymbol{\beta}\Big) - \log(y!).$$

Since

$$\boldsymbol{x}^\top\boldsymbol{\beta} = \sum_{j=0}^{p-1} x_j\beta_j \qquad \text{and} \qquad \frac{\partial}{\partial\beta_j}\boldsymbol{x}^\top\boldsymbol{\beta} = x_j,$$

the jth score component is

$$\begin{aligned} \frac{\partial\ell_i}{\partial\beta_j} &= yx_j - x_j\exp\Big(\boldsymbol{x}^\top\boldsymbol{\beta}\Big) \\ S_j(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\Big[y_ix_{ij} - x_{ij}\exp\Big(\boldsymbol{x}_i^\top\boldsymbol{\beta}\Big)\Big], \quad j = 0,\ldots,p-1. \end{aligned} \tag{4.1}$$

2. General exponential family method. The general generalized linear model score formula gives

$$\frac{\partial\ell_i}{\partial\beta_j} = (y_i - \mu_i)w_i\frac{\partial\eta_i}{\partial\mu_i}x_{ij} \qquad \text{and} \qquad w_i^{-1} = \mathrm{Var}(Y_i)\left(\frac{\partial\eta_i}{\partial\mu_i}\right)^2.$$

For the Poisson canonical link,

$$\mu_i = \exp\Big(\boldsymbol{x}_i^\top\boldsymbol{\beta}\Big), \qquad \eta_i = \log\mu_i, \qquad \text{and} \qquad \frac{\partial\eta_i}{\partial\mu_i} = \frac{1}{\mu_i} = \frac{1}{\exp\big(\boldsymbol{x}_i^\top\boldsymbol{\beta}\big)}.$$

Since $\mathrm{Var}(Y_i) = \exp\big(\boldsymbol{x}_i^\top\boldsymbol{\beta}\big)$,

$$w_i^{-1} = \mathrm{Var}(Y_i)\left(\frac{\partial\eta_i}{\partial\mu_i}\right)^2 = \exp\Big(\boldsymbol{x}_i^\top\boldsymbol{\beta}\Big)\left(\frac{1}{\exp\big(\boldsymbol{x}_i^\top\boldsymbol{\beta}\big)}\right)^2 = \frac{1}{\exp\big(\boldsymbol{x}_i^\top\boldsymbol{\beta}\big)}.$$

Therefore,

$$\begin{aligned} \frac{\partial\ell_i}{\partial\beta_j} &= (y_i - \exp\Big(\boldsymbol{x}_i^\top\boldsymbol{\beta}\Big))\exp\Big(\boldsymbol{x}_i^\top\boldsymbol{\beta}\Big)\frac{1}{\exp\big(\boldsymbol{x}_i^\top\boldsymbol{\beta}\big)}x_{ij} \\ &= y_ix_{ij} - x_{ij}\exp\Big(\boldsymbol{x}_i^\top\boldsymbol{\beta}\Big). \end{aligned}$$

Summing over i recovers (4.1).

3. The same result can also be written directly in vector form. Since $\boldsymbol{\mu} = \exp(\boldsymbol{X}\boldsymbol{\beta})$ componentwise, the score is

$$\boldsymbol{S}(\boldsymbol{\beta}) = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{\mu}),$$

whose jth component is exactly the expression in (4.1).

Estimating the Regression Coefficients

Recall that the observed and expected information matrices are

$$\boldsymbol{I}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}^\top} \qquad \text{and} \qquad I_{jk}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell}{\partial \beta_j\, \partial \beta_k},$$

and

$$\mathcal{I}(\boldsymbol{\beta}) = \mathbb{E}\left[\boldsymbol{I}(\boldsymbol{\beta})\right] \qquad \text{and} \qquad \mathcal{I}_{jk}(\boldsymbol{\beta}) = \mathbb{E}\left[-\frac{\partial^2 \ell}{\partial \beta_j\, \partial \beta_k}\right].$$

The Fisher information is usually simpler to calculate and invert than the observed information. Newton–Raphson and Fisher scoring both target $\hat{\boldsymbol{\beta}}$, the solution to $\boldsymbol{S}(\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$, and R uses Fisher scoring for generalized linear models.

The weighted least squares estimate from STAT 331 has the form

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

The generalized linear model estimation algorithm is called **iteratively reweighted least squares** (**IRWLS**) because the Fisher scoring update can be written in this same weighted least squares form, with the weights and adjusted response updated at each iteration.

Example 4.4 Continue the Poisson example with canonical link.

1. From (4.1),

$$S_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i x_{ij} - x_{ij} \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right) \right].$$

   Therefore, the observed information is

$$I_{jk}(\boldsymbol{\beta}) = -\frac{\partial}{\partial \beta_k} S_j(\boldsymbol{\beta}) = \sum_{i=1}^{n} x_{ij} \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right) x_{ik}.$$

Since this expression does not contain y_i, the expected information is the same:

$$\mathcal{I}_{jk}(\boldsymbol{\beta}) = \mathbb{E}\left[I_{jk}(\boldsymbol{\beta})\right] = \sum_{i=1}^{n} x_{ij} \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right) x_{ik}.$$

2. The general exponential family result gives

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

Here

$$\boldsymbol{W} = \text{Diag}(w_i) = \text{Diag}(\exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right)).$$

Thus $I_{jk} = \mathcal{I}_{jk}$ for the Poisson canonical link model. The observed information equals the expected information because we have selected the canonical link.

3. At iteration r,

$$\boldsymbol{W}(\hat{\boldsymbol{\beta}}^{(r)}) = \text{Diag}(w_i(\hat{\boldsymbol{\beta}}^{(r)})) = \text{Diag}(\exp\left(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}^{(r)}\right)) = \text{Diag}(\hat{\mu}_i^{(r)}),$$

where $\hat{\mu}_i^{(r)}$ is the fitted value for y_i at the current estimate $\hat{\boldsymbol{\beta}}^{(r)}$. The observations with the largest fitted values therefore receive the largest weights.

Lecture 5 — Tuesday, January 19, 2016

2. Logistic Regression and Dose-Response

Binary Data and Odds Ratios

For a probability $\pi \in (0, 1)$, the odds $\pi/(1-\pi)$ lie in $(0, \infty)$. Figure 10 shows that this transformation is one-to-one and monotonically increasing, with the odds increasing rapidly as π approaches one.

Consider two groups and let

$$\pi_1 = \mathbb{P}(\text{outcome} \mid \text{group 1})$$

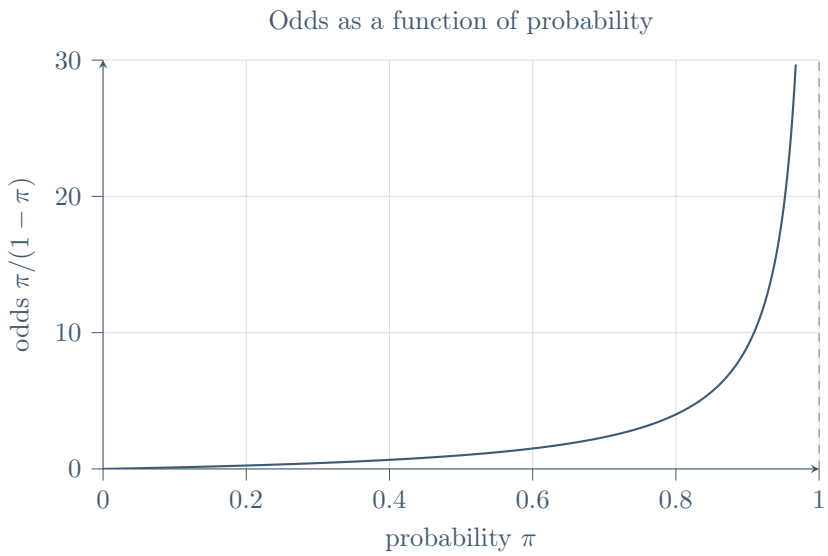

FIGURE 10

$$\pi_2 = \mathbb{P}(\text{outcome} \mid \text{group 2})$$

The odds ratio for group 1 relative to group 2 is

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

Example 5.1 If

$$\pi_1 = 0.5$$
$$\pi_2 = 0.25$$

then

$$\psi = \frac{0.5/0.5}{0.25/0.75} = 3 \qquad \text{and} \qquad \text{relative risk} = \frac{0.5}{0.25} = 2.$$

The odds ratio and the relative risk are not the same in general.

Suppose that $Y_1 \sim \text{Bin}(m_1, \pi_1)$ and $Y_2 \sim \text{Bin}(m_2, \pi_2)$ independently. The likelihood is

$$\begin{aligned} L(\pi_1, \pi_2) &= \mathbb{P}(Y_1 = y_1, Y_2 = y_2) \\ &= \mathbb{P}(Y_1 = y_1)\mathbb{P}(Y_2 = y_2) \end{aligned}$$

$$
\begin{aligned}
&= \binom{m_1}{y_1}\pi_1^{y_1}(1-\pi_1)^{m_1-y_1}\binom{m_2}{y_2}\pi_2^{y_2}(1-\pi_2)^{m_2-y_2} \\
&\propto \left(\frac{\pi_1}{1-\pi_1}\right)^{y_1}(1-\pi_1)^{m_1}\left(\frac{\pi_2}{1-\pi_2}\right)^{y_2}(1-\pi_2)^{m_2}\left[\frac{\pi_2/(1-\pi_2)}{\pi_2/(1-\pi_2)}\right]^{y_1} \\
&= \underbrace{\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right)^{y_1}}_{\psi=e^{\theta_1}}\underbrace{\left(\frac{\pi_2}{1-\pi_2}\right)^{y_1+y_2}}_{e^{(y_1+y_2)\theta_2}} \\
&\quad \times (1-\pi_1)^{m_1}\underbrace{(1-\pi_2)^{m_2}}_{(1+e^{\theta_2})^{-m_2}}.
\end{aligned}
$$

Reparameterize by setting

$$
\begin{aligned}
\theta_1 &= \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right) = \log\psi \\
\theta_2 &= \log\left(\frac{\pi_2}{1-\pi_2}\right)
\end{aligned}
$$

Remark 5.2 If

$$x = \log\left(\frac{y}{1-y}\right) = \text{logit}(y)$$

then

$$y = \frac{e^x}{1+e^x} = \text{expit}(x).$$

$$
\begin{aligned}
e^{\theta_2} &= \frac{\pi_2}{1-\pi_2} \\
e^{\theta_2} - \pi_2 e^{\theta_2} &= \pi_2 \\
e^{\theta_2} &= \pi_2(1+e^{\theta_2}) \\
\pi_2 &= \frac{e^{\theta_2}}{1+e^{\theta_2}} \\
1-\pi_2 &= \frac{1}{1+e^{\theta_2}}
\end{aligned}
$$

Thus the likelihood becomes

$$L(\theta_1,\theta_2) = (e^{\theta_1})^{y_1}(e^{\theta_2})^{y_1+y_2}(1+e^{\theta_1+\theta_2})^{-m_1}(1+e^{\theta_2})^{-m_2}$$

The parameters are now (θ_1, θ_2), and the target is $\psi = e^{\theta_1}$. The log-likelihood is

$$\ell(\theta_1, \theta_2) = y_1\theta_1 + (y_1 + y_2)\theta_2 - m_1 \log\Big(1 + e^{\theta_1+\theta_2}\Big) - m_2 \log\Big(1 + e^{\theta_2}\Big)$$

The score vector is

$$\boldsymbol{s}(\theta_1, \theta_2) = \begin{bmatrix} s_1(\theta_1, \theta_2) \\ s_2(\theta_1, \theta_2) \end{bmatrix},$$

where

$$\begin{aligned} s_1(\theta_1, \theta_2) &= \frac{\partial \ell(\theta_1, \theta_2)}{\partial \theta_1} = y_1 - m_1 \left(\frac{e^{\theta_1+\theta_2}}{1 + e^{\theta_1+\theta_2}} \right) \\ s_2(\theta_1, \theta_2) &= (y_1 + y_2) - m_1 \left(\frac{e^{\theta_1+\theta_2}}{1 + e^{\theta_1+\theta_2}} \right) - m_2 \left(\frac{e^{\theta_2}}{1 + e^{\theta_2}} \right) \end{aligned}$$

Solving

$$\boldsymbol{s}(\theta_1, \theta_2) = \mathbf{0}$$

gives the finite interior maximum likelihood estimates, provided $0 < y_j < m_j$ for $j = 1, 2$,

$$\begin{aligned} \hat{\theta}_1 &= \log \left(\frac{y_1/(m_1 - y_1)}{y_2/(m_2 - y_2)} \right) \\ \hat{\theta}_2 &= \log \left(\frac{y_2}{m_2 - y_2} \right) \end{aligned}$$

If a cell count is zero, the corresponding binomial maximum lies on the boundary and one or both log-odds estimates can be infinite; the Wald calculation below then does not apply without a separate small-sample or penalized treatment. Therefore, the estimate of the odds ratio is $\hat{\psi} = e^{\hat{\theta}_1}$. To do inference for ψ, we need the 2×2 information matrix. For example,

$$\begin{aligned} I_{11}(\theta_1, \theta_2) &= -\frac{\partial^2 \ell}{\partial \theta_1^2} \\ &= -\frac{\partial s_1(\theta_1, \theta_2)}{\partial \theta_1} \\ &= -\left\{ -m_1 \frac{[e^{\theta_1+\theta_2}](1 + e^{\theta_1+\theta_2}) - [e^{\theta_1+\theta_2}](e^{\theta_1+\theta_2})}{(1 + e^{\theta_1+\theta_2})^2} \right\} \\ &= m_1 \frac{e^{\theta_1+\theta_2}}{(1 + e^{\theta_1+\theta_2})^2} \end{aligned}$$

Notation 5.3 We write

$$\boldsymbol{I} = \begin{bmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{I}^{-1} = \begin{bmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{bmatrix}.$$

Proposition 5.4 (Schur complement) For a nonsingular 2×2 information matrix $\boldsymbol{I}$ with $I_{22} \neq 0$,

$$I^{11} = [I_{11} - I_{12} I_{22}^{-1} I_{21}]^{-1}.$$

Sketch proof for a 2×2 matrix. Let

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}.$$

Then

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

$$A^{11} = \frac{d}{ad - bc} = \frac{1}{a - bc/d} = [a - bd^{-1}c]^{-1} = [A_{11} - A_{12} A_{22}^{-1} A_{21}]^{-1}.$$

□

Returning to the binomial problem, the Schur complement formula gives

$$I^{11} = [I_{11} - I_{12} I_{22}^{-1} I_{21}]^{-1}.$$

In this case,

$$\begin{aligned} I_{11} &= m_1 \pi_1 (1 - \pi_1) \\ I_{21} = I_{12} &= m_1 \pi_1 (1 - \pi_1) \\ I_{22} &= m_1 \pi_1 (1 - \pi_1) + m_2 \pi_2 (1 - \pi_2) \end{aligned}$$

To use the Wald result for inference on θ_1, we need I^{11}:

$$I^{11} = \left[m_1 \pi_1 (1 - \pi_1) - \frac{\{m_1 \pi_1 (1 - \pi_1)\}^2}{m_1 \pi_1 (1 - \pi_1) + m_2 \pi_2 (1 - \pi_2)} \right]^{-1}$$

$$
\begin{aligned}
&= \left[\frac{m_1\pi_1(1-\pi_1)m_2\pi_2(1-\pi_2)}{m_1\pi_1(1-\pi_1)+m_2\pi_2(1-\pi_2)}\right]^{-1} \\
&= \frac{1}{m_1\pi_1(1-\pi_1)} + \frac{1}{m_2\pi_2(1-\pi_2)} \\
&= \frac{1}{m_1\pi_1} + \frac{1}{m_1(1-\pi_1)} + \frac{1}{m_2\pi_2} + \frac{1}{m_2(1-\pi_2)}
\end{aligned}
$$

Remark 5.5 The algebra in the previous display uses

$$\frac{1}{m_1\pi_1} + \frac{1}{m_1(1-\pi_1)} = \frac{(1-\pi_1)+\pi_1}{m_1\pi_1(1-\pi_1)} = \frac{1}{m_1\pi_1(1-\pi_1)}.$$

We evaluate $I^{11}(\hat{\theta}_1, \hat{\theta}_2)$ using the invariance property:

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

where $\hat{\pi}_1 = y_1/m_1$ and $\hat{\pi}_2 = y_2/m_2$.

The corresponding 2×2 table is

y_1	$m_1 - y_1$	m_1
y_2	$m_2 - y_2$	m_2

Thus the estimated asymptotic Wald variance is

$$\widehat{\text{Var}}(\hat{\theta}_1) = \widehat{\text{Var}}(\log\hat{\psi}) = \frac{1}{y_1} + \frac{1}{m_1-y_1} + \frac{1}{y_2} + \frac{1}{m_2-y_2}.$$

Example 5.6 In the prenatal care data, the outcome is fetal mortality and the explanatory

variable is intensive versus regular care. The pooled data are

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

The fitted mortality probabilities are

$$\hat{\pi}_1 = \frac{20}{336} = 0.060 \qquad \text{and} \qquad \hat{\pi}_2 = \frac{46}{419} = 0.11,$$

where group 1 is intensive care and group 2 is regular care.

$$\hat{\psi} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \frac{20/316}{46/373} = \frac{20 \cdot 373}{46 \cdot 316} = 0.5132.$$

To obtain a confidence interval for $\hat{\psi}$, first obtain a confidence interval for $\hat{\theta}_1 = \log \hat{\psi}$:

$$\hat{\theta}_1 = \log \hat{\psi} = -0.6671$$

$$\begin{aligned} \text{Var}(\hat{\theta}_1) &= \frac{1}{y_1} + \frac{1}{m_1 - y_1} + \frac{1}{y_2} + \frac{1}{m_2 - y_2} \\ &= \frac{1}{20} + \frac{1}{316} + \frac{1}{46} + \frac{1}{373} \\ &= 0.07758 \end{aligned}$$

$$\begin{aligned} \text{95\% confidence interval for } \theta_1: \quad \hat{\theta}_1 &\pm 1.96\sqrt{\text{Var}(\hat{\theta}_1)} \\ &= -0.6671 \pm 1.96\sqrt{0.07758} \\ &= (-1.2130, -0.1211) \end{aligned}$$

Exponentiating the endpoints gives the 95% confidence interval for ψ:

$$(e^{-1.2130}, e^{-0.1211}) = (0.297, 0.886).$$

Thus the odds of fetal mortality in the intensive treatment group are 0.51 times the odds in the regular treatment group, with 95% confidence interval $(0.30, 0.89)$. This suggests a strong association between mortality and level of care in the pooled data.

The conclusion changes after stratifying by clinic:

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

The estimates of the odds ratio are $\hat{\psi}_A = 0.80$ for clinic A and $\hat{\psi}_B = 1.01$ for clinic B. Thus the analyses within clinics show a weak association in clinic A and no association in clinic B.

This discrepancy is explained by confounding. There is a strong association between level of care and clinic: in clinic A, a patient is much more likely to receive intensive care. There is also a strong association between mortality and clinic: clinic A has lower mortality than clinic B, perhaps because clinic B treats more high-risk patients. This is an instance of Simpson's paradox, with clinic acting as a confounder. Rather than estimate several separate odds ratios, we would like to estimate a single odds ratio adjusted for clinic. This motivates binomial regression models.

For binomial generalized linear models, several link functions map probabilities in $(0, 1)$ to the real line. Figure 11 compares the logit, complementary log-log, and probit links. The logit link is the canonical link for the binomial distribution and is the link used to set up logistic regression.

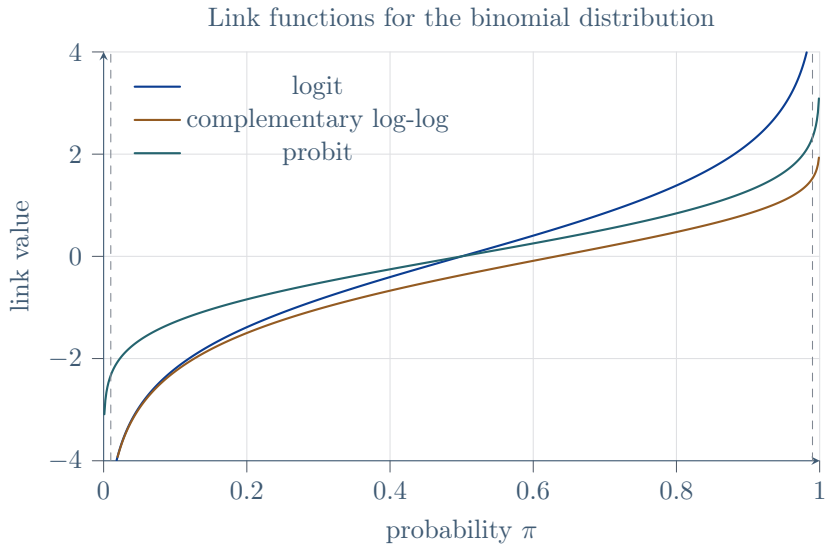

FIGURE 11

Lecture 6 — Thursday, January 21, 2016

Logistic Regression

The binomial distribution is a member of the exponential family. In a binomial regression model,

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

and π_i is the response probability for subject or group i. The explanatory variables are collected in

$$\boldsymbol{x}_i = (x_{i0}, x_{i1}, \ldots, x_{i,p-1})^\top,$$

where $x_{i0} = 1$ supplies the intercept, and the linear predictor is

$$\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}.$$

The logit link is the canonical link for the binomial distribution, giving the logistic regression model

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

In the simple logistic regression model,

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

Assume that x_{i1} is binary, with

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

Then the fitted log odds are as follows:

x_{i1}	$\log(\pi_i/(1-\pi_i))$
1	$\beta_0 + \beta_1$
0	β_0

Remark 6.1

$$\beta_0 = \log\left(\frac{\pi_0}{1-\pi_0}\right),$$

so β_0 is the log odds of response for group 0.

Subtracting the group 0 row from the group 1 row gives

$$\begin{aligned} \beta_1 &= \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_0}{1-\pi_0}\right) \\ &= \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}\right) \\ &= \log\psi. \end{aligned}$$

Therefore, β_1 is the log odds ratio of response for group 1 versus group 0, and

$$\exp(\beta_1) = \psi.$$

Since $\mathbb{E}[Y_i] = m_i\pi_i$, the fitted response probability and fitted number of responses are

$$\hat{\pi}_i = \frac{e^{\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}}}{1 + e^{\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}}} = \text{expit}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) \qquad \text{and} \qquad \hat{\mu}_i = m_i\hat{\pi}_i.$$

For the prenatal care data, the clinic-only model is

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

where $x_{i1} = 1$ for clinic A and $x_{i1} = 0$ for clinic B. The fitted log odds are

Clinic	$\boldsymbol{x}_i^\top$	$\log(\pi_i/(1-\pi_i))$
A	$(1,1)$	$\beta_0 + \beta_1$
B	$(1,0)$	β_0

Thus β_0 is the log odds of fetal mortality for mothers treated at clinic B, and β_1 is the log odds ratio for clinic A versus clinic B.

For the prenatal care data, consider the main effects model

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

Here $x_{i1} = 1$ for clinic A and $x_{i1} = 0$ for clinic B, while $x_{i2} = 1$ for intensive care and $x_{i2} = 0$ for regular care. The resulting fitted log odds are

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

With the coding used here, β_0 is the log odds of fetal mortality for mothers treated at clinic B with regular care. The coefficient β_1 is obtained either as row 2 minus row 4, giving the log odds ratio for clinic A versus clinic B at regular care, or as row 1 minus row 3, giving the log odds ratio for clinic A versus clinic B at intensive care. Thus β_1 is the log odds ratio of fetal mortality for mothers treated at clinic A versus clinic B at the same level of care. Here mortality is the response, clinic A versus clinic B is the comparison, and level of care is the control factor.

Similarly, β_2 is row 1 minus row 2, giving the log odds ratio for intensive care versus regular care at clinic A, and it is also row 3 minus row 4, giving the log odds ratio for intensive care versus regular care at clinic B.

The interaction model is

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

where $x_{i3} = x_{i1}x_{i2}$. Then β_1 is the log odds ratio for clinic A versus clinic B at regular care, while $\beta_1 + \beta_3$ is the log odds ratio for clinic A versus clinic B at intensive care. In the main effects model these two odds ratios are constrained to be the same, whereas β_3 allows the association between mortality and clinic to depend on level of care. Similarly, β_2 is the log odds ratio for intensive care versus regular care at clinic B, while $\beta_2 + \beta_3$ is the corresponding log odds ratio at clinic A. If $\beta_3 = 0$, the interaction model reduces to the main effects model.

For the interaction model, the fitted log odds are

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

Prenatal Care Data

The pooled data initially suggested an association between mortality and level of care, with $\hat{\psi} = 0.51$.

After stratifying by clinic, the association vanished: $\hat{\psi}_A = 0.80$ and $\hat{\psi}_B = 1.01$.

The qualitative dependence structure is summarized in Figure 12.

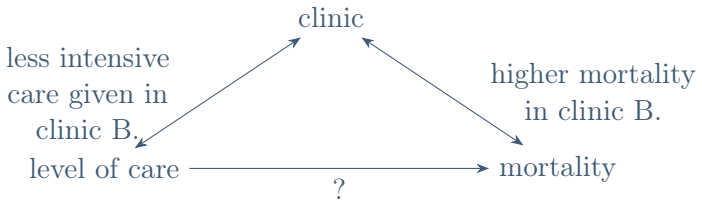

FIGURE 12

If clinic is ignored, intensive care appears to be associated with lower mortality.

The 2×2 tables for the two clinics are as follows.

		D	S	
A	I	16	293	309
	R	12	176	188

		D	S	
B	I	4	23	27
	R	34	197	231

In R, the response variable is supplied as `resp = (y, m-y)`, where `y` is the number of deaths and `m-y` is the number of survivors. If the data are not supplied in this form, then the response vector must be created before fitting the model. Help for the fitting function is available with `?glm`.

```
model1 = glm(resp ~ loc, family = binomial(link = "logit"),
             data = prenatal.dat)
```

Here `model1` is the fitted object, `resp` is the response $(y, m-y)$, `loc` is the covariate in the linear predictor, `family = binomial` specifies the binomial distribution, `link = "logit"` specifies the link, and `data = prenatal.dat` specifies the data object.

The selected R fits used for the prenatal care analysis are summarized below.

Model	Formula	Residual deviance	Residual degrees of freedom
1	`resp ~ loc`	10.8144	2
2	`resp ~ clinic + loc`	0.1069	1
3	`resp ~ loc + clinic + loc:clinic`	0.0000	0
4	`resp ~ clinic`	0.3148	2

The fitted coefficients most often used below are

Model	Coefficient	Estimate	standard error
1	`loc`	−0.6671	0.2785
2	`clinic`	−0.9863	0.3089
2	`loc`	−0.1503	0.3302
3	`loc`	0.0076	0.5727
3	`clinic`	−0.9287	0.3514
3	`loc:clinic`	−0.2296	0.6949
4	`clinic`	−1.0624	0.2621

For model 1 of the prenatal example, test

$$H_0 : \beta_2 = 0 \qquad \text{versus} \qquad H_a : \beta_2 \neq 0.$$

$$\frac{\hat{\beta}_2 - 0}{\text{standard error}(\hat{\beta}_2)} = \frac{-0.6671}{0.2785} = -2.3949.$$

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

where $Z \sim \mathcal{N}(0, 1)$.

Therefore, we reject the null hypothesis that $\beta_2 = 0$. A confidence interval for β_2 is

$$\hat{\beta}_2 \pm 1.96 \text{ standard error}(\hat{\beta}_2).$$

The corresponding confidence interval for the odds ratio is

$$\exp\Big(\hat{\beta}_2 \pm 1.96 \text{ standard error}(\hat{\beta}_2)\Big) = \exp(-0.6671 \pm 1.96(0.2785)) = (0.30, 0.89),$$

the same interval obtained from the 2×2 table calculation.

It is important not to compute this interval as $e^{\hat{\beta}_2} \pm 1.96 e^{\text{standard error}(\hat{\beta}_2)}$.

For model 2,

$$\exp\Big(\hat{\beta}_2\Big) = 0.86.$$

This is the odds ratio for mortality under intensive care versus regular care at the same clinic. Adjusting for clinic by including it in the model removes the association between mortality and level of care seen in model 1.

Lecture 7 — Tuesday, January 26, 2016

Prenatal Care Model Comparisons

The estimates of the odds ratio within each clinic were

$$\hat{\psi}_A = 0.80$$
$$\hat{\psi}_B = 1.01$$

Exercise 7.1 Recreate the estimates $\hat{\psi}_A$ and $\hat{\psi}_B$ using model 3: interaction model.

$$\psi_A : (\text{row } 1 - \text{row } 2) = \exp\left(\hat{\beta}_2 + \hat{\beta}_3\right) \qquad \psi_B : (\text{row } 3 - \text{row } 4) = \exp\left(\hat{\beta}_2\right)$$

Model 3 is the interaction model

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

It has $p = 4$ parameters and there are $n = 4$ data points, with $Y_i \sim \text{Bin}(m_i, \pi_i)$. Therefore, model 3 is saturated, with $n = p$. In this case, allowing boundary fitted probabilities when $y_i \in \{0, m_i\}$,

$$\hat{\pi}_i = \text{expit}(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}) = \frac{y_i}{m_i},$$

the binomial maximum likelihood estimate.

For logistic regression, the likelihood can be written as

$$\ell(\pi) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) \right].$$

Using the logit link,

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

and

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

Substitution gives

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[y_i\boldsymbol{x}_i^\top\boldsymbol{\beta} - m_i\log\left(1+e^{\boldsymbol{x}_i^\top\boldsymbol{\beta}}\right)\right].$$

Likelihood Ratio Tests

Let $\hat{\theta}$ be the estimate under a reduced model with p parameters and let $\tilde{\theta}$ be the estimate under a containing full model with $q > p$ parameters. Under the reduced model and the usual regularity conditions, the likelihood ratio statistic is

$$D = -2\left(\ell(\hat{\theta}) - \ell(\tilde{\theta})\right) \dot{\sim} \chi^2_{q-p}.$$

When the full model is saturated, this statistic is the residual deviance of the reduced model. Since the unconstrained model fits at least as well as the constrained model, $\ell(\hat{\theta}) \leq \ell(\tilde{\theta})$, and hence $D \geq 0$.

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

For logistic regression, the saturated model has $\tilde{\pi}_i = y_i/m_i$, while an unsaturated p-dimensional model has $\hat{\pi}_i = \text{expit}(\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}})$. The deviance statistic is

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

$$= 2\sum_{i=1}^{n}\left[\underbrace{y_i}_{O_{i1}}\log\left(\frac{y_i}{m_i\hat{\pi}_i}\right)_{E_{i1}} + \underbrace{(m_i-y_i)}_{O_{i2}}\log\left(\frac{m_i-y_i}{m_i(1-\hat{\pi}_i)}\right)_{E_{i2}}\right]$$

Here and below, a term $0\log(0/E)$ is interpreted as 0.

Under H_0, $D \dot{\sim} \chi^2_{n-p}$. The statistic has the familiar form

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

Example 7.2 Model 3 is saturated, so

$$D = -2(\ell(\tilde{\pi}) - \ell(\tilde{\pi})) = 0.$$

For model 4, the clinic-only model, $D = 0.3148$. The degrees of freedom are

$$n - p = 4 - 2 = 2,$$

as reported by R.

For a deviance likelihood ratio test of model 1 against model 2, the null hypothesis is that model 1 is adequate compared to model 2. The alternative hypothesis is that model 1 is not adequate. In other words, we are asking whether the simpler p-dimensional model 1 can be used instead of the more complex q-dimensional model 2, where $p < q$.

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

where $\hat{\pi}$ comes from the reduced model, $\tilde{\pi}$ comes from the full model, and $\tilde{\tilde{\pi}}$ comes from the saturated model. Here D_0 is the deviance for model 1 and D_A is the deviance for model 2.

Under H_0,

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

This agrees with the degrees-of-freedom difference because $(n-p)-(n-q)=q-p$.

Remark 7.3 If the reduced and full models are each correctly specified, their residual deviances have the respective approximations $D_0 \dot{\sim} \chi^2_{n-p}$ and $D_A \dot{\sim} \chi^2_{n-q}$. The nested-model test itself uses the direct likelihood ratio result $\Delta D \dot{\sim} \chi^2_{q-p}$ under the reduced model.

There are two main types of tests for logistic regression in this section. A Wald test has the form

$$H_0 : \beta_k = 0 \qquad \text{versus} \qquad H_a : \beta_k \neq 0.$$

A deviance likelihood ratio test can test several coefficients at once:

$$H_0 : \beta_k = \beta_{k+1} = \cdots = 0$$

against the alternative that at least one of $\beta_k, \beta_{k+1}, \ldots, \beta_{q-1}$ is nonzero.

Example 7.4 Model 4 is the clinic-only model

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

and model 2 is the main effects model

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

Model 4 is nested in model 2, and we test whether model 4 is adequate compared to model 2:

$$H_0 : \beta_2 = 0 \qquad \text{versus} \qquad H_a : \beta_2 \neq 0.$$

The deviance likelihood ratio statistic is

$$\Delta D = D_4 - D_2 = 0.3148 - 0.1069 = 0.2079.$$

The p-value is

$$\mathbb{P}(\chi^2_{3-2} > 0.2079) = 0.648 > 0.05.$$

We do not reject the null hypothesis that model 4 is adequate compared to model 2. Equivalently, we do not reject $\beta_2 = 0$, and we do not reject the hypothesis of no association between level of care and mortality after adjusting for clinic. This is the same conclusion as the Wald test.

Recall that in multiple linear regression the residuals are

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}},$$

and the standardized residuals are approximately $\mathcal{N}(0,1)$. We want analogous residuals for generalized linear models.

$$D(\hat{\pi}) = \sum_{i=1}^{n} d_i \qquad \text{and} \qquad D(\hat{\pi}) \dot{\sim} \chi^2_{n-p},$$

where d_i is the contribution to deviance from subject i. The deviance residuals are

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

Under an adequate model, deviance residuals should be roughly centred at zero and have a stable scale, so we inspect them for systematic patterns and unusually large values. Because they use fitted parameters, they need not be independent or exactly $\mathcal{N}(0,1)$.

The Pearson residuals for binomial data are

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

Under a good model, Pearson residuals should likewise be roughly centred at zero and have a stable scale. If either $m_i\hat{\pi}_i$ or $m_i(1-\hat{\pi}_i)$ is small, then the normal and chi-square approximations used for residual and deviance diagnostics should be treated with caution.

Lecture 8 — Thursday, January 28, 2016

Neuroblastoma Example

For nested logistic regression models, the reduced model under H_0 is

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

and the full model under H_a is

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

Then

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

Example 8.1 In the neuroblastoma example, the outcome is $\mathbb{1}_{\{\text{survival of at least 2 years}\}}$. The explanatory variables are age, with three levels, and stage, with five levels. There are $n = 15$ observations, and each table entry is of the form y/m, where y is the number of children who survived and m is the number diagnosed in that combination of age and stage.

Age in months	Stage I	Stage II	Stage III	Stage IV	Stage 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

The marginal totals are

Age in months	Stage I	Stage II	Stage III	Stage IV	Stage V	Total
0–11	11/12	15/16	2/4	5/18	18/19	51/69 (0.74)
12–23	3/4	3/7	5/8	0/25	1/3	12/47 (0.26)
24+	4/5	4/12	3/15	3/93	2/5	16/130 (0.12)
Total	18/21 (0.86)	22/35 (0.62)	10/27 (0.37)	8/136 (0.06)	21/27 (0.78)	79/246 (0.32)

The data show higher survival at younger ages at diagnosis, overall survival of 32%, and higher survival at lower stages of disease, except for stage V.

Since age has three levels, it requires two indicator variables. Since stage has five levels, it requires four indicator variables. The lowest group is used as the comparison group: for age 0–11 months, $x_{i1} = x_{i2} = 0$, and for stage I,

$$x_{i3} = x_{i4} = x_{i5} = x_{i6} = 0.$$

Model	Deviance	p	$n - p$
Age and stage	9.625	7	8
Age only	83.583	3	12
Stage only	42.446	5	10

For the age-and-stage model, the fitted coefficient estimates are

Coefficient	Estimate	standard error
β_0	3.3175	0.7721
β_1 (age 12–23)	−2.1181	0.5736
β_2 (age 24+)	−2.6130	0.5017
β_3 (stage II)	−1.2529	0.7837
β_4 (stage III)	−1.7759	0.8003
β_5 (stage IV)	−4.3678	0.7902
β_6 (stage V)	−1.0222	0.8644

To assess model fit, we ask whether the residuals are roughly centred at zero, have a stable scale, and show no systematic pattern. Values beyond about ±1.96 deserve attention, although this is a diagnostic guide rather than an exact normal-reference rule. Figure 13 shows the deviance and Pearson residuals for the fitted neuroblastoma model with age and stage.

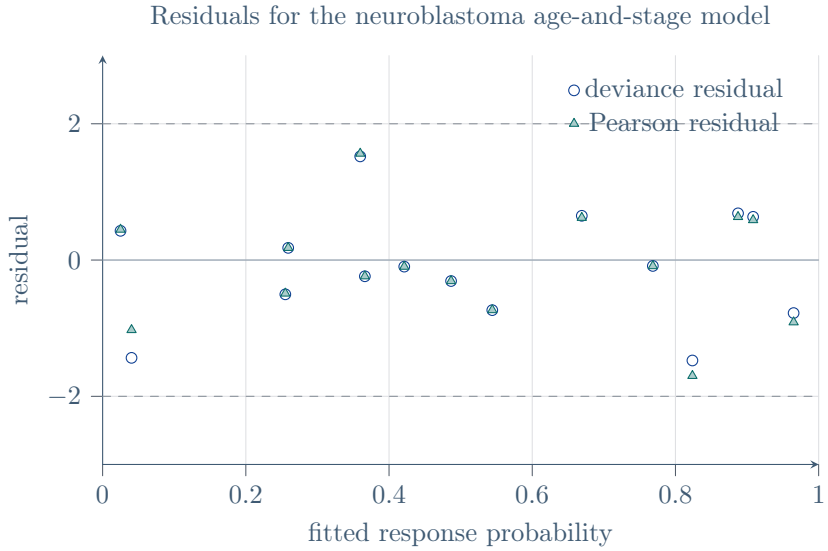

FIGURE 13

First ask whether stage is important after controlling for age. The null hypothesis is that model 2 is adequate compared to model 1, so stage is not associated with survival after controlling for age:

$$H_0 : \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0.$$

The alternative hypothesis is that model 2 is not adequate, so stage is associated with survival

after controlling for age; that is, at least one of $\beta_3, \beta_4, \beta_5, \beta_6$ is nonzero. The test statistic is

$$\Delta D = D_0 - D_A = D_2 - D_1 = 83.583 - 9.625 \dot{\sim} \chi^2_4 \text{ under } H_0.$$

$$p\text{-value} = \mathbb{P}(\chi^2_{(4)} > 73.958) < 0.001.$$

Therefore, we reject H_0 and conclude that stage is associated with survival. Stage should be kept in the regression model.

Next ask whether age is important. This compares model 3 against model 1:

$$H_0 : \beta_1 = \beta_2 = 0 \qquad \text{versus} \qquad H_a : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0.$$

$$\Delta D = D_0 - D_A = D_3 - D_1 = 42.446 - 9.625 \dot{\sim} \chi^2_2 \text{ under } H_0.$$

$$p\text{-value} = \mathbb{P}(\chi^2_{(2)} > 32.821) < 0.001.$$

Therefore, we reject H_0 and conclude that age is associated with survival. Age should be kept in the regression model.

Finally, ask whether an age-by-stage interaction is needed. The interaction model is saturated:

number of parameters	1	intercept
	+ 2	age parameters
	+ 4	stage parameters
	+ (2)(4)	interaction parameters
	$15 = n =$	observations.

Thus $D_4 = 0$. The null hypothesis is that model 1 is adequate compared to model 4, while the alternative is that model 1 is not adequate. The test statistic is

$$\Delta D = D_0 - D_A = D_1 - D_4 = 9.625 \dot{\sim} \chi^2_8 \text{ under } H_0.$$

$$p\text{-value} = \mathbb{P}(\chi^2_{(8)} > 9.625) = 0.29.$$

Therefore, we do not reject H_0; model 1 is adequate. We do not need an age-by-stage interaction, so model 1 is selected for analysis and interpretation.

Because several cells are sparse and one observed cell count is zero, the saturated fit reaches the boundary and the chi-square approximation for this interaction test should be treated cautiously. A parametric bootstrap likelihood ratio test would provide a more reliable small-sample check; the displayed calculation is the usual asymptotic analysis.

Remark 8.2 The null deviance in R is the deviance for the intercept-only model.

Two odds ratio interpretations are useful. First, holding age fixed but unspecified, the odds ratio for stage IV versus stage I is

$$\exp\Big(\hat{\beta}_5\Big) = \exp(-4.368) = 0.013,$$

where $x_5 = \mathbb{1}_{\{\text{stage=IV}\}}$. Controlling for age, the odds of surviving two years among children diagnosed at stage IV are 0.013 times the odds among children diagnosed at stage I. A Wald confidence interval is

$$\exp(-4.368 \pm 1.96(0.7902)) = (0.003, 0.060).$$

Second, holding stage fixed but unspecified, let ψ be the odds ratio for age 24+ months versus age 12–23 months. Then

$$\hat{\psi} = \exp\Big(\hat{\beta}_2 - \hat{\beta}_1\Big) = 0.61.$$

Thus survival is higher for children aged 12–23 months than for children aged 24+ months.

To get a confidence interval for $\exp\Big(\hat{\beta}_2 - \hat{\beta}_1\Big)$, we need

$$\mathrm{Var}(\hat{\beta}_2 - \hat{\beta}_1) = \underbrace{\mathrm{Var}(\hat{\beta}_2) + \mathrm{Var}(\hat{\beta}_1)}_{\substack{\text{squares of the standard errors} \\ \text{reported by R}}} - \underbrace{2\,\mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_1)}_{\substack{\text{from the covariance matrix} \\ \boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}) \\ \text{reported as } \texttt{cov.unscaled}}}.$$

Equivalently, use the contrast vector

$$\boldsymbol{c}^\top \hat{\boldsymbol{\beta}} = \hat{\beta}_2 - \hat{\beta}_1 \qquad \text{and} \qquad \boldsymbol{c} = (0, -1, 1, 0, 0, 0, 0).$$

$$\begin{aligned}\text{Var}(\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}) &= \underset{1\times p}{\boldsymbol{c}^\top}\underset{p\times p}{\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})}\underset{p\times 1}{\boldsymbol{c}} \\ &= 0.2649 \text{ from R.}\end{aligned}$$

Since $\hat{\beta}_2 - \hat{\beta}_1 = -0.495$, the approximate 95% confidence interval for $\beta_2 - \beta_1$ is

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

Exponentiating gives the confidence interval

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

for the odds ratio.

Exercise 8.3 Check the preceding calculation using the `cov.unscaled` output.

The same calculation gives confidence intervals for other functions of the linear predictor. Since $\hat{\boldsymbol{\beta}}$ is a maximum likelihood estimate,

$$\hat{\boldsymbol{\beta}} \text{ is approximately } \mathcal{N}_p\left(\boldsymbol{\beta}, \boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})\right).$$

Thus

$$\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} \text{ is approximately } \mathcal{N}\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}, \boldsymbol{x}_i^\top\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i\right),$$

and

$$\frac{\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} - \boldsymbol{x}_i^\top\boldsymbol{\beta}}{\sqrt{\boldsymbol{x}_i^\top\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i}}$$

is approximately standard normal. Therefore an approximate 95% confidence interval for $\eta_i = \boldsymbol{x}_i^\top\boldsymbol{\beta}$ is

$$\boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}} \pm 1.96\sqrt{\boldsymbol{x}_i^\top\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i} = (\hat{\eta}_L, \hat{\eta}_U).$$

If an odds ratio is expressed as $\exp\left(\boldsymbol{c}^\top\boldsymbol{\beta}\right)$, then an approximate 95% confidence interval is

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

Finally, if the response probability is

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

then transforming the endpoints of the interval for η_i gives

$$\left(\frac{e^{\hat{\eta}_L}}{1+e^{\hat{\eta}_L}}, \frac{e^{\hat{\eta}_U}}{1+e^{\hat{\eta}_U}}\right).$$

Lecture 9 — Tuesday, February 02, 2016

Bioassay and Dose-Response Models

In a bioassay experiment, several groups of subjects are exposed to varying levels of a toxin or drug, and the number of subjects showing a binary response is recorded within a fixed period of time. The stimulus is the applied dose, often measured as

$$x = \log z,$$

where z is the concentration. The response is commonly of the form died or survived. Each subject is assumed to have a threshold dose above which the response occurs. This threshold is called the tolerance, and since tolerances vary across individuals, they are modelled with a population distribution.

Let x denote the dose of the stimulus, and let $f(x)$ be the density for tolerance in the population. Figure 14 shows the response probability as the area under the tolerance density up to the applied dose.

If the dose x_0 is applied to the population, the proportion that responds is

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

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

The purpose of a dose-response model is to describe the relationship between the dose x and the probability of response π. For each group $j = 1, \ldots, J$, let m_j denote the number of subjects in the group, let x_j denote the dose applied to subjects in the group, and let Y_j denote the number of subjects with the response. The model is

$$Y_j \sim \text{Bin}(m_j, \pi_j), \quad j = 1, \ldots, J,$$

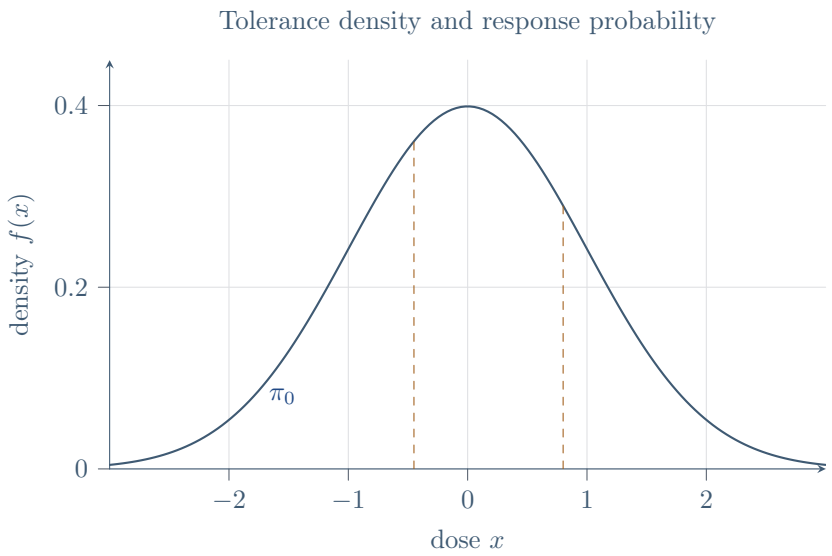

FIGURE 14

with the random variables Y_j independent, where π_j is the probability of response in group j.

Since

$$\mathbb{E}\left[Y_j/m_j\right] = \pi_j,$$

we model π_j as a function $\pi_j(x_j)$ of the continuous dose covariate.

Link Functions for Dose Response

Since $0 \leq \pi(x) \leq 1$, it is natural to introduce a link function g and write

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

Equivalently,

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

Figure 15 shows the typical increasing shape of a fitted dose-response curve.

1. The fitted value must satisfy $0 \leq \pi(x) \leq 1$.
2. The fitted dose-response curve should be nondecreasing in x.
3. A cumulative distribution function is therefore a natural choice for the inverse link.

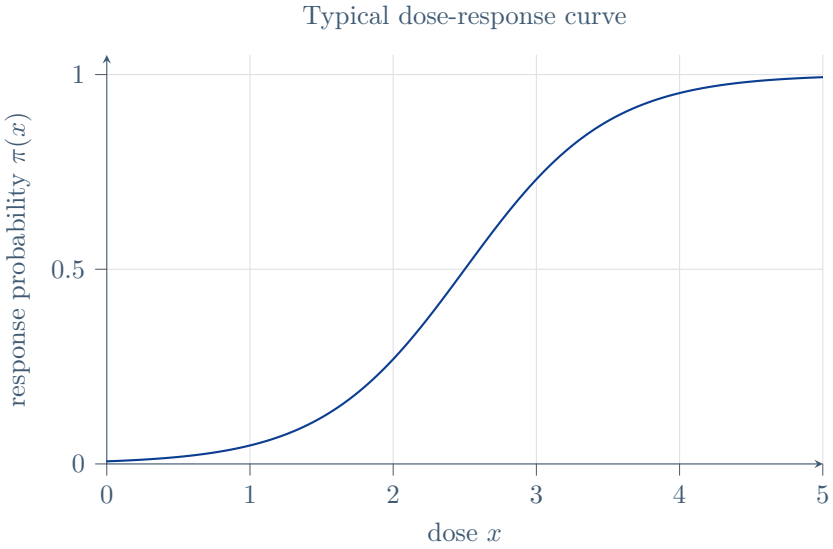

FIGURE 15

Thus we often take

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

where F^* is a cumulative distribution function and $\beta_1 > 0$ when the response probability is meant to increase with dose.

The tolerance distribution gives the same probability in the form

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

where $\pi(x)$ is the probability of response at dose x, and f is the density of the tolerance distribution.

Remark 9.1 Equating the model $\pi(x) = F^*(\beta_0 + \beta_1 x)$ with the tolerance distribution integral determines the link function. The density f controls how quickly the probability of response changes with the dose.

Equating these two representations defines the tolerance density as

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

so the tolerance density determines the local rate at which the response probability increases with dose.

The modelling procedure is as follows.

1. Choose a tolerance distribution.
2. Find the implied link function.
3. Fit a binary/binomial regression model.

For example, suppose that the tolerance distribution is $\mathcal{N}(\mu, \sigma^2)$. Then

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

where Φ is the cumulative distribution function of $\mathcal{N}(0,1)$. Setting this equal to the inverse link gives

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

Therefore $g(\cdot) = \Phi^{-1}(\cdot)$, which is called the probit link, and the regression equation is

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

The interpretation of β_0 and β_1 is in terms of the tolerance distribution parameters μ and σ^2, not in terms of log odds ratios. Figure 16 indicates the role of μ as the centre of the tolerance distribution.

μ is the dose at which 50% of the population responds, and σ^2 is the variance in the response threshold.

For the probit link,

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

Thus

$$\Phi\left(\frac{x}{\sigma} - \frac{\mu}{\sigma}\right) = \Phi(\beta_0 + \beta_1 x),$$

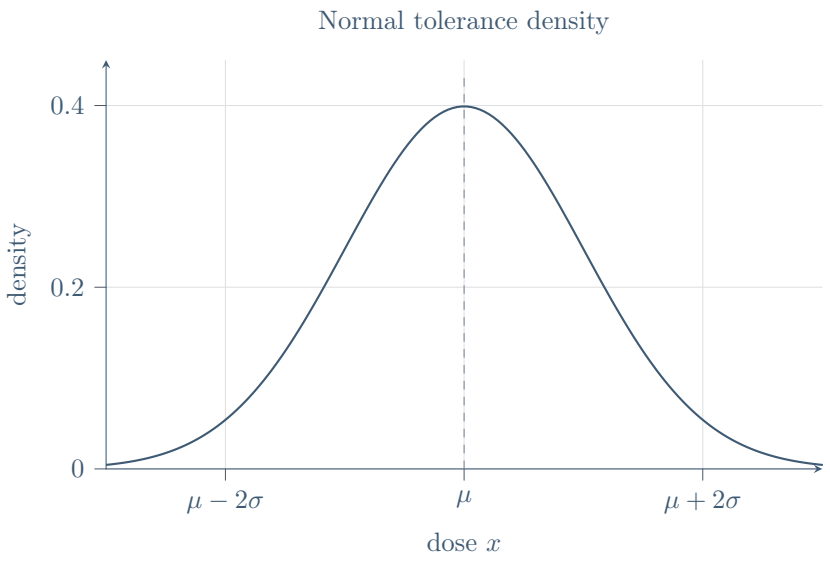

FIGURE 16

so

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

This gives the relationship between the tolerance parameters (μ, σ^2) and the regression parameters (β_0, β_1).

Example 9.2 Let δ be the median effective dose, so that $\pi(\delta) = 0.50$. Then

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

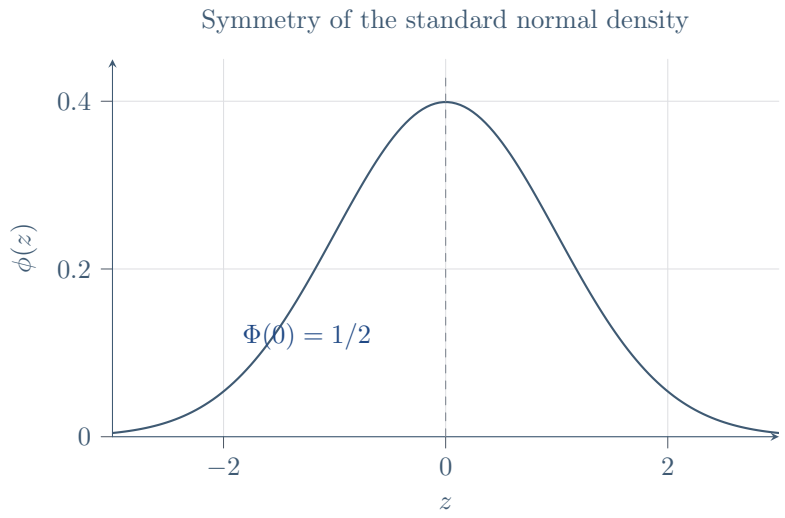

FIGURE 17

Figure 17 shows the symmetry behind $\Phi^{-1}(0.50) = 0$. This equality follows from the identity

$$\Phi(x) = \int_{-\infty}^{x} f(s)\,\mathrm{d}s.$$

Since $\Phi(0) = 0.50$, we have $0 = \Phi^{-1}(0.50)$.

Tolerance distribution	Inverse link $\pi = g^{-1}(\eta)$	Link $g(\pi) = \eta$
Normal	$\Phi(\eta)$	$\Phi^{-1}(\pi)$
Logistic	$e^{\eta}/(1+e^{\eta})$	$\log(\pi/(1-\pi))$
Extreme value	$1-\exp(-e^{\eta})$	$\log(-\log(1-\pi))$

For the logistic tolerance distribution, this table comes from

$$\pi(x) = \int_{-\infty}^{x} \frac{\beta_1 e^{\beta_0+\beta_1 s}}{(1+e^{\beta_0+\beta_1 s})^2}\,\mathrm{d}s = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}},$$

which gives the logit link. When the tolerance distribution is an extreme value distribution,

$$\pi(x) = \int_{-\infty}^{x} \beta_1 e^{\beta_0+\beta_1 s - e^{\beta_0+\beta_1 s}}\,\mathrm{d}s = \int_{-\infty}^{\eta} e^{\nu-e^{\nu}}\,\mathrm{d}\nu = 1-\exp(-e^{\eta}),$$

where $\eta = \beta_0 + \beta_1 x$. This gives the complementary log-log link

$$\eta = \log(-\log(1-\pi)).$$

Example 9.3 (Beetle mortality) Bliss's beetle mortality experiment exposed groups of beetles to varying concentrations of carbon disulphide gas. The dose is the logarithm of concentration.

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

The response is represented in R by `cbind(y, m-y)`, and three binomial regression models are fit using the same linear predictor $\eta = \beta_0 + \beta_1 x$:

1. Model 1 uses the logit link.
2. Model 2 uses the probit link.
3. Model 3 uses the complementary log-log link.

The selected fitted coefficients are

Link	$\hat{\beta}_0$	$\hat{\beta}_1$	Residual deviance
logit	−60.7175	34.2703	11.2322
probit	−34.9353	19.7279	10.1198
complementary log-log	−39.5723	22.0412	3.4464

Each residual deviance is on 6 degrees of freedom. These models have different links, so a likelihood ratio deviance test is not an appropriate comparison. Instead, compare the

deviance residual plots in Figure 18 and the fitted dose-response curves in Figure 19, where

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

is compared with the observed proportions y_i/m_i. The complementary log-log link gives the best fit by these diagnostics.

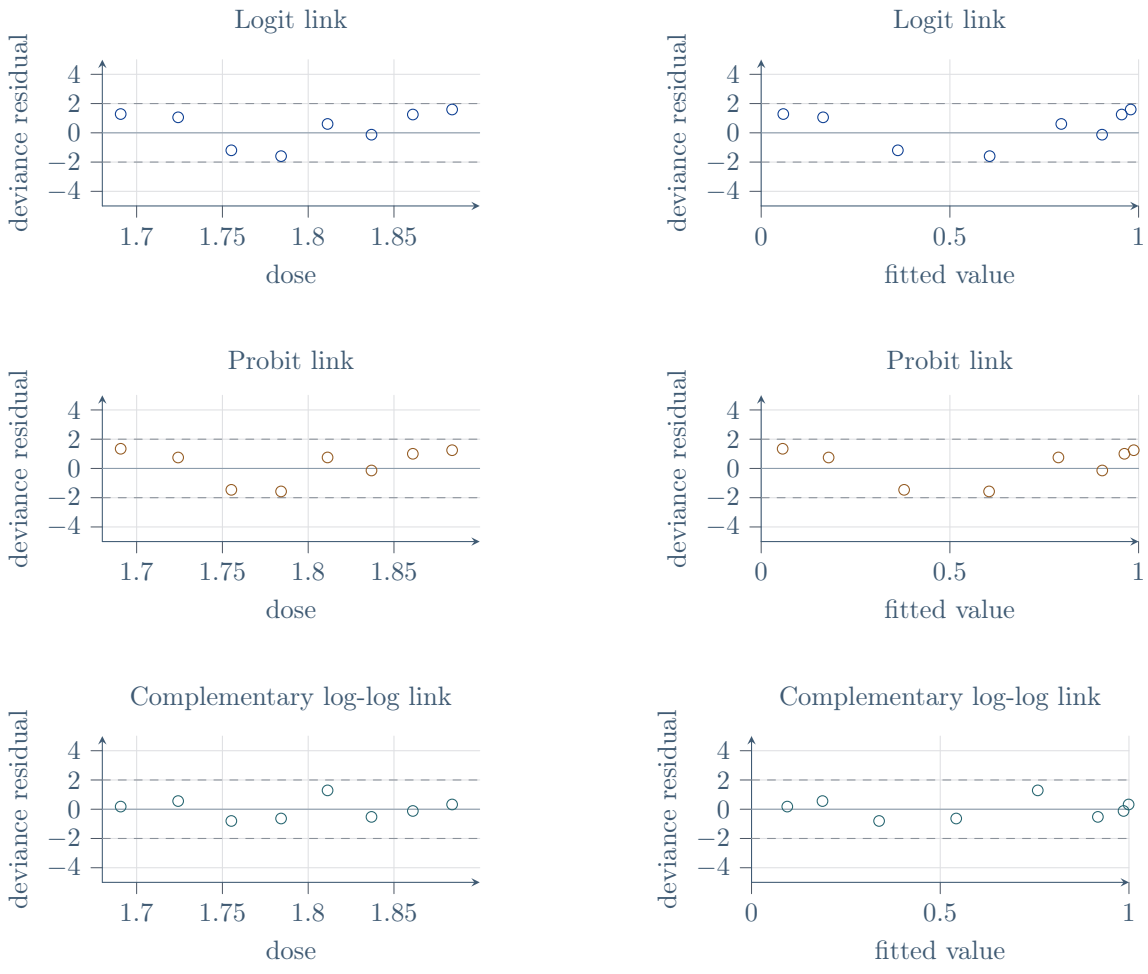

FIGURE 18

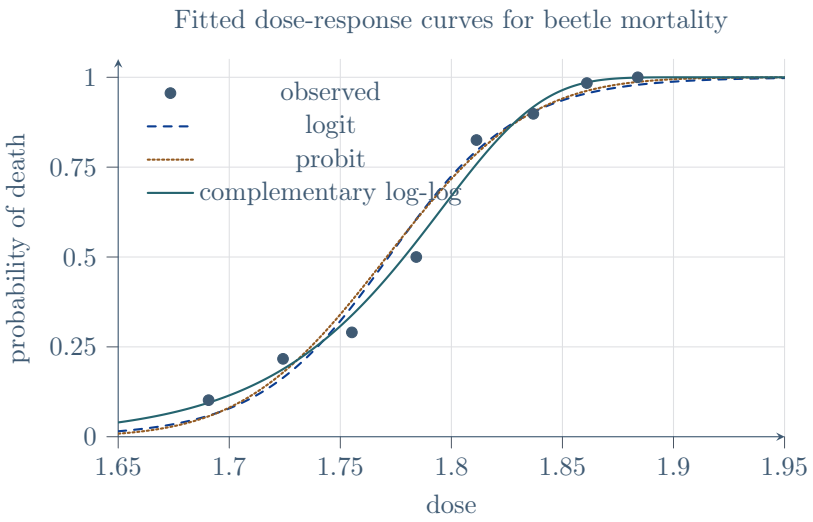

FIGURE 19

For the probit model,

$$\hat{\beta}_0 = -34.935266 \qquad \text{and} \qquad \hat{\beta}_1 = 19.727938.$$

Therefore the maximum likelihood estimate of the median lethal dose is

$$\hat{\delta} = \frac{-\hat{\beta}_0}{\hat{\beta}_1} = \frac{34.935}{19.728} = 1.77.$$

Exercise 9.4 Find $\delta_{0.25}$ and $\delta_{0.75}$, the twenty-fifth and seventy-fifth percentiles of the lethal dose for the probit beetle mortality model.

For a general p, solve

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

Figure 20 marks the standard normal quantiles used for $p = 0.25$ and $p = 0.75$.

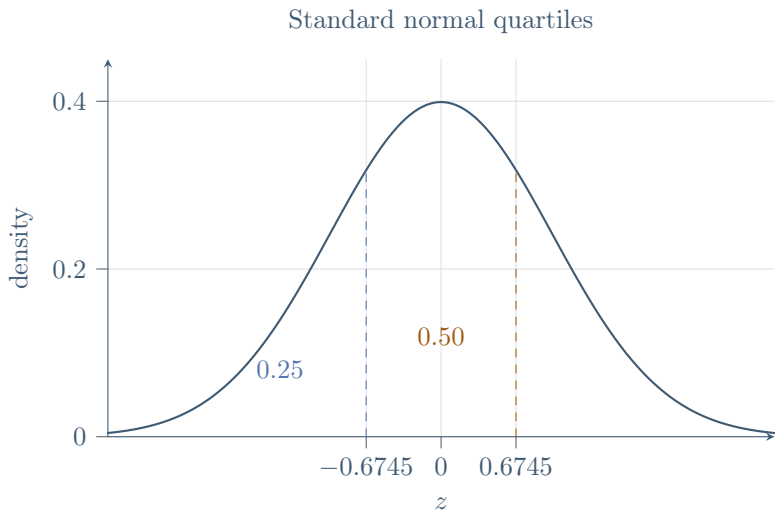

FIGURE 20

The needed standard normal quantiles are

$$\Phi^{-1}(0.25) = -0.6745$$
$$\Phi^{-1}(0.75) = 0.6745.$$

Therefore

$$\hat{\delta}_{0.25} = \frac{\Phi^{-1}(0.25) - \hat{\beta}_0}{\hat{\beta}_1} = \frac{-0.6745 + 34.9353}{19.7279} = 1.737,$$
$$\hat{\delta}_{0.75} = \frac{\Phi^{-1}(0.75) - \hat{\beta}_0}{\hat{\beta}_1} = \frac{0.6745 + 34.9353}{19.7279} = 1.805.$$

Lecture 10 — Thursday, February 04, 2016

3. Poisson and Log-Linear Regression

Poisson Generalized Linear Models

Recall that if $Y \sim \text{Poisson}(\mu)$, then

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

Thus the Poisson distribution is a member of the exponential family

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

with

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

Consequently,

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

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

Now suppose $Y_i \sim \text{Poisson}(\mu_i)$ independently for $i = 1, \ldots, n$. The response vector and mean vector are

$$\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^\top \qquad \text{and} \qquad \boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^\top.$$

The log-likelihood is

$$\ell(\mu) = \sum_{i=1}^{n} \left[y_i \log\mu_i - \mu_i - \log y_i!\right].$$

Using the canonical link, the log-linear regression model is

$$\begin{aligned} \log\mu_i &= \boldsymbol{x}_i^\top\boldsymbol{\beta} \\ \mu_i &= \exp\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}\right). \end{aligned}$$

The goal is to estimate and make inference about the regression coefficients $\boldsymbol{\beta}$. Substituting the model into the log-likelihood gives

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\left[y_i\boldsymbol{x}_i^\top\boldsymbol{\beta} - \exp\left(\boldsymbol{x}_i^\top\boldsymbol{\beta}\right) - \log y_i!\right] \\ &= \sum_{i=1}^{n}\left[y_i\sum_{j=0}^{p-1}x_{ij}\beta_j - \exp\left(\sum_{j=0}^{p-1}x_{ij}\beta_j\right) - \log y_i!\right].\end{aligned}$$

The jth component of the score vector is

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

The (j,k) element of the observed information matrix is

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

This expression does not depend on the observed responses, so the observed and expected information matrices coincide here. These formulas may also be obtained from the general exponential family score and information formulas.

For a likelihood ratio deviance test, let $\tilde{\mu}_i$ denote the maximum likelihood estimate under the saturated model and let $\hat{\mu}_i$ denote the maximum likelihood estimate under a p-dimensional constrained model. The null hypothesis is that the constrained p-dimensional model is adequate compared with the saturated n-dimensional model, while the alternative hypothesis is that the constrained model is not adequate.

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

asymptotically under H_0.

For nested Poisson regression models, with a reduced model having p parameters and a full model having q parameters, the likelihood ratio deviance statistic is

$$\Delta D = D_0 - D_A \dot{\sim} \chi^2_{q-p} \quad \text{under } H_0,$$

and the p-value is

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

For the Poisson distribution, the saturated estimate is $\tilde{\mu}_i = y_i$. Therefore

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

The convention $0\log(0/\hat{\mu}_i) = 0$ is used when $y_i = 0$. This is the residual deviance reported by R.

Since $\hat{\mu}_i = \exp\left(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}\right)$, the same statistic can be written as

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

The deviance residual is

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

where $d_i \geq 0$. Under an adequate model, these residuals should be roughly centred at zero and have a stable scale, but fitted residuals need not be independent or exactly standard normal.

Remark 10.1 The term

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

has the same $O_i \log(O_i/E_i)$ form that appears in the binomial deviance, with observed counts $O_i = y_i$ and expected counts $E_i = \hat{\mu}_i$.

Also,

$$2\sum_{i=1}^{n}(y_i - \hat{\mu}_i) = 0$$

whenever

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\mu}_i.$$

This equality holds when the intercept β_0 is included in the model $\log \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$.

Counting and Poisson Processes

A counting process $N(t)$ is a nondecreasing integer-valued function of time such that

$$N(0) = 0 \qquad \text{and} \qquad N(t) = \text{the number of events in } (0, t].$$

Example 10.2 Suppose events occur at times $2, 4, 5$, and 7. Then $N(t)$ is the step function shown in Figure 21.

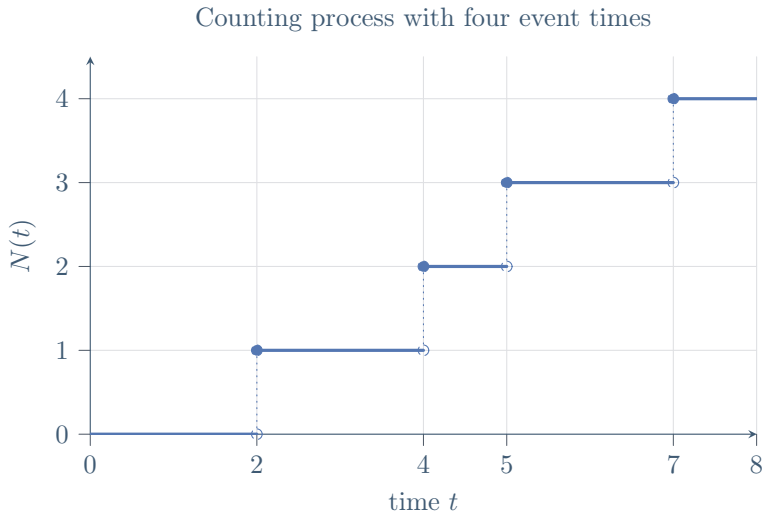

FIGURE 21

A counting process is a Poisson process if it satisfies the following two conditions.

1. It has independent increments: for every finite collection of pairwise disjoint time intervals, the corresponding event counts are mutually independent. In particular, if $s_1 < t_1 < s_2 < t_2$,

then $N(t_1) - N(s_1)$ is independent of $N(t_2) - N(s_2)$.

Figure 22 illustrates two disjoint time intervals whose event counts are independent.

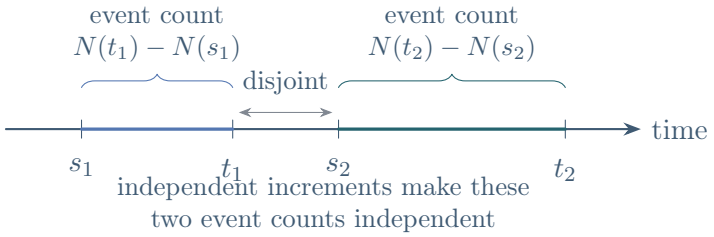

FIGURE 22

2. The number of events in $(0, t]$ has distribution

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

for $t \geq 0$ and $\lambda > 0$. Thus $N(t) \sim \text{Poisson}(\lambda t)$, where λ is the event rate and t is the duration of observation.

For a Poisson process,

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

Using the log link gives

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

Equivalently,

$$\mu_i(t_i) = t_i \exp\left(\boldsymbol{x}_i^\top \boldsymbol{\beta}\right) = t_i \lambda_i.$$

The regression equation is therefore

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

Remark 10.3 The term $\log t_i$ is an **offset term**. It explains variation in event counts due to different observation durations, but it enters deterministically and has no estimated regression coefficient in front of it.

Ship Damage Data

McCullagh and Nelder discuss a data set recording the number of damage incidents in cargo ships. Damage occurs in the forward section of the vessels, and the purpose of the analysis is to identify risk factors for the accident rate.

1. The explanatory variables are ship type, with five levels A through E; year of construction, with four levels 1960–1964, 1965–1969, 1970–1974, and 1975–1979; and year of operation, with two levels 1960–1974 and 1975–1979.
2. The variable months is the total duration of operation, denoted t_i.
3. The response y_i is the total number of damage incidents over t_i months of operation, so y_i is a count.

The data are

type	cyr	oyr	months	y	type	cyr	oyr	months	y
1	1	1	127	0	3	2	2	676	1
1	1	2	63	0	3	3	1	783	6
1	2	1	1095	3	3	3	2	1948	2
1	2	2	1095	4	3	4	2	274	1
1	3	1	1512	6	4	1	1	251	0
1	3	2	3353	18	4	1	2	105	0
1	4	2	2244	11	4	2	1	288	0
2	1	1	44882	39	4	2	2	192	0
2	1	2	17176	29	4	3	1	349	2
2	2	1	28609	58	4	3	2	1208	11
2	2	2	20370	53	4	4	2	2051	4
2	3	1	7064	12	5	1	1	45	0
2	3	2	13099	44	5	2	1	789	7
2	4	2	7117	18	5	2	2	437	7
3	1	1	1179	1	5	3	1	1157	5
3	1	2	552	1	5	3	2	2161	12
3	2	1	781	0	5	4	2	542	1

The first task is to decide whether the total duration of operation should enter as an offset.

Model 1 includes the main effects and the offset log(months):

$$\texttt{y} \sim \texttt{typeft + cyrft + oyrft + offset(log(months))}.$$

Model 2 includes the same main effects but treats log(months) as an ordinary covariate:

$$\texttt{y} \sim \texttt{typeft + cyrft + oyrft + log(months)}.$$

The fitted summaries are

Model	Residual deviance	Residual degrees of freedom	Coefficient of log(months)
1	38.6951	25	1, fixed as an offset
2	37.8043	24	0.9027 (0.1018)

where the number in parentheses is the standard error of the estimated coefficient in Model 2.

For Model 1, the main effect coefficient estimates are

Coefficient	Estimate	standard error
β_0	-6.4059	0.2174
β_1 (type B)	-0.5433	0.1776
β_2 (type C)	-0.6874	0.3290
β_3 (type D)	-0.0760	0.2906
β_4 (type E)	0.3256	0.2359
β_5 (construction 1965–1969)	0.6971	0.1496
β_6 (construction 1970–1974)	0.8184	0.1698
β_7 (construction 1975–1979)	0.4534	0.2332
β_8 (operation 1975–1979)	0.3845	0.1183

The offset formulation corresponds to testing

$$H_0 : \beta_9 = 1 \qquad \text{against} \qquad H_a : \beta_9 \neq 1$$

in Model 2. This is not the default test displayed in the R coefficient table, which tests $H_0 : \beta_9 = 0$.

$$z = \frac{\hat{\beta}_9 - 1}{\text{standard error}(\hat{\beta}_9)} = \frac{0.9027 - 1}{0.1018} = -0.9558,$$
$$p\text{-value} = 2\mathbb{P}(Z > 0.9558) = 0.34, \quad Z \sim \mathcal{N}(0, 1).$$

We do not reject H_0. The data are not inconsistent with the unit exposure coefficient required by a time-homogeneous rate model, so the more parsimonious offset formulation is retained. We use Model 1 as the starting point for model selection.

Lecture 11 — Tuesday, February 09, 2016

Ship Damage Model Selection

The next question is whether a smaller model can be used in place of Model 1.

$$\log \mu_i(t_i) = \log \lambda_i + \log t_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \underbrace{\log t_i}_{\text{offset}}.$$

Example 11.1 Model 1 has residual deviance $D_1 = 38.695$ on 25 degrees of freedom and has linear predictor

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

where t_i is measured in months.

There are two goals: select the model with the best fit and then interpret it. The following reduced models are nested in Model 1, so deviance tests can be used.

Model 1	type, construction year, operation year,
Model 3a	construction year, operation year,
Model 3b	type, operation year,
Model 3c	type, construction year.

Model 3a removes ship type. The hypotheses are

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

against the alternative that at least one of these coefficients is nonzero. The deviance test statistic is

$$\Delta D = D_{3a} - D_1 = 62.365 - 38.695 = 23.67,$$

which is compared with χ^2_4 under H_0. The p-value is 9.30×10^{-5}, so we reject the null hypothesis that ship type is unimportant. Ship type should not be removed from Model 1.

Model 3b removes construction year. The hypotheses are

$$H_0 : \beta_5 = \beta_6 = \beta_7 = 0$$

against the alternative that at least one of these coefficients is nonzero. The deviance test statistic is

$$\Delta D = D_{3b} - D_1 = 70.103 - 38.695 = 31.408,$$

which is compared with χ^2_3. The p-value is 6.97×10^{-7}, so construction year should not be removed.

Model 3c removes operation year. The hypotheses are

$$H_0 : \beta_8 = 0 \qquad \text{against} \qquad H_a : \beta_8 \neq 0.$$

The deviance test statistic is

$$\Delta D = D_{3c} - D_1 = 49.355 - 38.695 = 10.660,$$

which is compared with χ^2_1. The p-value is 0.00109, so operation year should not be removed. We therefore retain all three main effects.

It remains to check whether interaction terms should be added to Model 1.

Model 4 adds the interaction between type and construction year. This adds $4 \cdot 3 = 12$ regression parameters.

The null hypothesis is that the interaction between type and construction year is unimportant:

$$H_0 : \beta_9 = \beta_{10} = \cdots = \beta_{20} = 0.$$

The alternative is that at least one of these interaction coefficients is nonzero.

$$\Delta D = D_1 - D_4 = 38.695 - 14.587 = 24.108,$$

which is compared with χ^2_{12}. The p-value is 0.0197, so the deviance test rejects the main effects

model in favour of Model 4. However, the fitted Model 4 has very large standard errors and extreme coefficient estimates, indicating overparameterization due to sparse cells; for instance, ship type 4 has no events in construction year groups 1 or 2. We therefore do not select Model 4.

Model 5 adds the interaction between type and operation year. Its deviance test p-value is 0.293632, so we do not reject the main effects model. Model 6 adds the interaction between construction year and operation year. Its deviance test p-value is 0.409127, so again we do not reject the main effects model.

We therefore select Model 1 as the most appropriate compromise among the candidates considered. Model 4 attains a smaller deviance, but its unstable estimates make that nominal improvement unpersuasive. The residual plot for Model 1 is also acceptable.

We now interpret Model 1.

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

Relative Rate Inference

The model is based on

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

where λ_i is the expected number of events per unit time. Ratios of the rates λ_i therefore compare accident rates.

First, estimate the relative rate of accidents for ships of type A versus ships of type C, controlling for construction period and operation period. The corresponding rows of the design matrix are

type	construction	operation	$\boldsymbol{x}_i^\top \boldsymbol{\beta} = \log \lambda_i$
A	-	-	$\beta_0 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$
C	-	-	$\beta_0 + \beta_2 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$

Subtracting the second row from the first gives

$$\log \lambda_A - \log \lambda_C = -\beta_2,$$
$$\log \left(\frac{\lambda_A}{\lambda_C} \right) = -\beta_2,$$

$$\frac{\lambda_A}{\lambda_C} = \exp(-\beta_2).$$

Thus

$$\widehat{\lambda_A/\lambda_C} = \exp\Big(-\hat{\beta}_2\Big) = \exp(0.6874) = 1.990.$$

For ships constructed in the same period and operated in the same period, the estimated accident rate for ships of type A is 1.99 times the accident rate for ships of type C.

A 95% confidence interval for this relative rate is

$$\begin{aligned}\exp\Big(-\hat{\beta}_2 \pm 1.96 \text{ standard error}(\hat{\beta}_2)\Big) &= \exp(0.6874 \pm 1.96(0.3290)) \\ &= (e^{0.0426}, e^{1.332}) \\ &= (1.04, 3.79).\end{aligned}$$

Testing equality of the two accident rates is equivalent to testing

$$H_0 : \beta_2 = 0 \qquad \text{against} \qquad H_a : \beta_2 \neq 0.$$

Using $U \sim \mathcal{N}(0,1)$,

$$p = 2\mathbb{P}\left(U > \frac{|\hat{\beta}_2|}{\text{standard error}(\hat{\beta}_2)}\right) = 2\mathbb{P}\left(U > \frac{|-0.6874|}{0.3290}\right) = 2\mathbb{P}(U > 2.089) = 0.0367.$$

We reject H_0, so the data provide evidence that the accident rates differ between ships of type A and ships of type C after adjusting for construction period and operation period.

Second, estimate the relative rate of accidents for ships of type E versus ships of type B, again controlling for construction period and operation period. The log relative rate is

$$\begin{aligned}\log(\lambda_E/\lambda_B) &= \beta_4 - \beta_1 \\ \frac{\lambda_E}{\lambda_B} &= \exp(\beta_4 - \beta_1).\end{aligned}$$

The log relative rate is now a linear combination of two regression parameters. Let

$$\boldsymbol{c} = (0, -1, 0, 0, 1, 0, 0, 0, 0)^\top.$$

Then $\boldsymbol{c}^\top \hat{\boldsymbol{\beta}} = \hat{\beta}_4 - \hat{\beta}_1$, and

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

Equivalently, using the covariance entries reported by R,

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

Remark 11.2 If a correlation matrix rather than a covariance matrix is reported, use

$$\text{cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

to recover the covariance.

$$\hat{\beta}_4 - \hat{\beta}_1 = 0.3256 - (-0.5433) = 0.8689,$$

so

$$\widehat{\lambda_E/\lambda_B} = \exp\Big(\hat{\beta}_4 - \hat{\beta}_1\Big) = \exp(0.8689) = 2.38.$$

A 95% confidence interval is

$$\exp(0.8689 \pm 1.96(0.198)) = (e^{0.481}, e^{1.257}) = (1.62, 3.51).$$

Thus, for ships constructed and operated in the same periods, ships of type E have an estimated accident rate 2.38 times that of ships of type B, with 95% confidence interval $(1.62, 3.51)$.

The null hypothesis of no difference between types E and B is

$$H_0 : \beta_4 - \beta_1 = 0 \qquad \text{against} \qquad H_a : \beta_4 - \beta_1 \neq 0.$$

The Wald p-value is

$$2\mathbb{P}\left(U > \frac{|0.8689|}{0.198}\right) < 0.001,$$

so we reject H_0.

Third, 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. Here type B is type 2, 1970 is in construction year

group 3, and 1975–1979 is in operation year group 2.

The model gives

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

For this group,

$$\boldsymbol{x}_i = (1, 1, 0, 0, 0, 0, 1, 0, 1)^\top,$$

so

$$\log \hat{\lambda}_i = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_6 + \hat{\beta}_8 = -5.7463,$$

with 95% confidence interval $(-5.979, -5.513)$.

$$t_i = 10 \text{ ships} \times 5 \text{ years} \times 12 \text{ months per year} = 600 \text{ ship-months}.$$

Therefore

$$\hat{\mu}_i = \hat{\lambda}_i t_i = \exp(-5.7463)(600) = 1.92.$$

Transforming the confidence interval for $\log \lambda_i$ gives the 95% confidence interval

$$(600e^{-5.9789}, 600e^{-5.5138}) = (1.52, 2.42).$$

The estimated mean number of accidents is 1.92, with 95% confidence interval $(1.52, 2.42)$. This is an interval for the expected count, not a prediction interval for the realized future count.

Lecture 12 — Tuesday, February 23, 2016

Poisson Approximation to Binomial Data

Poisson generalized linear models can also be used to approximate binomial data when the population size is large and the event probability is small. Suppose

$$Y \sim \text{Bin}(m, \pi) \qquad \text{and} \qquad \mathbb{E}[Y] = m\pi.$$

Set $\mu = m\pi$, so that $\pi = \mu/m$. Then

$$\begin{aligned} f(y) &= \binom{m}{y}\pi^y(1-\pi)^{m-y} \\ &= \frac{m(m-1)\dots(m-y+1)}{y!}\left(\frac{\mu}{m}\right)^y\left(1-\frac{\mu}{m}\right)^{m-y} \\ &= \frac{m(m-1)\dots(m-(y-1))}{m\cdot m\cdot\dots\cdot m}\frac{\mu^y}{y!}\left(1-\frac{\mu}{m}\right)^m\left(1-\frac{\mu}{m}\right)^{-y} \end{aligned}$$

Remark 12.1 Recall $\lim_{n\to\infty}(1+a/n)^n = e^a$.

As $m \to \infty$ with $\mu = m\pi$ fixed,

1. $\{m(m-1)\cdots(m-y+1)\}/m^y \to 1$.
2. $(1-\mu/m)^m \to e^{-\mu}$.
3. $(1-\mu/m)^{-y} \to 1$.

Thus

$$f(y) \to \frac{\mu^y e^{-\mu}}{y!},$$

so Y is approximately Poisson($\mu = m\pi$).

Using a Poisson generalized linear model with log link,

$$\log\mu = \log(m\pi) = \log\pi + \log m = \boldsymbol{x}^\top\boldsymbol{\beta} + \underbrace{\log m}_{\text{offset}}.$$

The regression coefficients have a log relative rate interpretation, which is often easier to interpret than the log odds ratio interpretation from logistic regression.

Time Nonhomogeneous Poisson Processes

A time-homogeneous Poisson process has

$$\mathbb{E}[N_i(t)] = \mu_i(t) = \lambda_i t$$

with constant event rate λ_i. In a Poisson process that is nonhomogeneous in time, the rate varies with time. There are many possible models for $\lambda(t)$, including piecewise constant functions,

piecewise linear functions, splines, and polynomials, as in Figure 23.

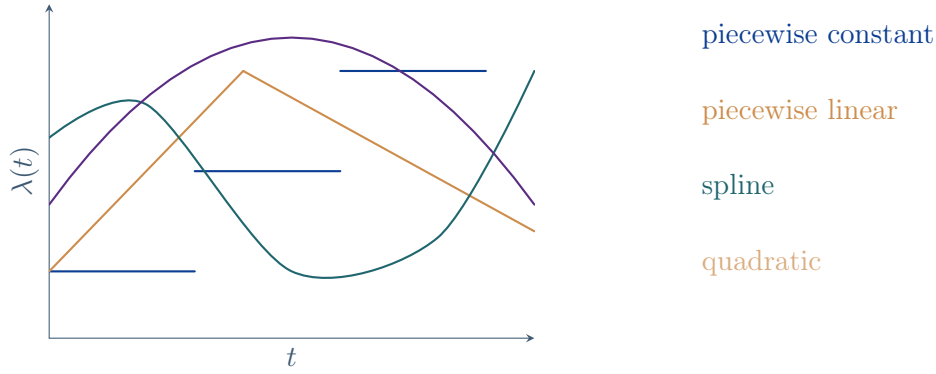

Figure 23

Example 12.2 (Rat tumor data) Gail, Santner, and Brown (1980) studied the development of mammary tumors in 48 rats over 122 days. There were 23 rats assigned to a treatment group and 25 assigned to a control group. Multiple tumors could develop in the same rat, and the day of each tumor was recorded. To fit a Poisson model, follow-up time is split into four intervals, and each rat contributes one line of data per interval. The variables are the tumor count in the interval, the interval length, and the treatment indicator. A basic Poisson GLM treats the interval counts as independent; because several rows come from the same rat, we should regard that as a working assumption and check for extra variation or within-rat association.

After the event time data are split into four intervals, the first five treated rats give rows of the following form:

id	interval	count	length	treatment
1	1	0	30	1
1	2	0	30	1
1	3	0	30	1
1	4	1	32	1
2	1	0	30	1
2	2	0	30	1
2	3	0	30	1
2	4	0	32	1
3	1	1	30	1
3	2	0	30	1
3	3	1	30	1
3	4	0	32	1
4	1	0	30	1
4	2	0	30	1
4	3	0	30	1
4	4	1	32	1
5	1	0	30	1
5	2	0	30	1
5	3	3	30	1
5	4	1	32	1

First fit a piecewise constant rate model for the control rats only. Figure 24 shows the corresponding structure of the log rate:

$$\log \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \log t_i,$$

where x_{i1}, x_{i2}, and x_{i3} indicate intervals 2, 3, and 4, and $\log t_i$ is the offset for interval length.

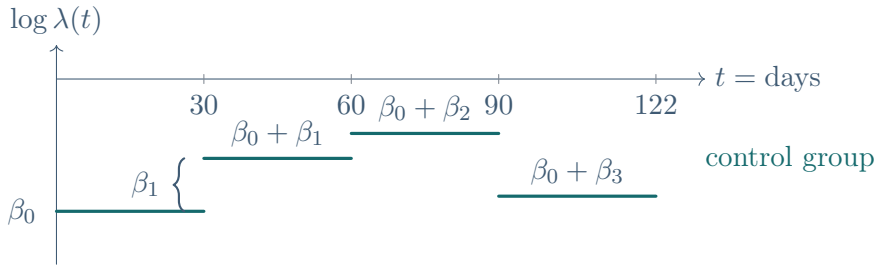

FIGURE 24

The fitted control-only coefficient table is

Term	Estimate	Standard error	z	p-value
Intercept	−3.09371	0.17137	−18.0530	$< 2 \times 10^{-16}$
interval 2	0.16254	0.23284	0.6981	0.4851
interval 3	0.30228	0.22593	1.3380	0.1809
interval 4	−0.15690	0.24794	−0.6328	0.5268

The residual deviance is 163.308 on 96 degrees of freedom.

The coefficient β_1 compares interval 2 with interval 1 among control rats:

Interval	$\log \mu_i(t)$
2	$\beta_0 + \beta_1 + \log t$
1	$\beta_0 + \log t$

$$\log \lambda_{\text{interval 2}} - \log \lambda_{\text{interval 1}} = \beta_1$$
$$\frac{\lambda_2}{\lambda_1} = e^{\beta_1}.$$

Thus e^{β_1} is the relative rate of tumors in interval 2 versus interval 1 for control rats. In the fitted control-only model, $e^{\hat{\beta}_1} = e^{0.16254} = 1.176$, and none of the interval coefficients is statistically significant.

Now fit both control and treatment groups with a piecewise constant baseline rate:

$$\log \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \log t_i,$$

where x_{i4} indicates the treatment group. Figure 25 shows the common piecewise baseline with a shift for the treatment group.

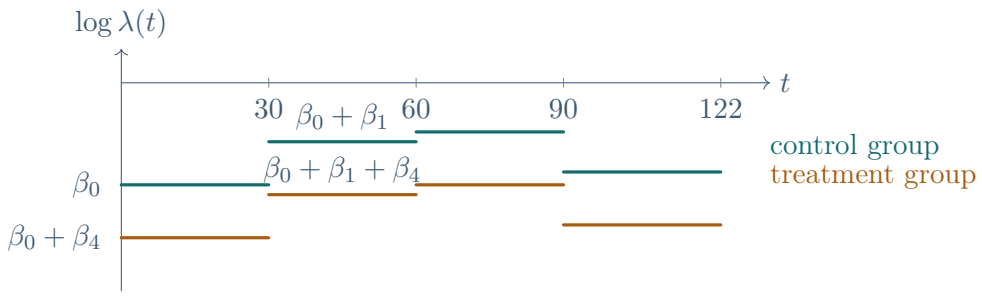

FIGURE 25

The fitted piecewise model for both groups has coefficient summary

Term	Estimate	Standard error	z	p-value
Intercept	−3.088179	0.150548	−20.5130	$< 2.2 \times 10^{-16}$
interval 2	0.171860	0.195454	0.8793	0.3792
interval 3	0.206350	0.193905	1.0642	0.2872
interval 4	−0.064533	0.203686	−0.3168	0.7514
treatment	−0.822988	0.151130	−5.4456	5.164×10^{-8}

The residual deviance is 266.323 on 187 degrees of freedom.

In this working-independence model,

$$e^{\beta_4} = \frac{\lambda_{\text{treatment}}}{\lambda_{\text{control}}}$$

is the relative tumor rate for a treated rat versus a control rat in the same time interval. The fitted value is $e^{\hat{\beta}_4} = e^{-0.8230} = 0.44$. Controlling for follow-up interval, the fitted rate of tumor development in treated rats is 0.44 times that of control rats. The interval coefficients remain statistically insignificant. Because observations from the same rat may be associated, the model-based standard errors and tests require the working Poisson-process assumption; robust, GEE, or random-effects methods are preferable if that assumption is doubtful.

Since the interval coefficients are not significant, consider the time-homogeneous model

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

This model has residual deviance 269.060 on 190 degrees of freedom, while the piecewise model has residual deviance 266.323 on 187 degrees of freedom. The likelihood ratio statistic for

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

is

$$\Delta D = 269.060 - 266.323 = 2.737,$$

which is compared with χ^2_3. The p-value is 0.4340, so we do not reject H_0. The time-homogeneous treatment model is adequate for these data. Figure 26 shows the resulting parallel model with a constant rate.

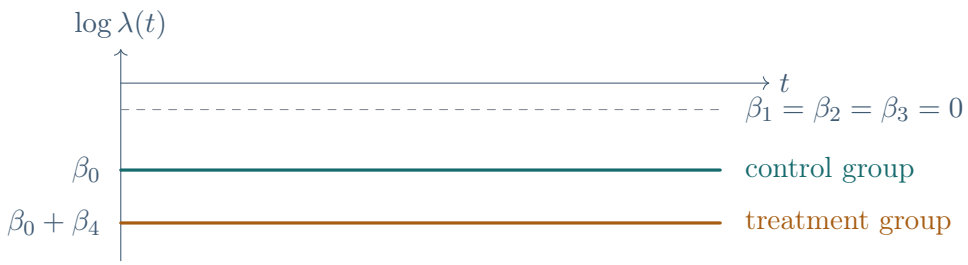

FIGURE 26

To check whether the treatment effect changes over time, fit the interaction model

$$\log \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i1} x_{i4} + \beta_6 x_{i2} x_{i4} + \beta_7 x_{i3} x_{i4} + \log t_i.$$

The model without the interaction is nested in this model by setting $\beta_5 = \beta_6 = \beta_7 = 0$. Figure 27 shows how the treatment effect can vary by time interval in the interaction model.

In the interaction model, the relative tumor rate in interval 2 versus interval 1 is e^{β_1} for the control group and $e^{\beta_1+\beta_5}$ for the treatment group. The deviance test statistic for

$$H_0 : \beta_5 = \beta_6 = \beta_7 = 0$$

is

$$\Delta D = 266.323 - 263.917 = 2.406,$$

which is compared with χ^2_3. The p-value is 0.4926, so there is no evidence that the treatment effect

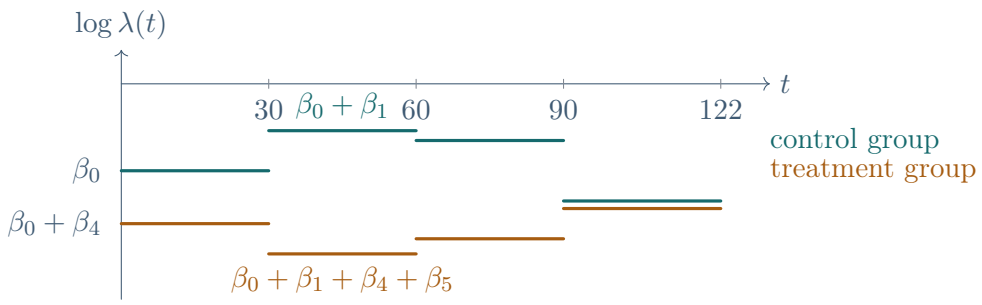

FIGURE 27

varies across the time intervals.

The interaction model has coefficient summary

Term	Estimate	Standard error	z	p-value
Intercept	−3.093712	0.171368	−18.0530	$< 2 \times 10^{-16}$
interval 2	0.162536	0.232840	0.6981	0.48514
interval 3	0.302284	0.225926	1.3380	0.18090
interval 4	−0.156902	0.247935	−0.6328	0.52684
treatment	−0.803850	0.316133	−2.5428	0.01100
interval 2 by treatment	0.031556	0.428219	0.0737	0.94126
interval 3 by treatment	−0.376186	0.443551	−0.8481	0.39637
interval 4 by treatment	0.286455	0.436610	0.6561	0.51177

The residual deviance is 263.917 on 184 degrees of freedom, and $\hat{\beta}_4 = -0.8038$ is close to the treatment coefficient in the model without the interaction.

For the rat tumor example, the estimated relative rate of tumor development for treated rats versus control rats is approximately 0.44. The treatment appears beneficial, since treated rats have fewer tumors. The piecewise constant rate function is not necessary here, and the selected model is

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

We should still examine residuals. The fitted model can also be used for prediction; for example, for a control rat observed for 70 days,

$$\log \hat{\mu} = \hat{\beta}_0 + \log(70).$$

Lecture 13 — Thursday, February 25, 2016

4. Contingency Tables and Log-Linear Models

Contingency Tables

Example 13.1 Let factor V be whether a birthday occurs on an odd or even day, and let factor W be the season of the birthday. This gives a 2×4 contingency table.

	Winter	Spring	Summer	Fall	Total
E	10	15	30	15	70
O					80

For an $I \times J$ contingency table, the observation in cell (i, j) is

$$Y_{ij} = y_{ij}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J.$$

The cell count y_{ij} records the number of individuals at level i of factor V and level j of factor W. The row totals, column totals, and grand total are

$$y_{i\cdot} = \sum_{j=1}^{J} y_{ij}, \qquad y_{\cdot j} = \sum_{i=1}^{I} y_{ij}, \qquad \text{and} \qquad y_{\cdot\cdot} = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij}.$$

The basic Poisson assumption is

$$Y_{ij} \sim \text{Poisson}(\mu_{ij})$$

independently over all cells. Then

$$Y_{\cdot\cdot} \sim \text{Poisson}(\mu_{\cdot\cdot}) \qquad \text{and} \qquad \mu_{\cdot\cdot} = \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}.$$

If the sample size is fixed by design, then we condition on $Y_{\cdot\cdot} = y_{\cdot\cdot}$.

Example 13.2 (Breast self-examination) Senie et al. (1981) investigated the relationship between age and frequency of breast self-examination in a sample of women.

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

Let A denote the event that $Y_{ij} = y_{ij}$ for all i, j. Then

$$\begin{aligned}
\mathbb{P}\left(A \mid Y_{\cdot\cdot} = y_{\cdot\cdot}\right) &= \frac{\mathbb{P}\left(A,\ Y_{\cdot\cdot} = y_{\cdot\cdot}\right)}{\mathbb{P}(Y_{\cdot\cdot} = y_{\cdot\cdot})} \\
&= \frac{\prod_{i=1}^{I}\prod_{j=1}^{J} \frac{\mu_{ij}^{y_{ij}} \exp(-\mu_{ij})}{y_{ij}!}}{\frac{\mu_{\cdot\cdot}^{y_{\cdot\cdot}} \exp(-\mu_{\cdot\cdot})}{y_{\cdot\cdot}!}} \\
&= \left(\frac{y_{\cdot\cdot}!}{\prod_{i=1}^{I}\prod_{j=1}^{J} y_{ij}!}\right)\left(\frac{\prod_{i=1}^{I}\prod_{j=1}^{J} \mu_{ij}^{y_{ij}}}{\mu_{\cdot\cdot}^{y_{\cdot\cdot}}}\right)\left(\frac{\exp\left(-\sum_{i=1}^{I}\sum_{j=1}^{J}\mu_{ij}\right)}{\exp(-\mu_{\cdot\cdot})}\right) \\
&= \frac{y_{\cdot\cdot}!}{\prod_{i=1}^{I}\prod_{j=1}^{J} y_{ij}!}\prod_{i=1}^{I}\prod_{j=1}^{J}\left(\frac{\mu_{ij}}{\mu_{\cdot\cdot}}\right)^{y_{ij}}.
\end{aligned}$$

Notation 13.3 Note that

$$\prod_{i=1}^{I}\prod_{j=1}^{J}\mu_{\cdot\cdot}^{y_{ij}} = \mu_{\cdot\cdot}^{\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}} = \mu_{\cdot\cdot}^{y_{\cdot\cdot}}.$$

This is the multinomial distribution with cell probabilities

$$\pi_{ij} = \frac{\mu_{ij}}{\mu_{\cdot\cdot}}.$$

Definition 13.4 The **standard multinomial distribution** has probability mass function

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

where

$$\sum_{i=1}^{k} x_i = n \qquad \text{and} \qquad \sum_{i=1}^{k} \pi_i = 1.$$

Here $x_1, \ldots, x_k$ are nonnegative integers, $\pi_i \geq 0$, and n is a nonnegative integer. The term $n!/(x_1! \cdots x_k!)$ is the multinomial coefficient.

Applying this to the contingency table gives

$$\mathbb{P}\left(Y_{ij} = y_{ij} \text{ for all } i, j \mid Y_{\cdot\cdot} = y_{\cdot\cdot}\right) = \frac{y_{\cdot\cdot}!}{\prod_{i=1}^{I} \prod_{j=1}^{J} y_{ij}!} \prod_{i=1}^{I} \prod_{j=1}^{J} \pi_{ij}^{y_{ij}},$$

where

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\mu_{ij}}{\mu_{\cdot\cdot}} = 1.$$

Ignoring constants, the multinomial log-likelihood is

$$\ell(\pi) = \sum_{i=1}^{I} \sum_{j=1}^{J} y_{ij} \log \pi_{ij} \qquad \text{and} \qquad \sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} = 1.$$

The null hypothesis is that factor V and factor W are independent:

$$H_0 : \pi_{ij} = \pi_{i\cdot} \pi_{\cdot j} \quad \text{for all } i, j.$$

Equivalently,

$$\mathbb{P}(\text{level } i \text{ of } V \text{ and level } j \text{ of } W) = \mathbb{P}(\text{level } i \text{ of } V)\, \mathbb{P}(\text{level } j \text{ of } W).$$

$$\pi_{i\cdot} = \sum_{j=1}^{J} \pi_{ij} = \mathbb{P}(\text{level } i \text{ of } V),$$

$$\pi_{\cdot j} = \sum_{i=1}^{I} \pi_{ij} = \mathbb{P}(\text{level } j \text{ of } W).$$

The alternative hypothesis is that $\pi_{ij} \neq \pi_{i\cdot}\pi_{\cdot j}$ for at least one cell (i, j).

We derive the likelihood ratio deviance test for independence. Under H_0,

$$\hat{\pi}_{i\cdot} = \frac{y_{i\cdot}}{y_{\cdot\cdot}} \qquad \text{and} \qquad \hat{\pi}_{\cdot j} = \frac{y_{\cdot j}}{y_{\cdot\cdot}},$$

so

$$\hat{\pi}_{ij} = \hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j} = \frac{y_{i\cdot}y_{\cdot j}}{y_{\cdot\cdot}^2}.$$

$$\ell(\hat{\pi}) = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \log\left(\frac{y_{i\cdot}y_{\cdot j}}{y_{\cdot\cdot}^2}\right).$$

Under the saturated alternative,

$$\tilde{\pi}_{ij} = \frac{y_{ij}}{y_{\cdot\cdot}},$$

and

$$\ell(\tilde{\pi}) = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \log\left(\frac{y_{ij}}{y_{\cdot\cdot}}\right).$$

The likelihood ratio deviance statistic is

$$\begin{aligned} D &= 2\left(\ell(\tilde{\pi}) - \ell(\hat{\pi})\right) \\ &= 2\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \log\left(\frac{y_{ij}}{y_{i\cdot}y_{\cdot j}/y_{\cdot\cdot}}\right) \\ &= 2\sum_{i=1}^{I}\sum_{j=1}^{J} O_{ij} \log\left(\frac{O_{ij}}{E_{ij}}\right), \end{aligned}$$

where $O_{ij} = y_{ij}$ and

$$\begin{aligned} E_{ij} &= \hat{\pi}_{ij}y_{\cdot\cdot} \\ &= \hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j}y_{\cdot\cdot} \end{aligned}$$

$$= \frac{y_{i\cdot}}{y_{\cdot\cdot}} \frac{y_{\cdot j}}{y_{\cdot\cdot}} y_{\cdot\cdot}$$
$$= \frac{y_{i\cdot} y_{\cdot j}}{y_{\cdot\cdot}}.$$

Thus E_{ij} is the expected cell count under independence. Under H_0,

$$D \dot{\sim} \chi^2_{(I-1)(J-1)}$$

asymptotically.

Example 13.5 For the breast self-examination table,

$$\hat{\mu}_{11} = E_{11} = \frac{y_{1\cdot} y_{\cdot 1}}{y_{\cdot\cdot}} = \frac{232 \cdot 350}{1216} = 66.78.$$

The likelihood ratio deviance statistic is $D = 25.19$. Since the table is 3×3, the reference distribution is χ^2_4, and

$$\mathbb{P}(\chi^2_4 > 25.19) < 0.001.$$

We reject the null hypothesis of independence between age and frequency of breast self-examination.

Now suppose that the row totals are fixed by design. In this case we condition on

$$Y_{i\cdot} = y_{i\cdot}, \quad i = 1, \ldots, I.$$

$$\mathbb{P}\left(Y_{ij} = y_{ij} \text{ for all } i, j \mid Y_{i\cdot} = y_{i\cdot} \text{ for all } i\right) = \prod_{i=1}^{I} \left[\frac{y_{i\cdot}!}{\prod_{j=1}^{J} y_{ij}!} \prod_{j=1}^{J} \left(\frac{\mu_{ij}}{\mu_{i\cdot}} \right)^{y_{ij}} \right].$$

This is called the **product multinomial distribution**. There is one multinomial distribution for each row, with

$$\pi_{ij} = \frac{\mu_{ij}}{\mu_{i\cdot}} \qquad \text{and} \qquad \sum_{j=1}^{J} \pi_{ij} = 1 \quad \text{for each } i.$$

Here π_{ij} is the conditional probability of being at level j of factor W, given level i of factor V. This differs from the multinomial case above, where $\pi_{ij} = \mu_{ij}/\mu_{\cdot\cdot}$.

Ignoring constants, the log-likelihood is

$$\ell(\pi) = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \log \pi_{ij} \qquad \text{and} \qquad \sum_{j=1}^{J} \pi_{ij} = 1 \quad \text{for each } i.$$

To test independence in this setting, the null hypothesis is

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

That is, the probability of being at level j of factor W is the same across all levels i of factor V. The alternative is that at least one row has a different conditional distribution.

Under H_0,

$$\hat{\pi}_{ij} = \hat{\pi}_j = \frac{y_{\cdot j}}{y_{\cdot\cdot}},$$

the overall sample proportion in column j. Under the saturated alternative,

$$\tilde{\pi}_{ij} = \frac{y_{ij}}{y_{i\cdot}},$$

the row-specific proportion in column j.

The likelihood ratio deviance statistic is the same as in the multinomial sampling scheme:

$$D = 2\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \log\left(\frac{y_{ij}}{y_{i\cdot}y_{\cdot j}/y_{\cdot\cdot}}\right).$$

Under H_0, D is approximately $\chi^2_{(I-1)(J-1)}$.

Example 13.6 (Fixed row totals in the breast self-examination study) Suppose instead that the investigators had sampled 400 women from each age group and then classified each woman by frequency of breast self-examination. The row totals are then fixed by design:

Age	Monthly	Occasionally	Never	Total
< 45	157	155	88	400
45–59	119	158	123	400
≥ 60	91	165	144	400
Total	367	478	355	1200

The unconstrained estimates are the row-specific proportions $\tilde{\pi}_{ij} = y_{ij}/y_{i\cdot}$:

Age	Monthly	Occasionally	Never
< 45	39.25%	38.75%	22.00%
45–59	29.75%	39.50%	30.75%
≥ 60	22.75%	41.25%	36.00%

Under the null hypothesis of independence, the common row probabilities are

$$\hat{\pi}_{ij} = \hat{\pi}_j = \frac{y_{\cdot j}}{y_{\cdot\cdot}},$$

so the constrained percentages are 30.58%, 39.83%, and 29.58% in each row. The corresponding expected counts are

Age	Monthly	Occasionally	Never	Total
< 45	122.33	159.33	118.33	400
45–59	122.33	159.33	118.33	400
≥ 60	122.33	159.33	118.33	400
Total	367	478	355	1200

The likelihood ratio deviance statistic is

$$D = 2\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \log\left(\frac{y_{ij}}{y_{i\cdot}y_{\cdot j}/y_{\cdot\cdot}}\right) = 32.1701.$$

Comparing this value with χ^2_4 gives $p < 0.001$, so the null hypothesis that age and frequency of breast self-examination are independent is again rejected.

Lecture 14 — Tuesday, March 01, 2016

Two-Way Contingency Tables

For a two-way contingency table, continue to assume

$$Y_{ij} \sim \text{Poisson}(\mu_{ij}), \quad i = 1, \ldots, I, \quad j = 1, \ldots, J,$$

independently over the cells. The two sampling schemes considered above lead to the same deviance statistic.

If the grand total is fixed, then

$$\pi_{ij} = \frac{\mu_{ij}}{\mu_{..}},$$

and independence is $H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}$ for all i, j. If the row totals are fixed, then

$$\pi_{ij} = \frac{\mu_{ij}}{\mu_{i\cdot}},$$

and independence is $H_0 : \pi_{1j} = \cdots = \pi_{Ij} = \pi_j$ for each j. In both cases,

$$D = 2\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij} \log\left(\frac{y_{ij}}{y_{i\cdot}y_{\cdot j}/y_{\cdot\cdot}}\right) \dot{\sim} \chi^2_{(I-1)(J-1)}$$

asymptotically under H_0.

We now express the same test as a Poisson log-linear model. Using indicator variables for factor V and factor W, the main effects model is

$$\log\mu = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{I-1}x_{I-1}}_{\text{Factor } V} + \underbrace{\beta_I x_I + \cdots + \beta_{I+J-2}x_{I+J-2}}_{\text{Factor } W}$$

Equivalently, we write

$$\log\mu_{ij} = u + u_i^V + u_j^W, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J,$$

with corner-point constraints $u_1^V = u_1^W = 0$. This notation suppresses the binary indicator variables. The relationship between the regression parameters is

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

Remark 14.1 Another common identifiability convention is to use analysis-of-variance constraints,

$$\sum_{i=1}^{I} u_i^V = 0 \qquad \text{and} \qquad \sum_{j=1}^{J} u_j^W = 0.$$

Testing independence corresponds to comparing the main effects model

$$H_0 : \log \mu_{ij} = u + u_i^V + u_j^W$$

with the saturated interaction model

$$H_a : \log \mu_{ij} = u + u_i^V + u_j^W + u_{ij}^{VW}.$$

With corner-point constraints, $u_1^V = u_1^W = 0$, $u_{1j}^{VW} = 0$ for every j, and $u_{i1}^{VW} = 0$ for every i. The main effects model has $1 + (I-1) + (J-1) = I + J - 1$ parameters, while the saturated model has IJ parameters.

The Poisson log-likelihood is

$$\ell(\mu) = \sum_{i=1}^{I} \sum_{j=1}^{J} [y_{ij} \log \mu_{ij} - \mu_{ij} - \log y_{ij}!]$$

For the main effects model,

$$\log \mu_{ij} = u + u_i^V + u_j^W \qquad \text{and} \qquad u_1^V = u_1^W = 0.$$

Substitution gives

$$\ell(u) = \sum_{i=1}^{I} \sum_{j=1}^{J} \left[y_{ij}(u + u_i^V + u_j^W) - \exp\big(u + u_i^V + u_j^W\big) - \log y_{ij}!\right]$$

Differentiating with respect to the intercept gives

$$\begin{aligned} \frac{\partial \ell}{\partial u} &= \sum_{i=1}^{I} \sum_{j=1}^{J} \left[y_{ij} - \exp\big(u + u_i^V + u_j^W\big)\right] \\ &= \sum_{i=1}^{I} \sum_{j=1}^{J} [y_{ij} - \mu_{ij}] \end{aligned}$$

$$= y_{\cdot\cdot} - \mu_{\cdot\cdot}.$$

Setting this score equation to zero gives $\hat{\mu}_{\cdot\cdot} = y_{\cdot\cdot}$, so the grand total is preserved.

Differentiating with respect to $u^V_{i^*}$ gives

$$\begin{aligned}\frac{\partial \ell}{\partial u^V_{i^*}} &= \sum_{j=1}^{J} \left[y_{i^* j} - \exp\left(u + u^V_{i^*} + u^W_j\right) \right] \\ &= y_{i^* \cdot} - \mu_{i^* \cdot}.\end{aligned}$$

For $i^* > 1$, setting this score equation to zero gives $\hat{\mu}_{i^* \cdot} = y_{i^* \cdot}$. The intercept equation then supplies the omitted baseline row:

$$\hat{\mu}_{1\cdot} = \hat{\mu}_{\cdot\cdot} - \sum_{i=2}^{I} \hat{\mu}_{i\cdot} = y_{\cdot\cdot} - \sum_{i=2}^{I} y_{i\cdot} = y_{1\cdot}.$$

Thus all row totals are preserved.

Similarly,

$$\begin{aligned}\frac{\partial \ell}{\partial u^W_{j^*}} &= \sum_{i=1}^{I} \left[y_{i j^*} - \exp\left(u + u^V_i + u^W_{j^*}\right) \right] \\ &= y_{\cdot j^*} - \mu_{\cdot j^*},\end{aligned}$$

for $j^* > 1$. Combining these equations with the fitted grand total likewise gives $\hat{\mu}_{\cdot 1} = y_{\cdot 1}$, so all column totals are preserved.

For the Poisson log-linear deviance test,

$$D = 2(\ell(\tilde{\mu}) - \ell(\hat{\mu}))$$

Remark 14.2 Under the saturated model, $\tilde{\mu}_{ij} = y_{ij}$, so the saturated model fits every cell perfectly.

Remark 14.3 Under the constrained independence model,

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

because the main effects model reproduces the observed row, column, and grand totals.

$$\begin{aligned}D &= 2\sum_{i=1}^{I}\sum_{j=1}^{J}\left[(y_{ij}\log\tilde{\mu}_{ij} - \tilde{\mu}_{ij}) - (y_{ij}\log\hat{\mu}_{ij} - \hat{\mu}_{ij})\right] \\ &= 2\sum_{i=1}^{I}\sum_{j=1}^{J}\left[y_{ij}\log\left(\frac{\tilde{\mu}_{ij}}{\hat{\mu}_{ij}}\right) - (\tilde{\mu}_{ij} - \hat{\mu}_{ij})\right] \\ &= 2\sum_{i=1}^{I}\sum_{j=1}^{J}\left[y_{ij}\log\left(\frac{y_{ij}}{y_{i\cdot}y_{\cdot j}/y_{..}}\right)\right] - 2\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}(\tilde{\mu}_{ij} - \hat{\mu}_{ij})}_{=0 \text{ since } \tilde{\mu}_{..}=\hat{\mu}_{..}=y_{..}}\end{aligned}$$

This is the same likelihood ratio deviance statistic obtained from the multinomial and product multinomial formulations. The degrees of freedom are

$$n - p = IJ - \{1 + (I-1) + (J-1)\} = (I-1)(J-1),$$

so $D \dot{\sim} \chi^2_{(I-1)(J-1)}$ under the independence model.

Example 14.4 (Melanoma study) A cross-sectional study classified 400 patients with malignant melanoma according to histological type and tumour site.

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

The question is whether histological type and tumour site are independent. The null model is the main effects model

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

where V is tumour type and W is tumour site. With treatment indicators, the same model can be written as

$$\log \mu = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}_{\text{tumour type}} + \underbrace{\beta_4 x_4 + \beta_5 x_5}_{\text{tumour site}}.$$

The correspondence is

$$u = \beta_0, \qquad u_2^V = \beta_1, \qquad u_3^V = \beta_2,$$
$$u_4^V = \beta_3, \qquad u_2^W = \beta_4, \qquad \text{and} \qquad u_3^W = \beta_5.$$

For the melanoma data, Model 1 is the main effects model $u + u_i^V + u_j^W$. It tests independence:

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \text{for all } i, j,$$
$$H_a : \pi_{ij} \neq \pi_{i\cdot}\pi_{\cdot j} \quad \text{for at least one cell } (i, j).$$

The fitted model has residual deviance $D = 51.795$ on 6 degrees of freedom, so

$$\mathbb{P}(\chi_6^2 > 51.795) < 0.001.$$

We reject the null hypothesis of independence, so the main effects model is not adequate compared with the saturated model. Equivalently, this is a test of the interaction terms:

$$H_0 : u_{ij}^{VW} = 0 \quad \text{for all } i, j,$$
$$H_a : u_{ij}^{VW} \neq 0 \quad \text{for at least one cell } (i, j).$$

The fitted main effects model has coefficient summary

Term	Estimate	Standard error	z	p-value
Intercept	1.75440	0.20400	8.5999	$< 2.2 \times 10^{-16}$
type 2	1.69400	0.18659	9.0786	$< 2.2 \times 10^{-16}$
type 3	1.30195	0.19342	6.7312	1.682×10^{-11}
type 4	0.49899	0.21741	2.2951	0.021725
site 2	0.44393	0.15537	2.8573	0.004273
site 3	1.20103	0.13831	8.6834	$< 2.2 \times 10^{-16}$

The fitted values preserve the row and column totals. For example, for Hutchinson's freckle,

$$y_{1\cdot} = 22 + 2 + 10 = 34$$
$$\hat{y}_{1\cdot} = 5.78 + 9.01 + 19.21 = 34.$$

The fitted values and deviance residuals are

Tumour type	Site	y_{ij}	$\hat{\mu}_{ij}$	Deviance residual
Hutchinson's freckle	Head and neck	22	5.780	5.135
Hutchinson's freckle	Trunk	2	9.010	−2.828
Hutchinson's freckle	Extremities	10	19.210	−2.316
Superficial spreading	Head and neck	16	31.450	−3.045
Superficial spreading	Trunk	54	49.025	0.699
Superficial spreading	Extremities	115	104.525	1.008
Nodular	Head and neck	19	21.250	−0.497
Nodular	Trunk	33	33.125	−0.022
Nodular	Extremities	73	70.625	0.281
Indeterminate	Head and neck	11	9.520	0.468
Indeterminate	Trunk	17	14.840	0.548
Indeterminate	Extremities	28	31.640	−0.660

The fitted values and residuals indicate that Hutchinson's freckle occurs more often on the head and neck than expected under independence and less often on the trunk and extremities. Superficial

spreading melanoma occurs less often on the head and neck than expected.

Model 2 is the type-only model $u + u_i^V$. It asks whether all tumour sites are equally likely after accounting for tumour type, or equivalently whether $u_j^W = 0$ for all j. The deviance difference against Model 1 is

$$\Delta D = 150.1 - 51.795 = 98.305,$$

which is compared with χ^2_2. The p-value is less than 0.001, so we reject the type-only model.

Model 3 is the site-only model $u + u_j^W$. It asks whether all tumour types are equally likely after accounting for site. The residual deviance is 196.9, and the deviance difference from Model 1 has p-value less than 0.001, so we reject the site-only model as well. The different tumour types do not occur equally often, and melanoma does not occur equally often at the different body sites.

The row and column percentages summarize the main pattern in the data:

Row percentages

Tumour type	Head and neck	Trunk	Extremities	All sites
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
All types	17.0	26.5	56.5	100

and

Column percentages

Tumour type	Head and neck	Trunk	Extremities	All sites
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
All types	100	100	100	100

Lecture 15 — Thursday, March 03, 2016

Eikosograms

An eikosogram is a visual display for categorical variables. The base is a unit square representing total probability 1. The square is divided into vertical strips according to the probabilities of the conditioning events, and each strip is divided horizontally according to conditional probabilities. The resulting rectangle areas equal joint probabilities.

For a single binary variable Y, the shaded area in Figure 28 represents $\mathbb{P}(Y = 0) = 1/3$.

FIGURE 28

$$\mathbb{P}(Y = 0) = \frac{1}{3}.$$

For two independent binary variables Y and Z, the conditional heights for $Y = 0$ are the same in each vertical strip. This is shown in Figure 29.

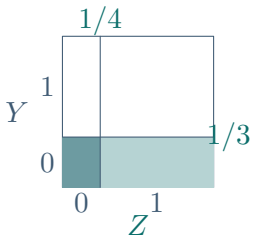

FIGURE 29

$$\mathbb{P}(Z = 0) = 1/4$$
$$\mathbb{P}(Y = 0 \mid Z = 0) = 1/3$$
$$\mathbb{P}(Y = 0 \mid Z = 1) = 1/3$$

		Z = 0	Z = 1	
Y	0	1	3	4
	1	2	6	8
		3	9	12

Thus Y is independent of Z. The joint probability is read as an area:

$$\mathbb{P}(Y=0,Z=0)=\frac{1}{12}=\left(\frac{1}{4}\right)\left(\frac{1}{3}\right).$$

For dependent data, the conditional heights need not agree across vertical strips. Figure 30 gives an example.

		X		
		0	1	
Y	0	1	3	4
	1	6	2	8
		7	5	12

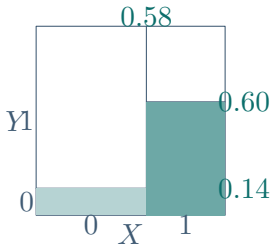

FIGURE 30

$$\begin{aligned}\mathbb{P}(X=0)&=7/12,\\ \mathbb{P}(Y=0\mid X=0)&=1/7,\\ \mathbb{P}(Y=0\mid X=1)&=3/5.\end{aligned}$$

$$\begin{aligned}\mathbb{P}(Y=0)&=4/12\\ &=\underbrace{\mathbb{P}(Y=0\mid X=0)\mathbb{P}(X=0)}_{\text{area 1}}+\underbrace{\mathbb{P}(Y=0\mid X=1)\mathbb{P}(X=1)}_{\text{area 2}}\\ &=(1/7)(7/12)+(3/5)(5/12)\\ &=1/12+3/12\\ &=4/12.\end{aligned}$$

$$\mathbb{P}(Y=0,X=0)=\frac{1}{12}\neq\mathbb{P}(Y=0)\mathbb{P}(X=0)=\left(\frac{4}{12}\right)\left(\frac{7}{12}\right).$$

For the melanoma data, the eikosogram conditional on location shows the main contribution

against independence. For example,

$$\mathbb{P}(\text{Hutchinson's freckle} \mid \text{head and neck}) = \frac{22}{68} = 0.32,$$

which is much larger than expected under the independence model. Figure 31 compares the eikosogram expected under independence with the observed eikosogram when location is placed on the horizontal axis and tumour type is placed on the vertical axis.

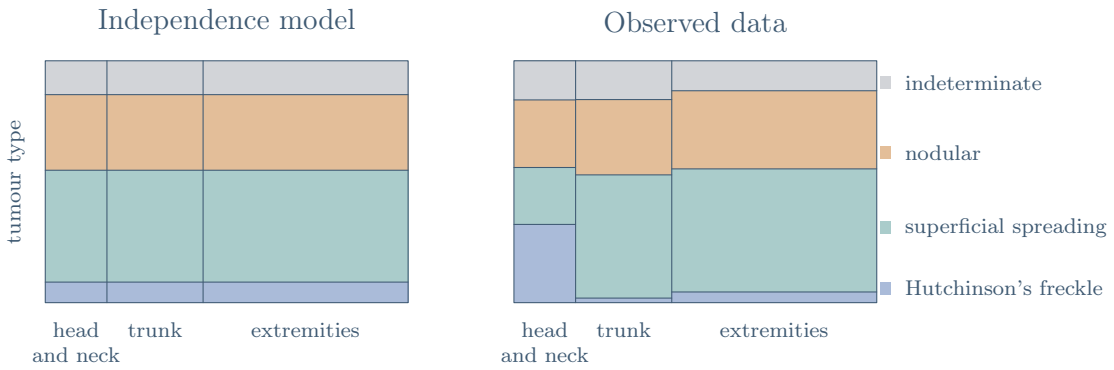

FIGURE 31

In the independence panel, the conditional distribution of tumour type is the same at each location. In the observed panel, the bottom band corresponding to Hutchinson's freckle is much taller in the head-and-neck strip and much shorter in the trunk and extremities strips. This visually matches the large positive and negative residuals in the fitted independence model.

Three-Way Tables

Figure 32 represents the indexing and marginal totals for a three-way table.

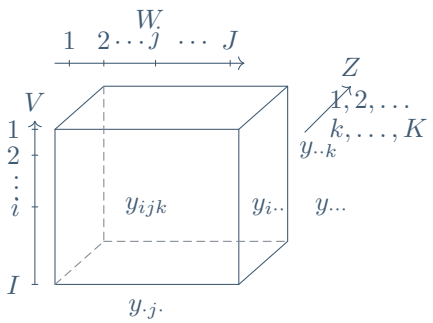

FIGURE 32

$$Y_{ijk} \sim \text{Poisson}(\mu_{ijk}), \quad \begin{matrix} i = 1, \ldots, I, \\ j = 1, \ldots, J, \\ k = 1, \ldots, K. \end{matrix}$$

If the grand total is fixed, the multinomial parameters are $\pi_{ijk} = \mu_{ijk}/\mu_{...}$, and they satisfy

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \pi_{ijk} = 1,$$

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

The saturated model is denoted (VWZ). In full, it 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,

$$u_1^V = u_1^W = u_1^Z = 0 \qquad \text{and} \qquad u_{1j}^{VW} = u_{i1}^{VW} = u_{1k}^{VZ} = u_{i1}^{VZ} = u_{1k}^{WZ} = u_{j1}^{WZ} = 0,$$

and

$$u_{1jk}^{VWZ} = u_{i1k}^{VWZ} = u_{ij1}^{VWZ} = 0.$$

This model has IJK parameters and therefore fits the data perfectly:

$$\tilde{\pi}_{ijk} = \frac{y_{ijk}}{y_{...}} \qquad \text{and} \qquad \tilde{\mu}_{ijk} = y_{...}\tilde{\pi}_{ijk} = y_{ijk}.$$

There are several different notions of independence for three-way tables. The factors V, W, and Z are mutually independent if

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

or equivalently

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

The corresponding log-linear model is the main effects model (V, W, Z). It can be written as

$$\log \mu_{ijk} = u + u_i^V + u_j^W + u_k^Z.$$

This model fits the one-factor marginal totals exactly, and its fitted values are

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

The number of parameters is

$$1 + (I-1) + (J-1) + (K-1) = I + J + K - 2,$$

so the residual degrees of freedom are $IJK - (I + J + K - 2)$.

A joint independence model can instead make factor W independent of the pair (V, Z):

$$\pi_{ijk} = \pi_{i\cdot k}\pi_{.j\cdot}.$$

Equivalently,

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

The corresponding log-linear model is (VZ, W):

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

This model fits the VZ combination totals $y_{i\cdot k}$ and the W marginal totals $y_{.j\cdot}$, with fitted values

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

The number of parameters is

$$1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1),$$

and the residual degrees of freedom are $IJK - (IK + J - 1)$. There are three joint independence models: (V, WZ), (VW, Z), and (VZ, W).

To describe conditional independence, define the conditional cell probabilities by

$$\pi_{ij|k} = \frac{\pi_{ijk}}{\pi_{..k}}.$$

Then V and W are conditionally independent given Z if

$$\pi_{ij|k} = \pi_{i\cdot|k}\pi_{.j|k}.$$

By definition,

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

Under conditional independence,

$$\pi_{ijk} = \pi_{i\cdot|k}\pi_{\cdot j|k}\pi_{..k} = \mathbb{P}(V = i \mid Z = k)\mathbb{P}(W = j \mid Z = k)\mathbb{P}(Z = k).$$

The corresponding log-linear model is (VZ, WZ):

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

This model fits the VZ and WZ combination totals $y_{i\cdot k}$ and $y_{\cdot jk}$. For each stratum with $y_{..k} > 0$, the fitted values are

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

An empty stratum has all observed and fitted counts equal to zero and contributes no information. The number of parameters is

$$1 + (I-1) + (J-1) + (K-1) + (I-1)(K-1) + (J-1)(K-1),$$

and the residual degrees of freedom are $IJK - (IK + JK - K)$. There are three conditional independence models: (VZ, WZ), (VW, VZ), and (VW, WZ).

The homogeneous association model is (VW, VZ, WZ):

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

This model contains every two-factor interaction but no three-factor interaction. The number of parameters is

$$1 + (I-1) + (J-1) + (K-1) + (I-1)(J-1) + (I-1)(K-1) + (J-1)(K-1),$$

and the residual degrees of freedom are $(I-1)(J-1)(K-1)$.

To interpret this model, fix a level k^* of factor Z. Then

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

For the two-way $I \times J$ table obtained by fixing $Z = k^*$, this is a saturated two-way model for V and W. More precisely, for any two levels i, i' of V and j, j' of W, the conditional log odds ratio is

$$\log\left(\frac{\mu_{ijk}\mu_{i'j'k}}{\mu_{ij'k}\mu_{i'jk}}\right).$$

All terms involving Z, VZ, or WZ cancel from this contrast, leaving only a contrast of the VW interaction parameters; it therefore does not depend on k. Thus the conditional odds-ratio association between V and W is the same at every level of Z, although their conditional marginal probabilities may change with Z.

The same argument holds after fixing levels of V or W. Homogeneous association means that the conditional odds ratios between any two factors are the same at every level of the third factor.

For a three-way log-linear model, the general Poisson deviance is

$$D = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left[O_{ijk}\log\left(\frac{O_{ijk}}{E_{ijk}}\right) - (O_{ijk} - E_{ijk})\right],$$

where $O_{ijk} = y_{ijk}$ and $E_{ijk} = \hat{\mu}_{ijk}$. If the model under the null hypothesis has q parameters, then D is approximately χ^2_{IJK-q} under that model. For nested models, the deviance difference

$$\Delta D = D_0 - D_A$$

is approximately χ^2_{p-q}, where p is the number of parameters in the larger alternative model and q is the number of parameters in the smaller null model.

Every hierarchical model considered here contains an intercept, so its fitted total equals the observed total. Consequently $\sum_{ijk}(O_{ijk} - E_{ijk}) = 0$, and the deviance simplifies to $2\sum_{ijk} O_{ijk}\log(O_{ijk}/E_{ijk})$, with $0\log(0/E)$ interpreted as zero.

Hierarchical Models

The hierarchy of three-way log-linear models is shown in Figure 33.

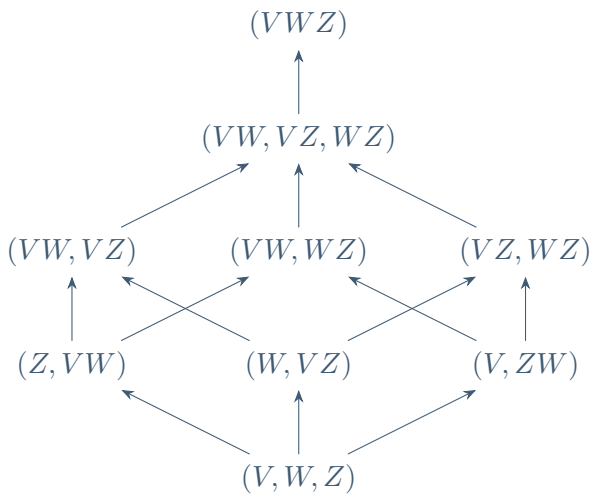

FIGURE 33

Lecture 16 — Tuesday, March 08, 2016

Three-Way Applications

The first application considers methods of suicide classified by sex, age group, and method. Let X denote sex, with $i = 1, 2$; let Y denote age group, with $j = 1, 2, 3$; and let Z denote method, with $k = 1, \ldots, 6$.

Sex	Age	Solid/liquid	Gas	Respiratory	Weapons	Jumping	Other
Male	10–40	398	121	455	155	55	124
Male	40–70	399	82	797	168	51	82
Male	> 70	93	6	316	33	26	14
Female	10–40	259	15	95	14	40	38
Female	40–70	450	13	450	26	71	60
Female	> 70	154	5	185	7	38	10

The goals are to select an appropriate log-linear model and then interpret it. The saturated model is (XYZ), the homogeneous association model is (XY, XZ, YZ), and the conditional independence candidates are (XY, XZ), (XY, YZ), and (XZ, YZ).

To compare the homogeneous association model with the saturated model, test

$$H_0 : u_{ijk}^{XYZ} = 0 \quad \text{for all three-factor interaction terms}$$

against the alternative that at least one three-factor interaction term is nonzero. The deviance difference is

$$\Delta D = D_2 - D_1 = 14.9 - 0 = 14.9,$$

which is compared with χ^2_{10}, since $(2-1)(3-1)(6-1) = 10$. The p-value is

$$\mathbb{P}(\chi^2_{10} > 14.9) = 0.1357.$$

We do not reject homogeneous association, so Model 2 is adequate.

The conditional independence models are not adequate. The analysis-of-deviance summary is

Model	Form	Residual deviance	Residual df	Comparison p-value
1	(XYZ)	0	0	not applicable
2	(XY, XZ, YZ)	14.9	10	0.14 vs. Model 1
3	(XY, XZ)	293.2	20	< 0.001 vs. Model 2
4	(XY, YZ)	389.0	15	< 0.001 vs. Model 2
5	(XZ, YZ)	154.7	12	< 0.001 vs. Model 2

Thus further simplification is not supported. We therefore select Model 2, the homogeneous association model:

$$\log \mu_{ijk} = u + u_i^X + u_j^Y + u_k^Z + u_{ij}^{XY} + u_{ik}^{XZ} + u_{jk}^{YZ}.$$

The fitted coefficients for Model 2, using corner-point constraints and treatment coding, are as follows. These coefficients give the numerical values used in the later interpretation.

Coefficient	Estimate	Standard error	Coefficient	Estimate	Standard error
$\hat{u}$	6.0182	0.0460	$\hat{u}_{23}^{XZ}$	−0.8652	0.0677
$\hat{u}_2^X$	−0.5123	0.0650	$\hat{u}_{24}^{XZ}$	−2.0041	0.1637
$\hat{u}_2^Y$	−0.0795	0.0615	$\hat{u}_{25}^{XZ}$	0.1147	0.1312
$\hat{u}_3^Y$	−1.4160	0.0894	$\hat{u}_{26}^{XZ}$	−0.6038	0.1290
$\hat{u}_2^Z$	−1.2085	0.0985	$\hat{u}_{22}^{YZ}$	−0.3784	0.1463
$\hat{u}_3^Z$	0.0668	0.0615	$\hat{u}_{32}^{YZ}$	−1.2327	0.3246
$\hat{u}_4^Z$	−0.9660	0.0898	$\hat{u}_{23}^{YZ}$	0.7037	0.0752
$\hat{u}_5^Z$	−1.9783	0.1216	$\hat{u}_{33}^{YZ}$	1.0640	0.0999
$\hat{u}_6^Z$	−1.2140	0.0945	$\hat{u}_{24}^{YZ}$	0.1409	0.1207
$\hat{u}_{22}^{XY}$	0.7254	0.0706	$\hat{u}_{34}^{YZ}$	−0.1289	0.1951
$\hat{u}_{23}^{XY}$	0.9026	0.0915	$\hat{u}_{25}^{YZ}$	−0.0268	0.1483
$\hat{u}_{22}^{XZ}$	−1.7096	0.1951	$\hat{u}_{35}^{YZ}$	0.5579	0.1805
			$\hat{u}_{26}^{YZ}$	−0.2857	0.1282
			$\hat{u}_{36}^{YZ}$	−0.8022	0.2329

The residual deviance for this fitted model is 14.9007 on 10 degrees of freedom, with Akaike information criterion 284.995. The plots of the deviance residuals in Figure 34 show the same conclusion graphically: Model 2 has comparatively small, patternless residuals, while the three conditional independence models leave much larger residual structure.

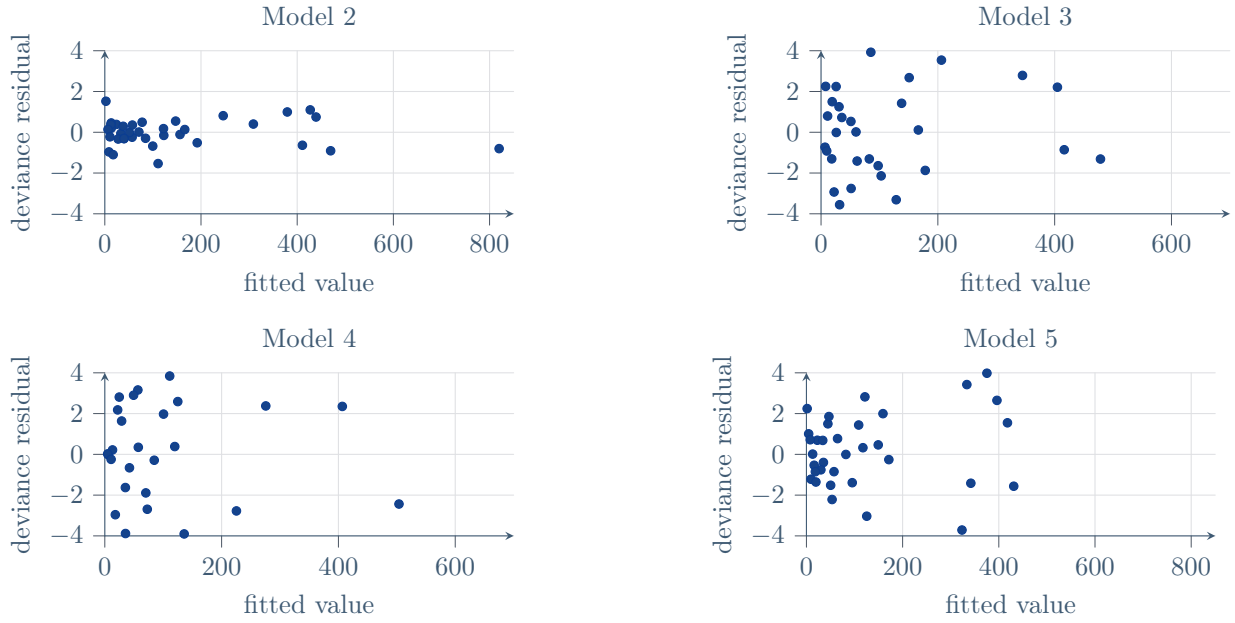

FIGURE 34

Example 16.1 (Seatbelt use, ejection, and fatality) The second application is a 2×2×2 table on seatbelt use, ejection, and fatality in automobile accidents. Let V denote seatbelt use, where $i = 1$ means not used and $i = 2$ means used. Let W denote ejection, where $j = 1$

means not ejected and $j = 2$ means ejected. Let Z denote injury status, where $k = 1$ means nonfatal and $k = 2$ means fatal.

Seatbelt use	Ejection	Nonfatal	Fatal
Used	Yes	1105	14
Used	No	411111	483
Not used	Yes	4624	497
Not used	No	157342	1008

Assume

$$Y_{ijk} \sim \text{Poisson}(\mu_{ijk}), \quad i, j, k = 1, 2.$$

Only the grand total $Y_{\cdots} = y_{\cdots}$ is fixed, so the multinomial formulation is appropriate after conditioning. The saturated model is (VWZ), and the homogeneous association model is (VW, VZ, WZ). Testing homogeneous association is the test

$$H_0 : u_{222}^{VWZ} = 0.$$

The fitted homogeneous association model has residual deviance 2.85402 on 1 degree of freedom, so

$$\mathbb{P}(\chi_1^2 > 2.85402) = 0.0911.$$

We do not reject H_0, so the homogeneous association model is adequate. Equivalently, the Wald test for $H_0 : u_{222}^{VWZ} = 0$ in the saturated model gives $p = 0.1127$.

For the homogeneous association model,

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

The fitted coefficients are

Coefficient	Estimate	Standard error	Coefficient	Estimate	Standard error
$\hat{u}$	11.9661	0.0025	$\hat{u}_{22}^{VW}$	−2.3996	0.0333
$\hat{u}_2^V$	0.9605	0.0030	$\hat{u}_{22}^{VZ}$	−1.7173	0.0540
$\hat{u}_2^W$	−3.5256	0.0149	$\hat{u}_{22}^{WZ}$	2.7978	0.0553
$\hat{u}_2^Z$	−5.0436	0.0312			

The conditional independence models are not adequate:

Model	Residual deviance	Degrees of freedom	Comparison
(VWZ)	0	0	not applicable
(VW, VZ, WZ)	2.85	1	$p = 0.09$ versus (VWZ)
(VW, VZ)	1680.41	2	$p < 0.001$ versus Model 2
(VW, WZ)	1144.64	2	$p < 0.001$ versus Model 2
(VZ, WZ)	7133.98	2	$p < 0.001$ versus Model 2

We therefore select (VW, VZ, WZ) for interpretation.

The fitted and observed values under the selected model are

Seatbelt use	Ejection	$\hat{\mu}_{ij1}$	y_{ij1}	$\hat{\mu}_{ij2}$	y_{ij2}
Used	Yes	1098.1	1105	20.9	14
Used	No	411117.9	411111	476.1	483
Not used	Yes	4630.9	4624	490.1	497
Not used	No	157335.1	157342	1014.9	1008

The corresponding observed row percentages are

Seatbelt use	Ejection	Nonfatal	Fatal
Used	Yes	98.7	1.3
Used	No	99.9	0.1
Not used	Yes	90.4	9.6
Not used	No	99.4	0.6

Find the odds ratio for a fatal accident, $Z = 2$, comparing passengers who were ejected, $W = 2$, with those who were not, $W = 1$, among passengers who did not use their seatbelt, $V = 1$.

Solution. The odds of a fatal accident for passengers with no seatbelt who were not ejected are

$$\frac{\mathbb{P}(Z = 2 \mid V = 1, W = 1)}{\mathbb{P}(Z = 1 \mid V = 1, W = 1)}.$$

The first odds can be written as

$$\begin{aligned}\frac{\mathbb{P}(Z=2 \mid V=1, W=1)}{\mathbb{P}(Z=1 \mid V=1, W=1)} &= \frac{\mathbb{P}(V=1, W=1, Z=2)/\mathbb{P}(V=1, W=1)}{\mathbb{P}(V=1, W=1, Z=1)/\mathbb{P}(V=1, W=1)} \\ &= \frac{\pi_{112}}{\pi_{111}} \\ &= \frac{\mu_{112}}{\mu_{111}}.\end{aligned}$$

Using the homogeneous association model, the relevant fitted means are as follows.

V	W	Z	$\log \mu_{ijk}$
1	1	2	$u + u_2^Z$
1	1	1	u

$$\log\left(\frac{\mu_{112}}{\mu_{111}}\right) = u_2^Z.$$

The odds of a fatal accident for passengers with no seatbelt who were ejected are μ_{122}/μ_{121}. The relevant fitted means are

V	W	Z	$\log \mu_{ijk}$
1	2	2	$u + u_2^W + u_2^Z + u_{22}^{WZ}$
1	2	1	$u + u_2^W$

$$\log\left(\frac{\mu_{122}}{\mu_{121}}\right) = u_2^Z + u_{22}^{WZ}$$

Let ψ denote the odds ratio for fatality comparing ejected versus not ejected passengers among those not wearing a seatbelt. Then

$$\log \psi = \log\left(\frac{\pi_{122}/\pi_{121}}{\pi_{112}/\pi_{111}}\right) = (u_2^Z + u_{22}^{WZ}) - u_2^Z = u_{22}^{WZ}.$$

Thus

$$\hat{\psi} = \exp\big(\hat{u}_{22}^{WZ}\big) = \exp(2.80) = 16.4.$$

Among passengers not wearing a seatbelt, the odds of fatality are estimated to be 16.4 times higher for ejected passengers than for passengers who were not ejected.

If the three-way interaction term were included, the odds ratio among passengers not wearing a seatbelt would not change. For passengers wearing a seatbelt, let ψ_{belt} denote the corresponding fatality odds ratio comparing passengers who were ejected with those who were not. Then

$$\log \psi_{\text{belt}} = u_{22}^{WZ} + u_{222}^{VWZ}.$$

Thus the association between fatality and ejection would depend on seatbelt use if $u_{222}^{VWZ} \neq 0$. In this example, we did not reject $u_{222}^{VWZ} = 0$.

More generally, log-linear parameters in contingency table models have interpretations as log odds ratios. Under the selected homogeneous association model, the three main odds ratio summaries are

Outcome	Comparison	Condition	Estimate
$Z = 2$	$W = 2$ versus $W = 1$	$V = 1$	$\exp\big(\hat{u}_{22}^{WZ}\big) = \exp(2.80) = 16.4$
$Z = 2$	$W = 2$ versus $W = 1$	$V = 2$	$\exp\big(\hat{u}_{22}^{WZ} + \hat{u}_{222}^{VWZ}\big) = \exp(2.80 + 0) = 16.4$
$Z = 2$	$V = 2$ versus $V = 1$	$W = 1$	$\exp\big(\hat{u}_{22}^{VZ}\big) = \exp(-1.72) = 0.18$
$Z = 2$	$V = 2$ versus $V = 1$	$W = 2$	$\exp\big(\hat{u}_{22}^{VZ} + \hat{u}_{222}^{VWZ}\big) = \exp(-1.72 + 0) = 0.18$
$W = 2$	$V = 2$ versus $V = 1$	$Z = 1$	$\exp\big(\hat{u}_{22}^{VW}\big) = \exp(-2.40) = 0.09$
$W = 2$	$V = 2$ versus $V = 1$	$Z = 2$	$\exp\big(\hat{u}_{22}^{VW} + \hat{u}_{222}^{VWZ}\big) = \exp(-2.40 + 0) = 0.09$

Thus the relative odds of fatality are much higher for ejected passengers, the relative odds of fatality are lower for passengers wearing seatbelts, and the relative odds of ejection are lower for passengers wearing seatbelts. Since the three conditional independence models were rejected, all three two-factor associations are needed. □

Lecture 17 — Tuesday, March 15, 2016

The selected model for the suicide data was the homogeneous association model,

$$\log \mu_{ijk} = u + u_i^X + u_j^Y + u_k^Z + u_{ij}^{XY} + u_{ik}^{XZ} + u_{jk}^{YZ}.$$

Pairwise eikosograms can be used to inspect the three two-factor interactions. In Figure 35, the strongest visible patterns are that older age groups are more likely to use respiratory methods,

males are more likely than females to use weapons, and males are more likely to belong to the youngest age group.

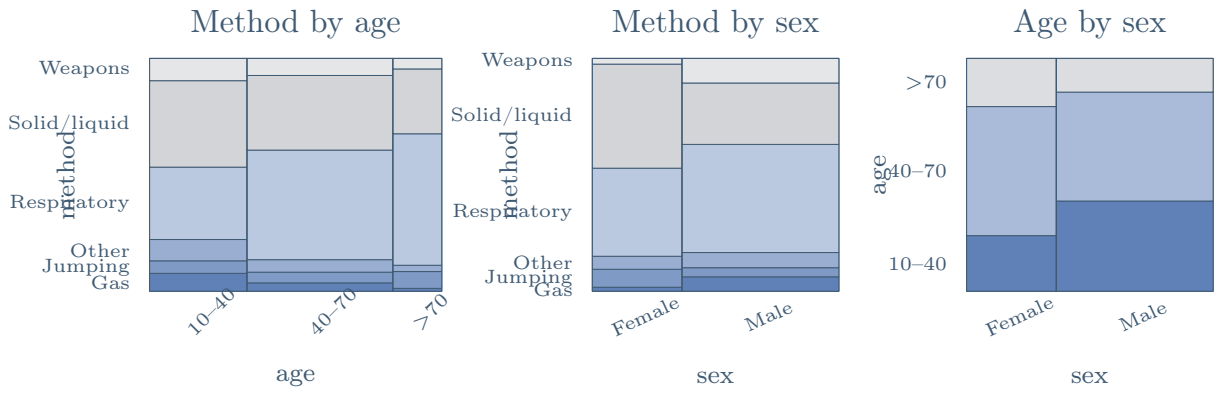

FIGURE 35

The corresponding three-factor eikosograms place method, sex, and age in turn on the vertical axis. Under homogeneous association, the conditional odds ratios for each pair are constant across levels of the third variable, even though the conditional proportions themselves may differ. Thus Figure 36 displays the method-by-age pattern separately for each sex, Figure 37 displays the sex composition across age groups separately for each method, and Figure 38 displays the age-by-sex pattern separately for each method. In Figure 38, the horizontal order within each method is female followed by male.

Example 17.1 For a male aged 40–70 years, sex is $X = 1$ and age group is $Y = 2$. The fitted mean for gas as the method of suicide is

$$\log \hat{\mu}_{122} = u + u_2^Y + u_2^Z + u_{22}^{YZ}.$$

The odds of using gas versus any other method are

$$\begin{aligned} \text{odds} &= \frac{\mathbb{P}(Z = 2 \mid X = 1, Y = 2)}{1 - \mathbb{P}(Z = 2 \mid X = 1, Y = 2)} \\ &= \frac{\pi_{122}/\pi_{12\cdot}}{1 - \pi_{122}/\pi_{12\cdot}} \\ &= \frac{\pi_{122}}{\pi_{12\cdot} - \pi_{122}} \\ &= \frac{\mu_{122}}{\mu_{12\cdot} - \mu_{122}}. \end{aligned}$$

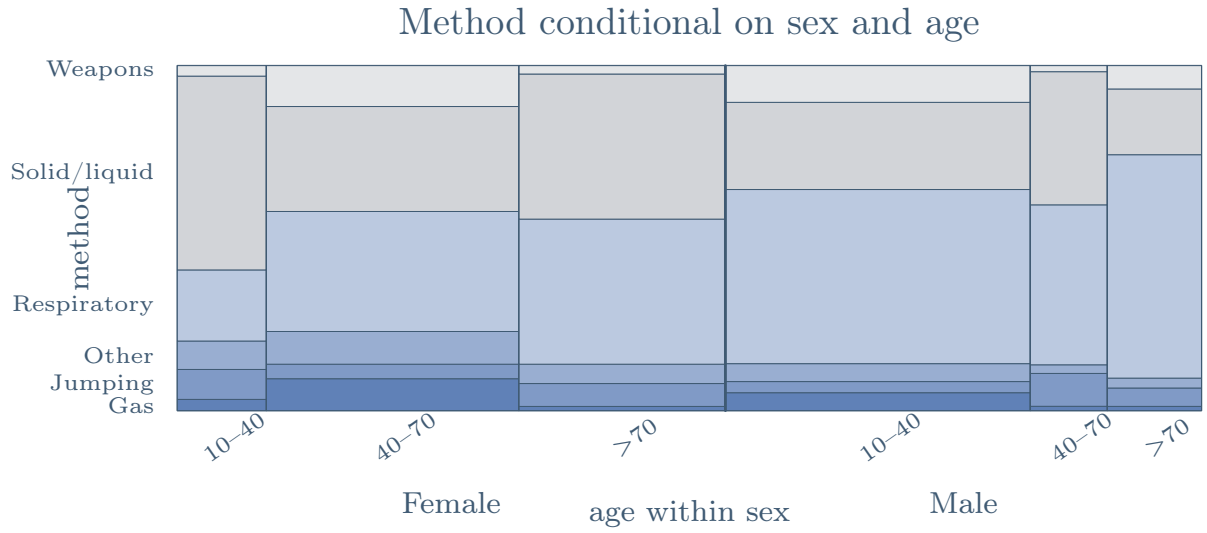

FIGURE 36

FIGURE 37

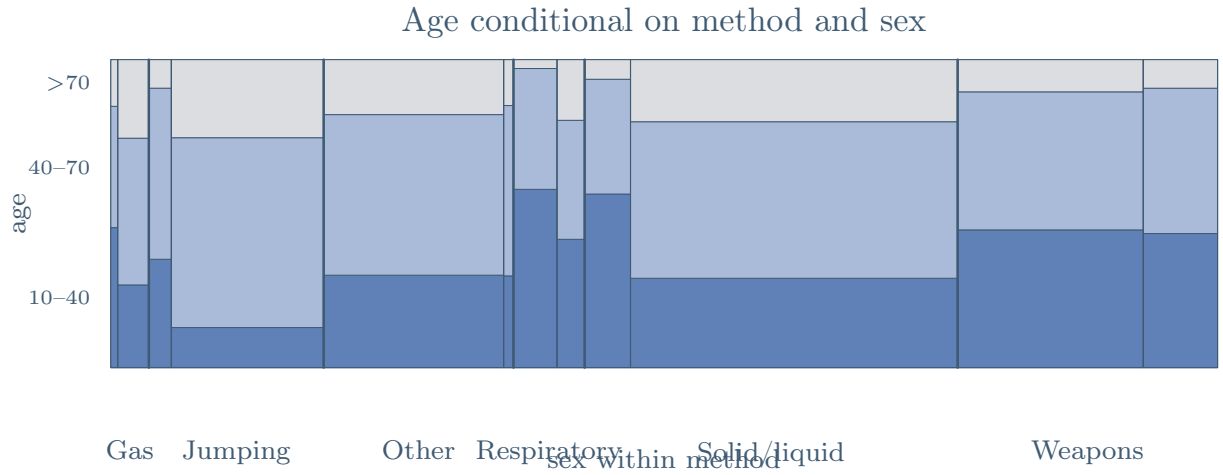

FIGURE 38

In terms of the log-linear parameters, this is

$$\frac{\exp\left(u + u_2^Y + u_2^Z + u_{22}^{YZ}\right)}{\sum_{\substack{k=1 \\ k\neq 2}}^{6} \exp\left(u + u_2^Y + u_k^Z + u_{2k}^{YZ}\right)}.$$

Notation 17.2 The corner-point constraints include $u_1^Z = 0$ and $u_{21}^{YZ} = 0$.

Under the homogeneous association model,

$$\log \mu_{ijk} = u + u_i^X + u_j^Y + u_k^Z + u_{ij}^{XY} + u_{ik}^{XZ} + u_{jk}^{YZ}.$$

Next estimate the relative probability of using jumping rather than weapons for a male aged 40–70 years.

$$\mathbb{P}(\text{jumping} \mid \text{male, 40–70}) = \mathbb{P}(Z = 5 \mid X = 1, Y = 2) = \frac{\pi_{125}}{\pi_{12\cdot}} = \frac{\mu_{125}/\mu_{\cdots}}{\mu_{12\cdot}/\mu_{\cdots}} = \frac{\mu_{125}}{\mu_{12\cdot}}$$
$$\mathbb{P}(\text{weapons} \mid \text{male, 40–70}) = \mathbb{P}(Z = 4 \mid X = 1, Y = 2) = \frac{\pi_{124}}{\pi_{12\cdot}} = \frac{\mu_{124}/\mu_{\cdots}}{\mu_{12\cdot}/\mu_{\cdots}} = \frac{\mu_{124}}{\mu_{12\cdot}}$$

$$\text{relative probability} = \frac{\pi_{125}/\pi_{12\cdot}}{\pi_{124}/\pi_{12\cdot}} = \frac{\mu_{125}}{\mu_{124}}.$$

sex	age	method	$\log \mu_{ijk}$
1	2	5	$u + u_2^Y + u_5^Z + u_{25}^{YZ}$
1	2	4	$u + u_2^Y + u_4^Z + u_{24}^{YZ}$

$$\log\left(\frac{\mu_{125}}{\mu_{124}}\right) = u_5^Z + u_{25}^{YZ} - u_4^Z - u_{24}^{YZ}$$
$$\frac{\pi_{125}}{\pi_{124}} = \exp\big(u_5^Z + u_{25}^{YZ} - u_4^Z - u_{24}^{YZ}\big).$$

$$\widehat{\text{relative probability}} = \exp((-1.978) + (-0.027) - (-0.966) - (0.141)) = \exp(-1.18) = 0.31.$$

Using the contrast covariance calculation $\sqrt{\boldsymbol{c}^\top \boldsymbol{I}^{-1}\boldsymbol{c}} = 0.138$, a 95% confidence interval is

$$(\exp(-1.18 - 1.96(0.138)), \exp(-1.18 + 1.96(0.138))) = (0.23, 0.40).$$

Thus, among males aged 40–70 years, the estimated probability of using jumping as a method of suicide is 0.31 times the probability of using weapons.

The null hypothesis of equal probabilities is

$$H_0 : \frac{\pi_{125}}{\pi_{124}} = 1.$$

Equivalently,

$$H_0 : u_5^Z + u_{25}^{YZ} - u_4^Z - u_{24}^{YZ} = 0.$$

Using a Wald test gives

$$p = \mathbb{P}\left(\chi_1^2 > \left(\frac{-1.18}{0.138}\right)^2\right) < 0.001.$$

We reject H_0. Males aged 40–70 years are significantly less likely to use jumping than weapons.

Multinomial Response

The coal mining data classify miners by covariate class V, $i = 1, \ldots, 8$, and pneumoconiosis response W, $j = 0, 1, 2$, where 0 is normal, 1 is mild disease, and 2 is severe disease. The continuous exposure variable x is the midpoint of the time interval spent working at the coalface.

Class	Normal	Mild	Severe	Time period	Row total
1	98	0	0	5.8	98
2	51	2	1	15.0	54
3	34	6	3	21.5	43
4	35	5	8	27.5	48
5	32	10	9	33.5	51
6	23	7	8	39.5	38
7	12	6	10	46.0	28
8	4	2	5	51.5	11

The scientific question is how disease status depends on exposure to the coalface. Although the data are recorded as a contingency table, disease status is the response of interest. Thus the row totals, corresponding to the numbers of miners in the time period categories, are treated as fixed.

The saturated model is

$$\log \mu_{ij} = u + u_i^V + u_j^W + u_{ij}^{VW}.$$

With corner-point constraints $u_1^V = 0$, $u_0^W = 0$, $u_{1j}^{VW} = 0$ for each j, and $u_{i0}^{VW} = 0$ for each i, this model has 24 parameters.

Instead, use a more parsimonious model with a linear exposure effect that depends on the response category:

$$\log \mu_{ij} = u + u_i^V + u_j^W + u_j^{WX} x.$$

The interaction between x and W allows a separate linear exposure effect for mild disease and severe disease. This model has 12 parameters.

With $W = 0$ as the baseline response category, the fitted model has residual deviance 13.9281 on

12 degrees of freedom and Akaike information criterion 127.332. The fitted coefficients are

Coefficient	Estimate	Standard error	Coefficient	Estimate	Standard error
$\hat{u}$	4.5514	0.1017	$\hat{u}_7^V$	−2.1773	0.2589
$\hat{u}_2^V$	−0.6399	0.1697	$\hat{u}_8^V$	−3.4825	0.3847
$\hat{u}_3^V$	−0.9291	0.1836	$\hat{u}_1^W$	−4.2917	0.5214
$\hat{u}_4^V$	−0.9147	0.1780	$\hat{u}_2^W$	−5.0598	0.5964
$\hat{u}_5^V$	−1.0060	0.1784	$\hat{u}_1^{WX}$	0.0836	0.0153
$\hat{u}_6^V$	−1.5273	0.2077	$\hat{u}_2^{WX}$	0.1093	0.0165

For covariate class 5, where $x_5 = 33.5$, estimate the relative probability of severe versus normal disease.

$$\frac{\mathbb{P}(\text{severe} \mid \text{class 5})}{\mathbb{P}(\text{normal} \mid \text{class 5})} = \frac{\mathbb{P}(W=2 \mid V=5)}{\mathbb{P}(W=0 \mid V=5)} = \frac{\pi_{52}}{\pi_{50}} = \frac{\mu_{52}/\mu_{5\cdot}}{\mu_{50}/\mu_{5\cdot}} = \frac{\mu_{52}}{\mu_{50}}.$$

V class	W response	x	$\log \mu_{ij}$
5	2	x_5	$u + u_5^V + u_2^W + u_2^{WX} x_5$
5	0	x_5	$u + u_5^V$

$$\log\left(\frac{\mu_{52}}{\mu_{50}}\right) = u_2^W + u_2^{WX} x_5.$$

$$\begin{aligned}
\widehat{\text{relative probability}} &= \frac{\hat{\mu}_{52}}{\hat{\mu}_{50}} \\
&= \exp\big(\hat{u}_2^W + \hat{u}_2^{WX} x_5\big) \\
&= \exp((-5.0598) + (0.1093)(33.5)) \\
&= \exp(-1.398) \\
&= 0.25.
\end{aligned}$$

$$\text{standard error}\left(\hat{u}_2^W + \hat{u}_2^{WX} x_5\right) = \sqrt{\operatorname{Var}(\hat{u}_2^W) + x_5^2 \operatorname{Var}(\hat{u}_2^{WX}) + 2x_5 \operatorname{Cov}(\hat{u}_2^W, \hat{u}_2^{WX})}$$

$$= \left[(0.3557) + (33.5)^2(0.000271) + 2(33.5)(-0.009354)\right]^{1/2}$$
$$= 0.182.$$

A 95% confidence interval for the relative probability is therefore

$$(\exp(-1.398 - 1.96(0.182)), \exp(-1.398 + 1.96(0.182))) = (0.17, 0.35).$$

The null hypothesis of equal probabilities is

$$H_0 : \frac{\pi_{52}}{\pi_{50}} = 1.$$

Equivalently,

$$H_0 : u_2^W + u_2^{WX} x_5 = 0.$$

$$p = \mathbb{P}\left(\chi_1^2 > \left(\frac{-1.398}{0.182}\right)^2\right) < 0.001.$$

We reject H_0.

Lecture 18 — Thursday, March 17, 2016

5. Overdispersion and Quasi-Likelihood

Poisson Overdispersion

For a Poisson response, the model mean and variance are tied together:

$$\mathbb{E}[Y] = \mu \qquad \text{and} \qquad \text{Var}(Y) = \mu.$$

This can be too restrictive for count data. When the observed variance is larger than the model variance, the data are said to be overdispersed. A simple way to represent this extra variation is

to introduce a dispersion parameter $\phi > 0$ and use the working variance

$$\text{Var}(Y) = \phi\mu.$$

If $\phi > 1$, the usual Poisson standard errors tend to be too small. Adjusting for overdispersion therefore widens confidence intervals and reduces the chance of rejecting $H_0 : \beta_j = 0$ solely because the fitted standard error is underestimated. After such an adjustment, the model selection process may need to be revisited, since terms that looked statistically significant under the naive Poisson fit may no longer be significant.

Ad Hoc Dispersion Adjustment

Consider independent observations from an exponential family with

$$f(y_i; \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i; \phi)\right) \qquad \text{and} \qquad a_i(\phi) = \frac{\phi}{w_i}.$$

Compare a constrained p-parameter model with fitted canonical parameters $\hat{\theta}_i$ against a larger q-parameter model with fitted canonical parameters $\tilde{\theta}_i$. The scaled likelihood ratio statistic is

$$D^* = 2\{\ell(\tilde{\theta}, \phi) - \ell(\hat{\theta}, \phi)\} = \frac{2}{\phi}\sum_{i=1}^{n} w_i\left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - \{b(\tilde{\theta}_i) - b(\hat{\theta}_i)\}\right] = \frac{D}{\phi},$$

where D is the usual deviance obtained when $\phi = 1$.

Asymptotically under the null hypothesis that the p-parameter model is adequate,

$$D^* \dot{\sim} \chi^2_{q-p}, \qquad \text{so} \qquad D \dot{\sim} \phi\chi^2_{q-p}.$$

In particular, when the larger model is saturated, $q = n$. Since $\mathbb{E}[\chi^2_{n-p}] = n - p$, a method-of-moments estimate of ϕ is

$$\hat{\phi} = \frac{D}{n-p}.$$

Thus $D \gg n - p$ is evidence of overdispersion, while D close to $n - p$ suggests ϕ close to 1.

The ad hoc adjustment keeps the fitted regression coefficients from the original generalized linear model, but rescales the estimated covariance matrix:

$$\widehat{\text{Cov}}_{\text{adj}}(\hat{\boldsymbol{\beta}}) = \hat{\phi}\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}).$$

Equivalently, each standard error is multiplied by $\sqrt{\hat{\phi}}$:

$$\text{standard error}_{\text{adjusted}}(\hat{\beta}_j) = \sqrt{\hat{\phi}}\,\text{standard error}(\hat{\beta}_j).$$

Mixed Poisson Model

The ad hoc adjustment is useful, but it does not by itself define a probability model. A model-based way to create overdispersion is to introduce a positive random effect u_i and assume

$$Y_i \mid u_i \sim \text{Poisson}(u_i\lambda), \qquad \mathbb{E}[u_i] = 1, \qquad \text{and} \qquad \text{Var}(u_i) = \kappa.$$

The symbol κ distinguishes this random-effect variance from the constant quasi-Poisson scale ϕ used in the preceding section. The random factor u_i inflates the conditional mean when $u_i > 1$ and deflates it when $u_i < 1$.

The unconditional mean is

$$\mathbb{E}[Y_i] = \mathbb{E}[\mathbb{E}[Y_i \mid u_i]] = \mathbb{E}[u_i\lambda] = \lambda.$$

The unconditional variance is

$$\begin{aligned}
\text{Var}(Y_i) &= \text{Var}(\mathbb{E}[Y_i \mid u_i]) + \mathbb{E}[\text{Var}(Y_i \mid u_i)] \\
&= \text{Var}(u_i\lambda) + \mathbb{E}[u_i\lambda] \\
&= \lambda^2\kappa + \lambda \\
&= \lambda(1 + \lambda\kappa).
\end{aligned}$$

Thus the variance is inflated by the factor $1 + \lambda\kappa$.

Now suppose

$$u_i \sim \text{Gamma}(\alpha, \beta),$$

where α is the shape parameter and β is the scale parameter. Then

$$\mathbb{E}[u_i] = \alpha\beta \qquad \text{and} \qquad \text{Var}(u_i) = \alpha\beta^2.$$

The requirements $\mathbb{E}[u_i] = 1$ and $\text{Var}(u_i) = \kappa$ are met by

$$\alpha = \frac{1}{\kappa} \qquad \text{and} \qquad \beta = \kappa.$$

The Gamma density integrates to one, so

$$1 = \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} u^{\alpha-1} e^{-u/\beta} \, \mathrm{d}u,$$

and therefore

$$\Gamma(\alpha)\beta^\alpha = \int_0^\infty u^{\alpha-1} e^{-u/\beta} \, \mathrm{d}u. \tag{18.1}$$

The marginal mass function of Y_i is obtained by integrating out u_i:

$$\begin{aligned} p(y_i; \lambda, \kappa) &= \int_0^\infty f(y_i \mid u_i, \lambda) g(u_i; \kappa) \, \mathrm{d}u_i \\ &= \int_0^\infty \frac{(u_i\lambda)^{y_i} e^{-u_i\lambda}}{y_i!} \frac{1}{\Gamma(\alpha)\beta^\alpha} u_i^{\alpha-1} e^{-u_i/\beta} \, \mathrm{d}u_i \\ &= \frac{\lambda^{y_i}}{y_i!\Gamma(\alpha)\beta^\alpha} \int_0^\infty u_i^{y_i+\alpha-1} e^{-u_i(\lambda+1/\beta)} \, \mathrm{d}u_i. \end{aligned}$$

The integrand is a Gamma kernel with shape $y_i + \alpha$ and scale $(\lambda + 1/\beta)^{-1} = \beta/(1+\beta\lambda)$. Using (18.1),

$$\begin{aligned} p(y_i; \lambda, \kappa) &= \frac{\lambda^{y_i}}{y_i!\Gamma(\alpha)\beta^\alpha} \Gamma(y_i + \alpha) \left(\frac{\beta}{1+\beta\lambda}\right)^{y_i+\alpha} \\ &= \frac{\Gamma(y_i+\alpha)}{y_i!\Gamma(\alpha)} \left(\frac{\lambda\beta}{1+\lambda\beta}\right)^{y_i} \left(\frac{1}{1+\lambda\beta}\right)^{\alpha}. \end{aligned}$$

Substituting $\alpha = 1/\kappa$ and $\beta = \kappa$ gives

$$p(y_i; \lambda, \kappa) = \frac{\Gamma(y_i + \kappa^{-1})}{y_i!\Gamma(\kappa^{-1})} \left(\frac{\lambda\kappa}{1+\lambda\kappa}\right)^{y_i} \left(\frac{1}{1+\lambda\kappa}\right)^{\kappa^{-1}}.$$

This is a negative binomial probability mass function. We use the parameterization

$$X \sim \mathrm{NegBin}(r, p), \qquad \mathbb{P}(X = x) = \frac{\Gamma(r+x)}{\Gamma(r)\Gamma(x+1)} (1-p)^x p^r,$$

for which

$$\mathbb{E}[X] = \frac{r(1-p)}{p} \qquad \text{and} \qquad \mathrm{Var}(X) = \frac{r(1-p)}{p^2}.$$

Thus

$$Y_i \sim \mathrm{NegBin}\left(\frac{1}{\kappa}, \frac{1}{1+\lambda\kappa}\right),$$

and

$$\mathbb{E}[Y_i] = \lambda \qquad \text{and} \qquad \operatorname{Var}(Y_i) = \lambda(1 + \lambda\kappa).$$

With covariates and a log link, write $\lambda_i = \exp\bigl(\boldsymbol{x}_i^\top \boldsymbol{\beta}\bigr)$. The negative binomial likelihood is then

$$L(\boldsymbol{\beta}, \kappa) = \prod_{i=1}^{n} \frac{\Gamma(y_i + \kappa^{-1})}{y_i! \Gamma(\kappa^{-1})} \left(\frac{\kappa e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}}{1 + \kappa e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \right)^{y_i} \left(\frac{1}{1 + \kappa e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}} \right)^{\kappa^{-1}}.$$

Unless κ is known, this is not an exponential family likelihood in the usual generalized linear model form. In practice, the parameters are found by iterative maximization, alternating between updating $\boldsymbol{\beta}$ and updating κ.

Example 18.1 (Univariate analysis of an epilepsy trial) Thall and Vail (1990) reported a clinical trial that followed 59 patients with epilepsy. Each patient was randomized to either standard therapy or a new drug designed to reduce the number of epileptic attacks. The calculations below use the data distributed as `MASS::epil`. For a first analysis, the response Y_{i1} is the number of attacks during the first two weeks after randomization. The explanatory variables are

$$\begin{aligned} x_{i1} &= \mathbb{1}_{\{\text{experimental treatment}\}}, \\ x_{i2} &= \text{seizure count during the eight-week baseline period}, \\ x_{i3} &= \text{age}. \end{aligned}$$

The Poisson model is

$$\log \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

The fitted Poisson coefficients are

Coefficient	Estimate	Standard error	p-value
$\hat{\beta}_0$	-0.2411	0.2733	0.3778
$\hat{\beta}_1$	-0.1189	0.0926	0.1994
$\hat{\beta}_2$	0.0257	0.0010	< 0.001
$\hat{\beta}_3$	0.0465	0.0078	< 0.001

Here β_1 is the log relative rate for treated versus control subjects, controlling for baseline seizure count and age. The Poisson fit gives

$$\hat{\beta}_1 = -0.1189 \qquad \text{and} \qquad \text{standard error}(\hat{\beta}_1) = 0.0926,$$

so the estimated relative rate is

$$\exp\left(\hat{\beta}_1\right) = \exp(-0.1189) = 0.888.$$

The naive 95% confidence interval is

$$\exp(-0.1189 \pm 1.96(0.0926)) = (0.74, 1.06).$$

However, the residual deviance is $D = 187.38$ on $59-4 = 55$ degrees of freedom, so $D \gg n-p$. The Poisson model therefore shows strong evidence of overdispersion.

The ad hoc estimate is

$$\hat{\phi} = \frac{187.38}{55} = 3.407,$$

and the adjusted standard error for the treatment coefficient is

$$\text{standard error}_{\text{adjusted}}(\hat{\beta}_1) = \sqrt{3.407}(0.0926) = 0.1710.$$

The adjusted Wald statistic is

$$\frac{|-0.1189|}{0.1710} = 0.70,$$

with corresponding two-sided p-value 0.49. The adjusted 95% confidence interval for the relative rate is

$$\exp(-0.1189 \pm 1.96(0.1710)) = (0.64, 1.24).$$

After the ad hoc adjustment, there is no statistically significant evidence of a treatment effect.

Lecture 19 — Tuesday, March 22, 2016

Epilepsy Trial Revisited

The same univariate epilepsy analysis can be fit with a negative binomial model. The fitted coefficients are

Coefficient	Estimate	Standard error	p-value
$\hat{\beta}_0$	-0.0528	0.4921	0.9146
$\hat{\beta}_1$	-0.2879	0.1833	0.1162
$\hat{\beta}_2$	0.0283	0.0031	< 0.001
$\hat{\beta}_3$	0.0393	0.0148	0.0081

The residual deviance is 63.888 on 55 degrees of freedom, and the estimated negative binomial size parameter is $\hat{\theta} = 3.281$, with standard error 0.957. In particular, the fitted treatment coefficient is

$$\hat{\beta}_1 = -0.2879, \qquad \text{standard error}(\hat{\beta}_1) = 0.1833, \qquad \text{and} \qquad \hat{\theta} = 3.281,$$

where the software uses $\theta = 1/\kappa$. Thus $\hat{\kappa} = 0.305$ for this negative binomial model. The estimated relative rate is

$$\exp\left(\hat{\beta}_1\right) = \exp(-0.2879) = 0.75,$$

with 95% confidence interval

$$\exp(-0.2879 \pm 1.96(0.1833)) = (0.52, 1.07).$$

This conclusion is consistent with the ad hoc adjustment: the treated group has a lower estimated seizure rate, but the treatment effect is not statistically significant at the 5% level.

The random-effect variance κ in the negative binomial model is not the same quantity as the ad hoc Poisson scale factor ϕ. In the ad hoc Poisson analysis,

$$\text{Var}(Y_i) = \phi\lambda_i,$$

whereas in the negative binomial model,

$$\text{Var}(Y_i) = \lambda_i(1 + \lambda_i\kappa) = \lambda_i\left(1 + \frac{\lambda_i}{\theta}\right).$$

The negative binomial standard errors already incorporate this variance structure. The coefficient estimates also need not equal the Poisson estimates, because the Poisson and negative binomial estimating equations weight the observations differently.

If the negative binomial model is fit using only the treatment indicator, then

$$\hat{\beta}_1 = -0.0866, \qquad \text{standard error}(\hat{\beta}_1) = 0.2922, \qquad \hat{\theta} = 0.874, \qquad \text{and} \qquad \hat{\kappa} = 1.144.$$

This fit has residual deviance 65.987 on 57 degrees of freedom and Akaike information criterion 388.35. Its $\hat{\kappa}$ is much larger than the value 0.305 from the model that also includes baseline seizure count and age. The difference is expected: baseline seizure count and age explain substantial between-subject variability in the seizure rate. Overdispersion can arise from omitted explanatory variables, from extra variation not explained by a Poisson model, or from dependent observations.

Theorem 19.1 (Delta method) Suppose $\sqrt{n}(X_n - \theta)$ converges in distribution to $\mathcal{N}(0, \sigma^2)$, and let g be differentiable at θ. Then

$$\sqrt{n}\{g(X_n) - g(\theta)\} \xrightarrow{\text{d}} \mathcal{N}\left(0, \sigma^2\{g'(\theta)\}^2\right).$$

Equivalently, the large-sample variance of $g(X_n)$ is approximately $\text{Var}(X_n)\{g'(\theta)\}^2$.

When inference is needed for $\kappa = 1/\theta$, the delta method is useful. For example,

$$\kappa = \frac{1}{\theta} = \exp(-\log\theta),$$

so the delta method can be used to obtain an approximate standard error and confidence interval for κ from the estimated variability of $\hat{\theta}$ or $\log\hat{\theta}$.

Clustered Count Data

The mixed Poisson construction also applies to clustered counts. Suppose cluster i contains observations Y_{ij}, $j = 1, \ldots, n_i$, and that

$$Y_{ij} \mid u_i \sim \text{Poisson}(u_i\lambda)$$

independently conditional on u_i. Observations in the same cluster are conditionally independent given u_i, while observations in different clusters are independent. If

$$\mathbb{E}[u_i] = 1 \qquad \text{and} \qquad \text{Var}(u_i) = \kappa,$$

then, as before,

$$\mathbb{E}[Y_{ij}] = \lambda \qquad \text{and} \qquad \text{Var}(Y_{ij}) = \lambda(1 + \lambda\kappa).$$

For two distinct observations $j \neq k$ in the same cluster,

$$\begin{aligned}\operatorname{Cov}(Y_{ij}, Y_{ik}) &= \operatorname{Cov}(\mathbb{E}[Y_{ij} \mid u_i], \mathbb{E}[Y_{ik} \mid u_i]) + \mathbb{E}[\operatorname{Cov}(Y_{ij}, Y_{ik} \mid u_i)] \\ &= \operatorname{Cov}(u_i\lambda, u_i\lambda) + \mathbb{E}[0] \\ &= \lambda^2\kappa.\end{aligned}$$

Thus the same random effect accounts both for overdispersion and for positive within-cluster correlation.

Example 19.2 (Joint analysis of the epilepsy trial) For the full epilepsy follow-up, each patient contributes four two-week seizure counts $Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4}$. Let

$$Y_{i\cdot} = \sum_{j=1}^{4} Y_{ij}$$

be the total number of seizures over the eight-week follow-up period. Let λ_i denote patient i's expected total over all four follow-up periods. Under the clustered mixed Poisson model with equal period-specific rates,

$$Y_{ij} \mid u_i \sim \text{Poisson}\left(\frac{u_i\lambda_i}{4}\right)$$

independently over j, so

$$Y_{i\cdot} \mid u_i \sim \text{Poisson}(u_i\lambda_i).$$

The joint marginal probability of the four counts is

$$\begin{aligned}&\mathbb{P}(Y_{i1} = y_{i1}, \ldots, Y_{i4} = y_{i4}) \\ &\quad= \int_0^\infty \prod_{j=1}^{4} f(y_{ij} \mid u_i; \lambda_i/4) g(u_i; \alpha, \beta)\, \mathrm{d}u_i \\ &\quad= \frac{\Gamma(y_{i\cdot} + \alpha)}{\Gamma(\alpha) \prod_{j=1}^{4} y_{ij}!} \left(\frac{\lambda_i\beta/4}{1 + \lambda_i\beta}\right)^{y_{i\cdot}} \left(\frac{1}{1 + \lambda_i\beta}\right)^{\alpha}.\end{aligned}$$

This factorization shows that $Y_{i\cdot}$ is sufficient for the cluster contribution to the likelihood, so the total counts can be analyzed with a negative binomial likelihood.

For the model

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3},$$

where x_{i1} is the treatment indicator, x_{i2} is the seizure count during the eight-week baseline period, and x_{i3} is age, the joint negative binomial coefficient estimates are

Coefficient	Estimate	Standard error	p-value
$\hat{\beta}_0$	1.9318	0.4084	< 0.001
$\hat{\beta}_1$	-0.1935	0.1543	0.210
$\hat{\beta}_2$	0.0277	0.0028	< 0.001
$\hat{\beta}_3$	0.0168	0.0125	0.180

The residual deviance is 63.667 on 55 degrees of freedom, and the estimated size parameter is $\hat{\theta} = 3.362$, with standard error 0.710. In particular, the joint negative binomial fit gives

$$\hat{\beta}_1 = -0.1935, \qquad \text{standard error}(\hat{\beta}_1) = 0.1543, \qquad \hat{\theta} = 3.362, \qquad \text{and} \qquad \hat{\kappa} = 0.2975.$$

The estimated relative rate over eight weeks for treated versus control subjects is

$$\exp\Big(\hat{\beta}_1\Big) = \exp(-0.1935) = 0.824,$$

with 95% confidence interval

$$\exp\Big(\hat{\beta}_1 \pm 1.96\,\text{standard error}(\hat{\beta}_1)\Big) = (0.609, 1.115).$$

The treated group has a lower estimated seizure rate, but the relative rate is not statistically significant.

For an untreated subject with baseline seizure count 11 and age 31,

$$\hat{\lambda}_i = \exp\Big(\hat{\beta}_0 + \hat{\beta}_1(0) + \hat{\beta}_2(11) + \hat{\beta}_3(31)\Big) = \exp(2.758) = 15.771.$$

Because λ_i is the expected four-period total, the model gives

$$\mathbb{E}[Y_{ij}] = \frac{\lambda_i}{4}, \qquad \text{Var}(Y_{ij}) = \frac{\lambda_i}{4}\left(1 + \frac{\lambda_i\kappa}{4}\right), \qquad \text{and} \qquad \text{Cov}(Y_{ij}, Y_{ik}) = \left(\frac{\lambda_i}{4}\right)^2 \kappa.$$

Therefore the within-subject correlation between the first and third responses is

$$\begin{aligned}\text{Corr}(Y_{i1}, Y_{i3}) &= \frac{\text{Cov}(Y_{i1}, Y_{i3})}{\{\text{Var}(Y_{i1})\,\text{Var}(Y_{i3})\}^{1/2}} \\ &= \frac{(\lambda_i/4)^2\kappa}{(\lambda_i/4)\{1 + (\lambda_i/4)\kappa\}}\end{aligned}$$

$$= \frac{(\lambda_i/4)\kappa}{1 + (\lambda_i/4)\kappa}.$$

Substituting the estimates gives

$$\widehat{\text{Corr}}(Y_{i1}, Y_{i3}) = \frac{(15.771/4)(0.2975)}{1 + (15.771/4)(0.2975)} = 0.540.$$

This indicates substantial positive correlation among observations from the same individual.

Lecture 20 — Thursday, March 24, 2016

Binomial Overdispersion

The same two broad approaches, ad hoc scaling and mixed modelling, can be used for binomial overdispersion. For a binomial response with m trials and response probability π, the usual model has

$$\mathbb{E}[Y] = m\pi \qquad \text{and} \qquad \text{Var}(Y) = m\pi(1-\pi).$$

Extra binomial variation often arises from unmodelled clustering, because the binomial model assumes independent and identically distributed binary outcomes. Examples of clusters include families, classes, neighbourhoods, litters, and repeated measurements on individuals. For example, if a class of 100 students is divided into study groups of size 4, then there are 25 clusters. Students in the same study group may have correlated outcomes, so the usual binomial independence assumptions may fail.

Suppose a population consists of independent clusters of size k, where k divides m, and suppose m individuals are sampled from m/k clusters. Let

$$Y_{ij} = \begin{cases} 1, & \text{if individual } j \text{ in cluster } i \text{ has the response,} \\ 0, & \text{otherwise,} \end{cases}$$

and define

$$Y_{i\cdot} = \sum_{j=1}^{k} Y_{ij} \qquad \text{and} \qquad Y_{\cdot\cdot} = \sum_{i=1}^{m/k} \sum_{j=1}^{k} Y_{ij}.$$

Overdispersion is induced by allowing each cluster to have its own response probability:

$$Y_{ij} \mid \pi_i \sim \text{Bin}(1, \pi_i) \quad \text{independently over } j,$$

and therefore

$$Y_{i\cdot} \mid \pi_i \sim \text{Bin}(k, \pi_i).$$

Assume the random probabilities π_i are iid across clusters and

$$\mathbb{E}[\pi_i] = \pi \qquad \text{and} \qquad \text{Var}(\pi_i) = \rho\pi(1-\pi).$$

Here π_i is the cluster-level random effect, $0 < \pi < 1$ is the overall response probability, and $0 \leq \rho \leq 1$ will be the intraclass correlation coefficient.

At the individual level,

$$\mathbb{E}[Y_{ij}] = \mathbb{E}[\mathbb{E}[Y_{ij} \mid \pi_i]] = \mathbb{E}[\pi_i] = \pi.$$

The marginal variance remains the Bernoulli variance:

$$\begin{aligned}
\text{Var}(Y_{ij}) &= \mathbb{E}[\text{Var}(Y_{ij} \mid \pi_i)] + \text{Var}(\mathbb{E}[Y_{ij} \mid \pi_i]) \\
&= \mathbb{E}[\pi_i(1-\pi_i)] + \text{Var}(\pi_i) \\
&= \mathbb{E}[\pi_i] - \mathbb{E}[\pi_i^2] + \mathbb{E}[\pi_i^2] - \{\mathbb{E}[\pi_i]\}^2 \\
&= \pi(1-\pi).
\end{aligned}$$

However, two different individuals in the same cluster are correlated:

$$\begin{aligned}
\text{Cov}(Y_{ij}, Y_{ik}) &= \mathbb{E}[\text{Cov}(Y_{ij}, Y_{ik} \mid \pi_i)] + \text{Cov}(\mathbb{E}[Y_{ij} \mid \pi_i], \mathbb{E}[Y_{ik} \mid \pi_i]) \\
&= \mathbb{E}[0] + \text{Cov}(\pi_i, \pi_i) \\
&= \text{Var}(\pi_i) \\
&= \rho\pi(1-\pi).
\end{aligned}$$

Therefore

$$\text{Corr}(Y_{ij}, Y_{ik}) = \frac{\rho\pi(1-\pi)}{\pi(1-\pi)} = \rho.$$

At the cluster level,

$$\mathbb{E}[Y_{i\cdot}] = \mathbb{E}[\mathbb{E}[Y_{i\cdot} \mid \pi_i]] = \mathbb{E}[k\pi_i] = k\pi.$$

The variance of the cluster count is

$$\begin{aligned}
\text{Var}(Y_{i\cdot}) &= \mathbb{E}[\text{Var}(Y_{i\cdot} \mid \pi_i)] + \text{Var}(\mathbb{E}[Y_{i\cdot} \mid \pi_i]) \\
&= \mathbb{E}[k\pi_i(1-\pi_i)] + \text{Var}(k\pi_i)
\end{aligned}$$

$$
\begin{aligned}
&= k\mathbb{E}[\pi_i] - k\mathbb{E}[\pi_i^2] + k^2 \operatorname{Var}(\pi_i) \\
&= k\mathbb{E}[\pi_i] - k\{\operatorname{Var}(\pi_i) + \mathbb{E}[\pi_i]^2\} + k^2 \operatorname{Var}(\pi_i) \\
&= k(k-1)\rho\pi(1-\pi) + k\pi(1-\pi) \\
&= k\pi(1-\pi)\{(k-1)\rho + 1\}.
\end{aligned}
$$

The usual binomial variance $k\pi(1-\pi)$ is multiplied by the dispersion factor

$$\sigma^2 = (k-1)\rho + 1.$$

At the level of the grand total,

$$\mathbb{E}[Y_{\cdot\cdot}] = \sum_{i=1}^{m/k} \mathbb{E}[Y_{i\cdot}] = \left(\frac{m}{k}\right) k\pi = m\pi,$$

and, because the clusters are independent,

$$\operatorname{Var}(Y_{\cdot\cdot}) = \sum_{i=1}^{m/k} \operatorname{Var}(Y_{i\cdot}) = \left(\frac{m}{k}\right) k\pi(1-\pi)\sigma^2 = m\pi(1-\pi)\sigma^2.$$

Methods for Binomial Overdispersion

An ad hoc binomial adjustment estimates

$$\hat{\sigma}^2 = \frac{D}{n-p}$$

when the deviance is much larger than its degrees of freedom. This is most natural when clusters have equal size, because the cluster-level variance inflation is

$$\sigma^2 = (k-1)\rho + 1.$$

With unequal cluster sizes, a common scale factor is generally inefficient, so a mixture model is usually preferable.

For a binomial mixture model, let

$$\pi_i \sim \text{Beta}(\gamma_1, \gamma_2), \quad \gamma_1 > 0, \quad \gamma_2 > 0.$$

The density is

$$g(\pi_i) = \frac{\Gamma(\gamma_1+\gamma_2)}{\Gamma(\gamma_1)\Gamma(\gamma_2)}\pi_i^{\gamma_1-1}(1-\pi_i)^{\gamma_2-1} = \frac{1}{B(\gamma_1,\gamma_2)}\pi_i^{\gamma_1-1}(1-\pi_i)^{\gamma_2-1}.$$

The beta mean and variance are

$$\mathbb{E}[\pi_i] = \frac{\gamma_1}{\gamma_1+\gamma_2} \qquad \text{and} \qquad \text{Var}(\pi_i) = \frac{\gamma_1\gamma_2}{(\gamma_1+\gamma_2)^2(1+\gamma_1+\gamma_2)}.$$

Matching these to $\mathbb{E}[\pi_i] = \pi$ and $\text{Var}(\pi_i) = \rho\pi(1-\pi)$ gives

$$\pi = \frac{\gamma_1}{\gamma_1+\gamma_2} \qquad \text{and} \qquad \rho = \frac{1}{1+\gamma_1+\gamma_2}.$$

Allowing unequal cluster sizes, suppose

$$Y_{i\cdot} \mid \pi_i \sim \text{Bin}(k_i, \pi_i).$$

The marginal distribution of the cluster count is

$$\begin{aligned}
\mathbb{P}(Y_{i\cdot} = y_{i\cdot}) &= \int_0^1 \mathbb{P}(Y_{i\cdot} = y_{i\cdot} \mid \pi_i) g(\pi_i;\gamma_1,\gamma_2)\,\mathrm{d}\pi_i \\
&= \int_0^1 \binom{k_i}{y_{i\cdot}} \pi_i^{y_{i\cdot}}(1-\pi_i)^{k_i-y_{i\cdot}} \frac{1}{B(\gamma_1,\gamma_2)}\pi_i^{\gamma_1-1}(1-\pi_i)^{\gamma_2-1}\,\mathrm{d}\pi_i \\
&= \binom{k_i}{y_{i\cdot}} \frac{1}{B(\gamma_1,\gamma_2)} \int_0^1 \pi_i^{y_{i\cdot}+\gamma_1-1}(1-\pi_i)^{k_i-y_{i\cdot}+\gamma_2-1}\,\mathrm{d}\pi_i \\
&= \binom{k_i}{y_{i\cdot}} \frac{B(y_{i\cdot}+\gamma_1, k_i-y_{i\cdot}+\gamma_2)}{B(\gamma_1,\gamma_2)}.
\end{aligned}$$

This is the beta-binomial distribution. Its mean and variance are

$$\mathbb{E}[Y_{i\cdot}] = k_i\pi,$$

and

$$\text{Var}(Y_{i\cdot}) = k_i\pi(1-\pi)\{1+(k_i-1)\rho\}.$$

If $\theta = 1/(\gamma_1+\gamma_2)$, then

$$\rho = \frac{\theta}{1+\theta},$$

so the variance can also be written as

$$\text{Var}(Y_{i\cdot}) = k_i\pi(1-\pi)\left(\frac{k_i\theta+1}{\theta+1}\right).$$

As $\theta \to 0$, this variance converges to the usual binomial variance $k_i\pi(1-\pi)$. Thus the binomial model is nested in the beta-binomial model, and testing $H_0 : \theta = 0$ is a test for overdispersion. Because $\theta = 0$ is on the boundary of the beta-binomial parameter space, the ordinary interior chi-square likelihood ratio reference law does not apply without modification. Maximum likelihood fitting of the beta-binomial model can also be difficult because of the nonlinear Gamma functions in the likelihood, motivating quasi-likelihood methods.

Lecture 21 — Tuesday, March 29, 2016

Quasi-Likelihood

Parametric regression models specify a full distribution for the responses, such as normal, binomial, Poisson, beta-binomial, or negative binomial. They also specify a linear predictor $\boldsymbol{x}_i^\top\boldsymbol{\beta}$, a link function $g(\mu_i) = \boldsymbol{x}_i^\top\boldsymbol{\beta}$, and then use likelihood theory to estimate and conduct inference for $\boldsymbol{\beta}$. This framework is powerful, but it relies on the assumed response distribution being correct.

Quasi-likelihood relaxes the full distributional assumption. Instead, assume the independent responses $Y_1, \ldots, Y_n$ satisfy

$$\mathbb{E}[Y_i] = \mu_i(\boldsymbol{\beta}), \qquad \mathrm{Var}(Y_i) = \phi V(\mu_i(\boldsymbol{\beta})), \qquad \text{and} \qquad g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top\boldsymbol{\beta}.$$

The function $V(\mu_i)$ is the variance function, and no complete parametric probability model is required. The resulting estimators need not be maximum likelihood estimators, because the chosen relationship between mean and variance may not correspond to a genuine probability distribution.

The covariance matrix of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ is

$$\phi\boldsymbol{V}(\boldsymbol{\mu}) = \phi \begin{pmatrix} V(\mu_1) & & 0 \\ & \ddots & \\ 0 & & V(\mu_n) \end{pmatrix}.$$

Let $\boldsymbol{D}$ be the $n \times p$ derivative matrix with entries

$$D_{ir} = \frac{\partial \mu_i}{\partial \beta_r}, \quad i = 1, \ldots, n, \quad r = 0, \ldots, p-1.$$

Equivalently, the ith row of $\boldsymbol{D}$ is

$$\boldsymbol{D}_i = \left(\frac{\partial \mu_i}{\partial \beta_0}, \ldots, \frac{\partial \mu_i}{\partial \beta_{p-1}} \right).$$

To motivate the quasi-score, recall first the usual score identities from likelihood theory. If $S(\theta)$ is the score for a regular one-parameter model, then

$$\mathbb{E}[S(\theta)] = 0, \qquad \mathbb{E}[S(\theta)^2] = I(\theta), \qquad \text{and} \qquad \text{Var}(S(\theta)) = I(\theta),$$

where $I(\theta)$ is the expected information. For an exponential family with natural parameter θ, this becomes

$$\text{Var}(S(\theta)) = \mathbb{E}[S(\theta)^2] = -\mathbb{E}\left[\frac{\partial}{\partial \theta} S(\theta) \right] = \frac{b''(\theta)}{a(\phi)}.$$

Now consider an exponential family for one observation,

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

with

$$\mathbb{E}[Y] = b'(\theta) = \mu \qquad \text{and} \qquad \text{Var}(Y) = a(\phi) b''(\theta) = a(\phi) V(\mu).$$

The score in θ is

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

Writing the score in terms of μ,

$$s(\mu) = \frac{\partial \ell}{\partial \theta} \frac{\partial \theta}{\partial \mu} = \left(\frac{y - b'(\theta)}{a(\phi)} \right) \left(\frac{\partial \mu}{\partial \theta} \right)^{-1} = \frac{y - \mu}{a(\phi) V(\mu)}.$$

This suggests the quasi-score component

$$U = \frac{Y - \mu}{\phi V(\mu)}$$

when only the mean and variance are specified. The function U has the key score-like properties:

1. Its expectation is zero:
$$\mathbb{E}[U] = \frac{1}{\phi V(\mu)} \mathbb{E}[Y - \mu] = 0.$$

2. Its variance is
$$\text{Var}(U) = \mathbb{E}[U^2] = \left(\frac{1}{\phi V(\mu)}\right)^2 \mathbb{E}[(Y-\mu)^2] = \frac{1}{\phi V(\mu)}.$$

3. Its expected negative derivative is
$$\begin{aligned} -\mathbb{E}\left[\frac{\partial U}{\partial \mu}\right] &= -\mathbb{E}\left[\frac{[-1]\phi V(\mu) - [\phi V'(\mu)](Y-\mu)}{\{\phi V(\mu)\}^2}\right] \\ &= \frac{1}{\phi V(\mu)} + \frac{\phi V'(\mu)}{\{\phi V(\mu)\}^2}\mathbb{E}[Y-\mu] \\ &= \frac{1}{\phi V(\mu)}. \end{aligned}$$

Hence $\text{Var}(U) = \mathbb{E}[U^2] = -\mathbb{E}[\partial U/\partial \mu]$, as for an ordinary likelihood score.

If $U(\mu)$ were a true score function for μ, then it would be the derivative of a log likelihood. This motivates defining the quasi-log likelihood for one observation by
$$Q(\mu; y) = \int_y^\mu \frac{y-t}{\phi V(t)}\,\mathrm{d}t.$$

For independent observations,
$$Q(\mu; y) = \sum_{i=1}^n Q_i(\mu_i; y_i) = \sum_{i=1}^n \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)}\,\mathrm{d}t.$$

The corresponding quasi-deviance contribution is
$$\begin{aligned} \Delta Q_i(y_i; \mu_i) &= 2\{Q_i(\tilde{\mu}_i; y_i) - Q_i(\mu_i; y_i)\} \\ &= -2\int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)}\,\mathrm{d}t \\ &= 2\int_{\mu_i}^{y_i} \frac{y_i - t}{\phi V(t)}\,\mathrm{d}t, \end{aligned}$$

where $\tilde{\mu}_i = y_i$ is the fitted mean under the saturated model. The total quasi-deviance is
$$\Delta Q(y; \mu) = \sum_{i=1}^n \Delta Q_i(y_i; \mu_i).$$

Differentiating Q with respect to β_j gives the jth quasi-score component

$$\frac{\partial Q(\mu; y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial Q_i(\mu_i; y_i)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \beta_j}.$$

In matrix form,

$$\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{D}^\top \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})/\phi.$$

The estimate $\hat{\boldsymbol{\beta}}$ is obtained by solving $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{0}$, typically by Newton–Raphson or another root-finding method. The information matrix is

$$\boldsymbol{I}(\boldsymbol{\beta}) = \boldsymbol{D}^\top \boldsymbol{V}^{-1} \boldsymbol{D}/\phi,$$

so the corresponding large-sample covariance approximation is

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) \simeq \hat{\phi} (\boldsymbol{D}^\top \hat{\boldsymbol{V}}^{-1} \boldsymbol{D})^{-1}.$$

The usual estimate of ϕ is based on the Pearson statistic:

$$\hat{\phi} = \frac{1}{n - p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

The numerator is the Pearson chi-squared statistic for goodness of fit. Thus quasi-likelihood provides estimating equations that do not require a full probability model, and it allows regression models with variance functions beyond those arising from the standard exponential family distributions.

Example 21.1 (Epilepsy trial with quasi-Poisson regression) For the univariate epilepsy analysis, the quasi-Poisson model uses

$$\mathbb{E}[Y_i] = \mu_i, \qquad \mathrm{Var}(Y_i) = \phi \mu_i, \qquad \text{and} \qquad \log \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}.$$

The fitted regression coefficients are the same as in the Poisson model, because the Poisson and quasi-Poisson score equations for β have the same root. The standard errors differ because the quasi-Poisson model estimates ϕ.

In `glm`, this fit can be expressed as a quasi model with log link and variance function $V(\mu) = \mu$. Other common quasi families use the identity link with constant variance, the logit link with

binomial variance $V(\mu) = \mu(1-\mu)$, or the log link with Poisson variance $V(\mu) = \mu$.

The fitted treatment coefficient is

$$\hat{\beta}_1 = -0.1189, \qquad \text{standard error}(\hat{\beta}_1) = 0.1764, \qquad \text{and} \qquad \hat{\phi} = 3.627.$$

The comparison of fitted models is

Model	Parameter	Estimate	Standard error	Adjusted standard error
Poisson	β_1	-0.1189	0.0926	0.1710
	ϕ	3.407		
Negative binomial	β_1	-0.2879	0.1833	
	θ	3.281	0.957	
Quasi-Poisson	β_1	-0.1189	0.1764	
	ϕ	3.627		

This estimated quasi-Poisson dispersion is comparable to the ad hoc estimate 3.407, and the quasi-Poisson standard error is comparable to both the ad hoc adjusted Poisson standard error and the negative binomial standard error. The treatment effect is not statistically significant under the quasi-Poisson model.

Example 21.2 (Pacific cod hatching data) An experiment studied the effects of salinity, temperature, and oxygen concentration on the probability that Pacific cod eggs hatch. These factors were varied over practical ranges, a known number of eggs was placed in each of four tanks at each factor setting, and the number of eggs that hatched was recorded. This gives four binomial samples for each experimental configuration.

The data are organized by factor configuration. In the following displayed portion of the data, each hatch/total entry is one tank:

Salinity	Temperature	Oxygen	Tank 1	Tank 2	Tank 3	Tank 4
14.00	2.70	3.60	224/283	180/245	160/235	182/320
14.00	2.70	8.60	231/325	237/283	171/207	178/270
14.00	9.30	3.60	159/240	163/229	234/349	295/385
14.00	9.30	8.60	186/314	214/297	97/298	74/244
26.00	2.70	3.60	5/217	5/316	2/243	3/224
26.00	2.70	8.60	143/292	186/316	159/301	138/264
26.00	9.30	3.60	19/262	18/263	36/277	44/290
26.00	9.30	8.60	74/293	152/307	68/181	45/167

A logistic regression model with all two-way interactions,

$$\begin{aligned}\text{logit}(\pi_i) = \beta_0 &+ \beta_1 \text{salinity}_i + \beta_2 \text{temperature}_i + \beta_3 \text{oxygen}_i \\ &+ \beta_4(\text{salinity}_i \text{temperature}_i) + \beta_5(\text{temperature}_i \text{oxygen}_i) + \beta_6(\text{salinity}_i \text{oxygen}_i),\end{aligned}$$

gives residual deviance

$$D = 4308.2$$

on 69 degrees of freedom, with null deviance 8209.9 on 75 degrees of freedom and Akaike information criterion 4746. The coefficient summary is

Term	Estimate	Standard error	Test statistic	p-value
Intercept	5.339730	0.279098	19.132	$< 2 \times 10^{-16}$
Salinity	−0.388016	0.013975	−27.764	$< 2 \times 10^{-16}$
Temperature	0.083883	0.030418	2.758	0.005821
Oxygen	−0.127545	0.035243	−3.619	0.000296
Salinity by temperature	0.009308	0.001277	7.288	3.15×10^{-13}
Temperature by oxygen	−0.043843	0.003054	−14.355	$< 2 \times 10^{-16}$
Salinity by oxygen	0.028164	0.001603	17.571	$< 2 \times 10^{-16}$

Every coefficient appears statistically significant under the ordinary logistic model. However, $D \gg n - p$, so the ordinary logistic model severely underestimates the variability.

The ad hoc dispersion estimate is

$$\hat{\phi} = \frac{D}{n-p} = \frac{4308.2}{69} = 62.44,$$

so standard errors are multiplied by

$$\sqrt{\hat{\phi}} = 7.90.$$

For example, the unadjusted 95% confidence interval for β_4 is

$$(0.0068, 0.0118),$$

whereas the adjusted interval is

$$(-0.0105, 0.0291).$$

A beta-binomial regression model accounts for overdispersion through a random cluster probability. Writing its intraclass correlation parameter as ρ, the variance structure is

$$\mathrm{Var}(Y_{i\cdot}) = k_i \pi_i (1 - \pi_i)\{1 + (k_i - 1)\rho\}.$$

For the full two-way interaction model, the beta-binomial fit gives

Term	Estimate	Standard error	Test statistic	p-value
Intercept	5.645924	2.050995	2.753	0.00591
Salinity	−0.421677	0.105007	−4.016	5.93×10^{-5}
Temperature	0.152677	0.228123	0.669	0.50332
Oxygen	−0.223837	0.264152	−0.847	0.39678
Salinity × temperature	0.008644	0.009475	0.912	0.36162
Temperature × oxygen	−0.049173	0.022930	−2.144	0.03200
Salinity × oxygen	0.034040	0.012323	2.762	0.00574

The residual deviance is 75.03 on 69 degrees of freedom, the Akaike information criterion is 96.17, and the estimated intraclass correlation parameter is

$$\hat{\rho} = 0.1982644.$$

This is not the same parameter as the ad hoc scale factor. Under the beta-binomial fit, the coefficient estimates and their standard errors both change because the overdispersion is part of the probability model.

A quasi-binomial model instead keeps the binomial mean structure but uses a scaled variance. For the full two-way interaction model it gives

Term	Estimate	Standard error	Test statistic	p-value
Intercept	5.339730	2.134676	2.501	0.01475
Salinity	−0.388016	0.106890	−3.630	0.00054
Temperature	0.083883	0.232652	0.361	0.71954
Oxygen	−0.127545	0.269557	−0.473	0.63759
Salinity × temperature	0.009308	0.009769	0.953	0.34399
Temperature × oxygen	−0.043843	0.023359	−1.877	0.06476
Salinity × oxygen	0.028164	0.012260	2.297	0.02464

The quasi-binomial dispersion estimate is

$$\hat{\phi} = 58.49932,$$

which is comparable to the ad hoc estimate 62.44. The residual deviance remains 4308.2 on 69 degrees of freedom, and the Akaike information criterion is not available because this is not a full likelihood model.

The salinity by temperature interaction is not significant in either overdispersion-aware full model, so the reduced model keeps the temperature by oxygen and salinity by oxygen interactions:

$$\begin{aligned}\text{logit}(\pi_i) = \beta_0 &+ \beta_1 \text{temperature}_i + \beta_2 \text{oxygen}_i + \beta_3 \text{salinity}_i \\ &+ \beta_4(\text{temperature}_i \text{oxygen}_i) + \beta_5(\text{oxygen}_i \text{salinity}_i).\end{aligned}$$

For this reduced model, the beta-binomial fit gives

Term	Estimate	Standard error	Test statistic	p-value
Intercept	4.69472	1.74939	2.684	0.00728
Temperature	0.30839	0.15378	2.005	0.04492
Oxygen	−0.24170	0.26184	−0.923	0.35595
Salinity	−0.36705	0.08472	−4.333	1.47×10^{-5}
Temperature × oxygen	−0.04647	0.02272	−2.046	0.04081
Oxygen × salinity	0.03388	0.01231	2.752	0.00593

The residual deviance is 75.824 on 70 degrees of freedom, the Akaike information criterion is 94.961, and the estimated intraclass correlation parameter is

$$\hat{\rho} = 0.1983778.$$

For the same reduced model, the quasi-binomial fit gives

Term	Estimate	Standard error	Test statistic	p-value
Intercept	4.25598	1.78228	2.388	0.019645
Temperature	0.25120	0.15428	1.628	0.107967
Oxygen	−0.13773	0.26739	−0.515	0.608116
Salinity	−0.32661	0.08341	−3.916	0.000207
Temperature × oxygen	−0.04085	0.02312	−1.767	0.081577
Oxygen × salinity	0.02756	0.01221	2.257	0.027156

Here the residual deviance is 4361.6 on 70 degrees of freedom, and

$$\hat{\phi} = 58.4616.$$

All three overdispersion adjustments lead to the same broad conclusion: several effects that appeared highly significant under the ordinary logistic regression model are no longer significant once the extra binomial variation is accounted for.

Appendix: Lecture and Tutorial Index