Binomial Distribution:

$$Y_1, \ldots, Y_n \overset{ind}{\sim} Bin(n_i, \pi_i),\ \mu_i = n_i\pi_i,$$

$$\Rightarrow \quad \ell(\boldsymbol{\pi}, \mathbf{y}) = \sum_i \left\{ \log \binom{n_i}{y_i} + y_i \log \pi_i + (n_i - y_i) \log(1 - \pi_i) \right\}$$

$$\Rightarrow \quad \ell(\boldsymbol{\mu}; \mathbf{y}) = \sum_i \left\{ \log \binom{n_i}{y_i} + y_i \log \left(\frac{\mu_i}{n_i} \right) + (n_i - y_i) \log \left(\frac{n_i - \mu_i}{n_i} \right) \right\}$$

$$\Rightarrow \quad D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2 \sum_i \left\{ \log \binom{n_i}{y_i} + y_i \log \left(\frac{y_i}{n_i} \right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i} \right) \right.$$
$$\left. - \left[\log \binom{n_i}{y_i} + y_i \log \left(\frac{\hat{\mu}_i}{n_i} \right) + (n_i - y_i) \log \left(\frac{n_i - \hat{\mu}_i}{n_i} \right) \right] \right\}$$

$$\Rightarrow \quad D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2 \sum_i \left\{ y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - \hat{\mu}_i} \right) \right\}$$

In general, for a sample from an E.D. family density

$$\begin{aligned} D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) &= \sum_i 2w_i \left\{ y_i(\tilde{\theta}_i - \hat{\theta}_i) - \left(b(\tilde{\theta}_i) - b(\hat{\theta}_i) \right) \right\} / \phi \\ &= D(\mathbf{y}; \hat{\boldsymbol{\mu}}) / \phi \end{aligned}$$

where $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ is the (unscaled) deviance, which is a function of the data only (it is free of ϕ), and $\tilde{\theta}_i = \theta_i(y_i)$, $\hat{\theta}_i = \theta_i(\hat{\mu}_i)$.

Asymptotic Distribution of the Deviance and X^2 Statistics:

From the asymptotic chi-square-ness of $2\log\lambda$ its tempting to conclude

$$D^*(\mathbf{y}, \hat{\boldsymbol{\mu}}) \overset{a}{\sim} \chi^2(n-p)$$

This is not necessarily true!

- Wilks' result is based on asymptotics under which $n \to \infty$ while t_1, t_2 stay fixed.
- However, in some cases the saturated model requires estimation of an increasing number of parameters as $n \to \infty$, so that standard theory does not apply. Therefore, we need to be careful in using the chi-square approximation to the distribution of the deviance, and in using the deviance as a true goodness-of-fit statistic.
- Similar comments apply to Pearson's generalized X^2 statistic:

$$X^2(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/w_i}$$

- Under some grouped data situations, both X^2 and D^* are asymptotically chi-square.

We assume here that data have been grouped as far as possible and that X^2 and D^* are computed as sums of independent contributions over g groups:

$$D^*(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{g} \{\ell_i(y_i; y_i) - \ell_i(\hat{\mu}_i; y_i)\}$$

$$X^2(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{g} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/w_i}$$

"Fixed-cells" Asymptotics:

Classical assumptions in asymptotic theory for grouped data imply

i. a fixed number of groups

ii. increasing sample sizes $n_i \to \infty$, $i = 1, \ldots, g$, such that $n_i/n \to \lambda_i$ where $\lambda_i > 0$, $i = 1, \ldots, g$, are fixed proportions

iii. a fixed "number of cells" k in each group (think of a discrete outcome variable in each group with k possible values)

iv. a fixed number of parameters being estimated in the "current model"

Under these assumptions and certain regularity conditions, then both X^2 and D^* are asymptotically chi-square under the null hypothesis that the current model holds with

$$X^2, D^* \overset{a}{\sim} \chi^2(t_1 - p)$$

where

$t_1 = g(k-1)$

= the number of independent parameters estimated under the full model

- This result holds for a more general class of goodness-of-fit statistic, **the power-divergence family**, within which X^2 and D^* are special cases (see F&T, sec. 3.4).

Sparseness and "Increasing-Cells" Asymptotics:

Several authors have considered the "increasing-cells" case, where as $n \to \infty$, $g \to \infty$ as well. (see F&T, sec. 3.4, for a review).

- Such asymptotics are more appropriate for many example encountered in practice where the number of covariate classes (e.g., the dimension of a contingency table) increases as the sample size grows.

- In this case, empty and small-count cells proliferate - hence the term sparseness asymptotics.

The essential result under sparseness is that power-divergence family g.o.f. statistics are no longer asymptotically chi-square distributed, and the special cases, X^2 and D^* are no longer asymptotically equivalent. Instead, both X^2 and D^* are asymptotically normal, with differing mean and variance parameters.

- In such situations, X^2 and D are not appropriate as goodness of fit statistics.

- See F&T, sec. 3.4 and the book by Read and Cressie (*Goodness-of-Fit Statistics for Discrete Multivariate Data*) for more information and the conditions under which these results hold.

Digression: Estimation of ϕ

For some ED distributions, $\phi = 1$, so no estimation is required. Otherwise, a Method of Moments (MOM) estimator can be used. Recall that we obtain a MOM estimator by setting the sample moment equal to population moment.

According to a GLM,

$$\text{var}(y_i - \mu_i) = \phi v(\mu_i)/w_i$$

or

$$\text{Population } 2^{\text{nd}} \text{ moment:} \quad \text{var}\left(\frac{y_i - \mu_i}{\sqrt{v(\mu_i)/w_i}}\right) = \phi$$

The corresponding sample moment is

$$\text{Sample } 2^{\text{nd}} \text{ moment:} \quad \hat{\text{var}}\left(\frac{y_i - \mu_i}{\sqrt{v(\mu_i)/w_i}}\right) = \frac{1}{n}\sum_i \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/w_i}$$

So a MOM estimator is given by

$$\hat{\phi}^* = \frac{1}{n}\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/w_i}$$

This estimator is typically modified slightly with a bias-correction:

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)/w_i}$$

- The division by $n - p$ rather than n is akin to using the sample variance $s^2 = \frac{1}{n-1}\sum_i (y_i - \hat{\mu})^2$ rather than the MLE $\frac{1}{n}\sum_i (y_i - \hat{\mu})^2$. s^2 is unbiased while the MLE is not. The subtraction of 1 from the denominator is an adjustment for having to estimate 1 parameter (μ). Here, we subtract p because we must estimate p parameters, $\beta_1, \ldots, \beta_p$.

- For grouped data the MOM estimator takes the form

$$\hat{\phi} = \frac{1}{g-p}\sum_{i=1}^{g}\frac{(y_i-\hat{\mu}_i)^2}{v(\hat{\mu}_i)/w_i} = \frac{X^2}{g-p} \quad \begin{matrix}\text{MOM}\\ \text{estimator}\end{matrix}$$

Alternatively, ϕ is sometimes estimated based on the deviance, replacing X^2 above with D:

$$\tilde{\phi} = \frac{D(\mathbf{y};\hat{\boldsymbol{\mu}})}{n-p} \quad \text{or} \quad \tilde{\phi} = \frac{D(\mathbf{y};\hat{\boldsymbol{\mu}})}{g-p} \quad \begin{matrix}\text{Deviance}\\ \text{estimator}\end{matrix}$$

(ungrouped data) (grouped data)

- A third approach is to estimate ϕ using ML, just as we do $\boldsymbol{\beta}$.

Back to Testing:

As we've seen, a test of

$$H_0 : \boldsymbol{\beta}_{p-q} = \mathbf{0} \qquad \text{versus} \qquad H_1 : \boldsymbol{\beta}_{p-q} \neq \mathbf{0}$$

in a model with $p-$dimensional regression parameter $\boldsymbol{\beta}_p$ can be based on

$$D_q^* - D_p^* \overset{a}{\sim} \chi^2(p-q)$$

and this test is equivalent to a LR Test.

One drawback to the LR approach is that it requires estimation of the model under both the null and alternative hypotheses. Alternative tests that only require estimation of one or the other model are the **Wald** and **Score Tests.**

Let $\boldsymbol{\beta}$ be of dimension p. For the general hypothesis,

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b} \qquad \text{versus} \qquad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{b},$$

where C is $s \times p$, the LR test statistic is

$$2 \log \lambda = 2 \left[\ell(\tilde{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}}) \right] \quad \begin{array}{l} \tilde{\boldsymbol{\beta}} = \text{unrestricted MLE} \\ \hat{\boldsymbol{\beta}} = \text{MLE under restrictions imposed by } H_0 \end{array}$$

The Wald test statistic is

$$(\mathbf{C}\tilde{\boldsymbol{\beta}} - \mathbf{b})^T \left[\overbrace{\mathbf{C} \underbrace{\mathbf{I}_n(\tilde{\boldsymbol{\beta}})^{-1}}} \mathbf{C}^T \right]^{-1} (\mathbf{C}\tilde{\boldsymbol{\beta}} - \mathbf{b}) \overset{a}{\sim} \chi^2(s)$$

- Notice that in the simple situation of testing $H_0 : \beta_j = 0$, the Wald statistic is just the square of $\frac{\tilde{\beta}_j}{\text{a.s.e.}(\tilde{\beta}_j)}$.
- The Wald test only requires estimation of the unrestricted model.

The score test statistic is

$$\left[\frac{\partial \ell}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) \right]^T \mathbf{I}_n(\hat{\boldsymbol{\beta}})^{-1} \left[\frac{\partial \ell}{\partial \boldsymbol{\beta}}(\hat{\boldsymbol{\beta}}) \right] \overset{a}{\sim} \chi^2(s)$$

- The score test is sometimes called the **Lagrange Multiplier Test** especially in the econometric literature.
- The score test only requires fitting the restricted model.

Comments:

1. Wald and score statistics are quadratic approximations to the LR test statistic. When loglikelihood is quadratic (e.g., normal distribution) all 3 statistics are equal: Wald=score=LR.

2. For finite samples the quality of the χ^2 approximation to the distribution depends on n, but also on the form of the log-likelihood function. This is true for all 3 statistics, but especially for the Wald and score statistics.

3. Score and Wald tests depend on likelihood only through first and second moments. Therefore, for models with an overdispersion parameter not implied by the assumed error distribution (i.e., quasilikelihood models), the Wald and score statistics are properly defined; LR statistic is not.

4. LRT and score tests are invariant to parameterization; Wald test is not.

5. For score, Wald statistics, observed information matrix or estimated expected information matrix may be substituted for the expected information matrix producing asymptotically equivalent test statistics.

Confidence Intervals:

Convenient approximate $100(1-\alpha)\%$ C.I.s for β_j can be formed as

$$\tilde{\beta}_j \pm z_{1-\alpha/2}\sqrt{i_{jj}}$$

where i_{jj} is the j^{th} diagonal element of an approximate variance-covariance matrix for $\tilde{\boldsymbol{\beta}}$, the MLE of $\boldsymbol{\beta}$. i_{jj} could be obtained from (expected info)$^{-1}$, (observed info)$^{-1}$, or (estimated expected info)$^{-1}$.

- This is a Wald confidence interval since it corresponds to inversion of the Wald test.
- We often can do better than Wald intervals (i.e., more precise intervals with better coverage properties) by inverting the LR test. We will use such **likelihood-based confidence intervals** in applications when available.

Tests in the Presence of Nuisance Parameters:

$$\text{Let} \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \quad \text{where} \quad \begin{array}{l} \boldsymbol{\beta}_1 \text{ is } (p-q)\times 1 \\ \boldsymbol{\beta}_2 \text{ is } q \times 1 \end{array}$$

and partition the design matrix accordingly: $\mathbf{X} = (\underbrace{\mathbf{X}_1}_{n\times(p-q)}, \underbrace{\mathbf{X}_2}_{n\times q})$, so that

$$\mathbf{X}\boldsymbol{\beta} = (\,\mathbf{X}_1 \quad \mathbf{X}_2\,)\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}.$$

Null hypothesis: $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$. $\boldsymbol{\beta}_1$ is a nuisance parameter.

Score Vector:

$$U(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{\beta}_1} \ell(\boldsymbol{\beta}) \\ \frac{\partial}{\partial \boldsymbol{\beta}_2} \ell(\boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} U_1(\boldsymbol{\beta}) \\ U_2(\boldsymbol{\beta}) \end{pmatrix}$$

Likelihood Equations:

$$\text{solve} \quad U \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \mathbf{0} \quad \Rightarrow \text{unrestricted MLEs } \tilde{\boldsymbol{\beta}}_1, \tilde{\boldsymbol{\beta}}_2$$

$$\text{solve} \quad U_1 \begin{pmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix} = \mathbf{0} \quad \Rightarrow \text{restricted MLEs } \hat{\boldsymbol{\beta}}_1, \mathbf{0}$$

Information Matrix - Partitioned According to Partition of $\boldsymbol{\beta}$:

$$\mathbf{I} \equiv \mathbf{I}_n(\boldsymbol{\beta}) = \begin{pmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{12}^T & \mathbf{I}_{22} \end{pmatrix},$$

where $\mathbf{I}_{11}$ is $(p-q) \times (p-q)$, $\mathbf{I}_{12}$ is $(p-q) \times q$ and $\mathbf{I}_{22}$ is $q \times q$.

$$\Rightarrow \qquad \mathbf{I}^{-1} = \begin{pmatrix} \mathbf{I}^{11} & \mathbf{I}^{12} \\ \mathbf{I}^{21} & \mathbf{I}^{22} \end{pmatrix},$$

where $\mathbf{I}^{22} = (\mathbf{I}_{22} - \mathbf{I}_{12}^T \mathbf{I}_{11}^{-1} \mathbf{I}_{12})^{-1}$ by the formula for the inverse of a partitioned matrix.

The score test statistic in this situation can be written

$$\begin{pmatrix} U_1\left((\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T\right) \\ U_2\left((\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T\right) \end{pmatrix}^T \mathbf{I}^{-1} \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} U_1\left((\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T\right) \\ U_2\left((\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T\right) \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{0} \\ U_2\left((\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T\right) \end{pmatrix}^T \mathbf{I}^{-1} \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} \\ U_2\left((\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T\right) \end{pmatrix}$$

$$= U_2\left((\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T\right)^T \mathbf{I}^{22} \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \mathbf{0} \end{pmatrix} U_2\left((\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T\right)$$

Summary of Tests with Nuisance Parameter:

$$\text{Score Test:} \qquad U_2\left((\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T\right)^T \mathbf{I}^{22}\begin{pmatrix}\hat{\boldsymbol{\beta}}_1 \\ \mathbf{0}\end{pmatrix} U_2\left((\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T\right)$$

$$\text{LRT:} \qquad 2\left[\ell\left((\tilde{\boldsymbol{\beta}}_1^T, \tilde{\boldsymbol{\beta}}_2^T)^T\right) - \ell\left((\hat{\boldsymbol{\beta}}_1^T, \mathbf{0}^T)^T\right)\right]$$

$$\text{Wald Test:} \qquad \tilde{\boldsymbol{\beta}}_2^T\left[\mathbf{I}^{22}\begin{pmatrix}\tilde{\boldsymbol{\beta}}_1 \\ \tilde{\boldsymbol{\beta}}_2\end{pmatrix}\right]^{-1}\tilde{\boldsymbol{\beta}}_2$$

- Compare each to $(1-\alpha)$ quantile of $\chi^2(q)$ for an approximate $\alpha-$level test of $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ versus $H_1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$.

Model Diagnostics

- In GLMs, as in CLMs, judging the fit of a model and determining the suitability of model assumptions is an essential part of any analysis. The tools of model diagnostics in CLMs, residuals, leverages, influence measures and the "Hat" matrix, all can be extended to GLMs.

The Hat Matrix:

For models with linear structure, the fitted values, $\hat{y}_i = \hat{\mu}_i$, can be written in matrix notation as

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

where $\mathbf{H}$ is an idempotent ($\mathbf{HH} = \mathbf{H}$), symmetric matrix with $\text{tr}(\mathbf{H}) = p$ and elements ≤ 1.

$$\text{For CLM:} \qquad \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \underbrace{\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T}_{=\mathbf{H}}\mathbf{y}$$

$$\text{and} \qquad \text{var}(\hat{y}_i) = h_{ii}\sigma^2$$

- $\mathbf{H}$ is known as the **hat matrix** because its the matrix that "puts the hat on" $\mathbf{y}$. Its diagonal elements h_{ii} $i = 1, \ldots, n$, are of special importance.

- The reciprocals of the h_{ii}'s are called the **effective replication**, and have rough interpretation as the number of observations providing information about $\hat{y}_i$.

For a GLM, at convergence, our regression parameter estimator satisfies

$$\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{z}$$

where

$$\mathbf{z} = \hat{\boldsymbol{\eta}} + \hat{\mathbf{D}}(\mathbf{y} - \hat{\boldsymbol{\mu}})$$

depends on $\hat{\boldsymbol{\beta}}$ as does $\hat{\mathbf{V}} = \text{diag}\left(\text{var}(y_1)(\partial\eta_1/\partial\mu_1)^2, \ldots, \text{var}(y_n)(\partial\eta_n/\partial\mu_n)^2\right)$ and $\hat{\mathbf{D}} = \text{diag}\left((\partial\eta_1/\partial\mu_1), \ldots, (\partial\eta_n/\partial\mu_n)\right)$.

$\hat{\boldsymbol{\beta}}$ is a WLS solution to the linear regression problem $\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, or equivalently, $\hat{\boldsymbol{\beta}}$ is an OLS solution to the linear regression problem,

$$\mathbf{z}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{e}^*$$

where $\mathbf{z}^* = \hat{\mathbf{V}}^{-1/2}\mathbf{z}$, $\mathbf{X}^* = \hat{\mathbf{V}}^{-1/2}\mathbf{X}$, and $\mathbf{e}^* = \hat{\mathbf{V}}^{-1/2}\mathbf{e}$.

The hat matrix corresponding to this model is

$$\begin{aligned}\mathbf{H} &= \mathbf{X}^*(\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T} = \hat{\mathbf{V}}^{-1/2}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}^{-T/2}\hat{\mathbf{V}}^{-1/2}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-T/2} \\ &= \hat{\mathbf{V}}^{-1/2}\mathbf{X}(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1/2}\end{aligned}$$

- Notice the GLM hat matrix $\mathbf{H}$ depends on $\hat{\boldsymbol{\beta}}$ through $\hat{\mathbf{V}}$.

Residuals in GLMs:

- In a CLM, residuals are typically defined as (possibly standardized) estimated error terms ($\hat{e}_i$'s).
- Obviously, this sort of a definition must be modified in the GLM context where we have no error term in the model.
- In R, the $\hat{\mu}_i$s can be obtained using the fitted() function and residuals of different types obtained with the residuals() and rstandard() functions.
- In SAS PROC GENMOD, residuals of different types can be obtained with the OBSTATS option on the MODEL statement.

1. Pearson Residual

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\text{var}}(y_i)/\hat{\phi}}} = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)/w_i}}$$

- r_i^P is defined so that $\hat{\phi}$ does not appear in the denominator. This omission is typically of little consequence because the pattern of residuals rather than their magnitude is important. However, $\sqrt{\hat{\text{var}}(y_i)}$ is sometimes used in the denominator. This is called the **studentized version**:

$$r_i^{P'} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} v(\hat{\mu}_i)/w_i}}$$

Normal dist'n: $r_i^P = y_i - \hat{\mu}_i$

Poisson dist'n: $r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$

2. Standardized Pearson Residuals:

   In the Pearson residuals, the denominator is just an estimate of $\text{var}(y_i)$, not an estimate of the variance of the entire numerator, $y_i - \hat{\mu}_i$. To take into account $\text{var}(\hat{\mu}_i)$ and $\text{cov}(y_i, \hat{\mu}_i)$ we can consider

$$r_i^{PS} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\text{var}}(y_i - \hat{\mu}_i)/\hat{\phi}}} = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)(1 - h_{ii})/w_i}}$$

   or its studentized version:

$$r_i^{PS'} = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} v(\hat{\mu}_i)(1 - h_{ii})/w_i}}$$

3. Deviance Residuals:

   Notice that generalized Pearson X^2 statistic is

$$X^2(\mathbf{y};\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n}(r_i^{P'})^2$$

   so that $r_i^{P'}$, is the signed square root of the i^{th} observation's contribution to this goodness of fit statistic. Similarly, we may define a deviance residual:

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

   where d_i is the contribution of the i^{th} observation to the (unscaled) deviance. That is $D(\mathbf{y};\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} d_i$, where

$$d_i = 2[\ell_i(y_i;y_i) - \ell_i(\hat{\mu}_i;y_i)].$$

- A **Standardized Deviance Residual** may be defined as

$$r_i^{DS} = \frac{\text{sgn}(y_i - \hat{\mu}_i)\sqrt{d_i}}{\sqrt{1-h_{ii}}}$$

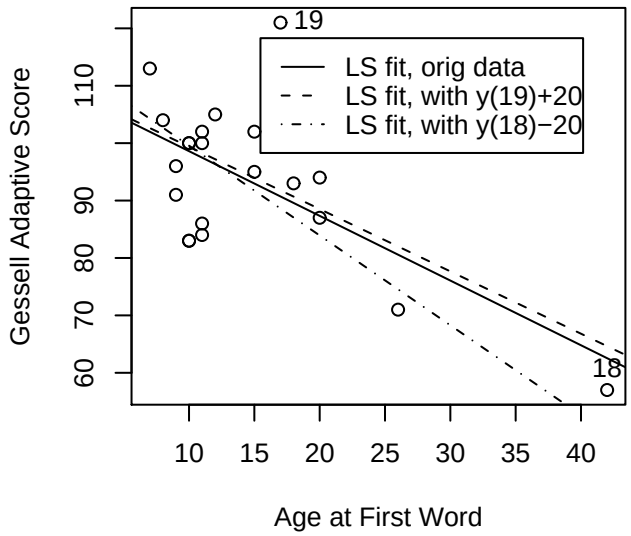
- r_i^D and r_i^{DS} each may be Studentized in an obvious way:

$$r_i^{D'} = \frac{r_i^D}{\sqrt{\hat{\phi}}}, \quad \text{and} \quad r_i^{DS'} = \frac{r_i^{DS}}{\sqrt{\hat{\phi}}}$$

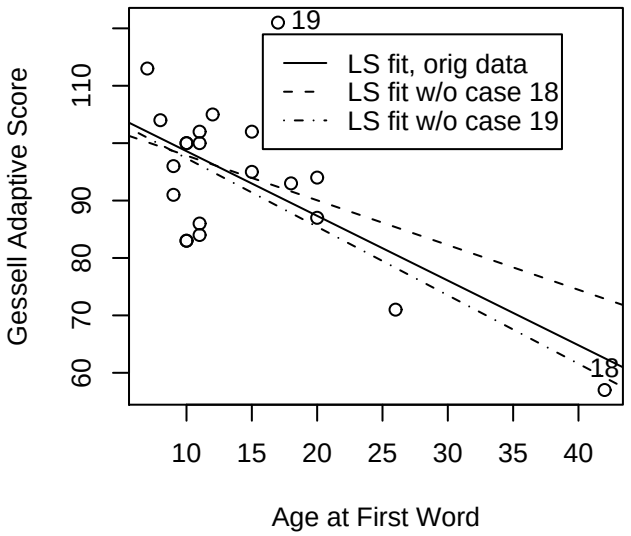
4. Others: Score Residuals, Likelihood Residuals, Anscombe Residuals (see M&N, sec. 2.4).

Influence and Leverage: Points with extreme **X** values have greater potential to be important in determining the fit in a regression problem.

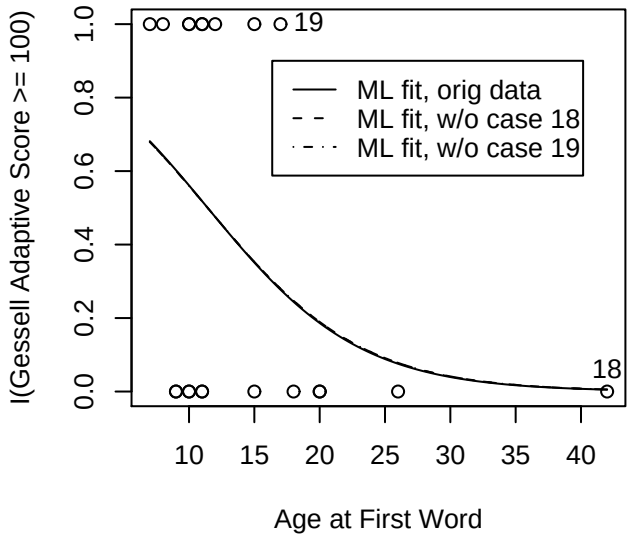
Case 18 has high leverage, 19 does not



Case 18 has high influence, 19 does not



Neither case 18 nor 19 has high influence



This phenomenon is known as leverage. A measure of the leverage of the i^{th} point is given by h_{ii}, the i^{th} diagonal element of the GLM hat matrix.

- Rough rule of thumb: Points with leverages $> 2p/n$ merit investigation.

A point with high leverage doesn't necessarily have a large influence on the fit. A measure of **influence** is **Cook's Distance:**

$$c_i = \frac{h_{ii}}{p(1-h_{ii})}(r_i^{PS'})^2 \qquad (*)$$

- c_i measures the change in the vector of fitted values that would result if we omitted observation i when fitting the model.
- Observations with high values of c_i are worthy of further investigation.
- Alternative forms of Cook's distance are possible with differently defined residuals substituted for $r_i^{PS'}$.
- PROC GENMOD does not automatically compute Cook's distances and does not output leverages (the h_{ii}'s) or Cook's distances. One can obtain the h_{ii}'s from the residuals, however. E.g.,

$$h_{ii} = 1 - \left(\frac{r_i^D}{\sqrt{\hat{\phi}} r_i^{DS'}} \right)^2$$

  and then compute Cook's distances from (*).

- PROC LOGISTIC does compute leverages and Cook's distances for binary response models (use the INFLUENCE option on the MODEL statement).
- In R, leverages and Cook's distances can be obtained with the influence.measures() function.

CLMs, Transformations, and Normal Error GLMs

In the CLM, to perform inference we assume that the error distribution is normal.

- Normality assumption can be checked with a normal quantile-quantile plot (Q-Q plot) of the residuals, r_i (could be raw residuals, or standardized in some way, it doesn't matter).
  - In a normal Q-Q plot the ordered residuals $r_{(i)}$ are plotted against the normal quantiles $z_i(a) = \Phi^{-1}\{(i-a)/(n+1-2a)\}$ where a is some suitably chosen constant. Usually a is 0 or $\frac{1}{2}$.
- A formal test of normality based on the Q-Q plot is given by Filliben (1975, *Technometrics*). Test statistic is the Pearson correlation coefficient for the normal Q-Q plot which uses $a = .3175$ (for reasons explained in his paper). Percentage points for this test statistic are given in the Filliben paper.
- Other normality tests exist (e.g., Shapiro-Wilk test, Kolmogorov-Smirnov test) but such tests are of limited use because they tend to have low power in small samples where normality of most concern and too much power in large samples where it is not.

What if normality assumption doesn't hold?

One possibility: Try a transformation. E.g., *log* or reciprocal transformation may be approximately normal when original data are skewed.

- Transformations are also useful for other reasons: to induce constant variance, or remove interactions (make covariate effects additive on the transformed scale).

Box-Cox Transformation Family

Suppose we have a *positive-valued* random variable Y which is non-normal.

The "simple" family of power transformations is given by

$$g_S(Y;\lambda) = \begin{cases} Y^\lambda, & \text{if } \lambda \neq 0 \\ \log Y, & \text{if } \lambda = 0 \end{cases}$$

where λ is the transformation parameter.

To avoid the discontinuity at $\lambda = 0$, Box and Cox altered the simple family slightly. The **Box-Cox family of power transformations** is given by

$$g(Y;\lambda) = \begin{cases} \frac{Y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log Y, & \text{if } \lambda = 0 \end{cases}$$

- In the CLM, these two families are equivalent because g is just a linear transformation of g_S.
- It is crucial to include in the model an intercept, or constant term, when using g. Otherwise the model is not scale invariant. (This is not an issue when using g_S.)

The choice of transformation can be made by estimating λ from the data. This can be done in either the simple family or the Box-Cox family, but we will use Maximum Likelihood to estimate λ, and for ML, the Box-Cox family is mathematically more convenient.

We assume there exists a λ so that

$$Z = g(Y;\lambda) \sim N(\mathbf{x}^T\boldsymbol{\beta}^*, \sigma^{*2})$$

A method for obtaining the MLE in a multiparameter problem like this one is to maximize the **profile likelihood** or **concentrated likelihood** in the parameter of interest (λ in this case).

Profile Likelihood (generic):

Suppose we have a parameter vector $\boldsymbol{\theta} \in \Theta$ that can be partitioned $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$.

Under the assumption that L is uniquely maximized with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ for every $\mathbf{y}$, the solution $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\alpha}}^T)^T$ to the maximization problem

$$\arg\max_{\boldsymbol{\beta},\boldsymbol{\alpha}} \log L(\boldsymbol{\beta}, \boldsymbol{\alpha}; \mathbf{y}) = \arg\max_{\boldsymbol{\beta},\boldsymbol{\alpha}} \ell(\boldsymbol{\beta}, \boldsymbol{\alpha}; \mathbf{y})$$

can be obtained via the following 2-step procedure:

1. Maximize the log-likelihood with respect to $\boldsymbol{\alpha}$ while treating $\boldsymbol{\beta}$ as a fixed constant to yield $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}$. Notice that $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}$ is a function of $\boldsymbol{\beta}$.

2. Define the profile (log-)likelihood to be the log likelihood evaluated at $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}$ obtained in step 1:

$$p\ell(\boldsymbol{\beta}; \mathbf{y}) \equiv \ell(\boldsymbol{\beta}, \hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}; \mathbf{y}).$$

   Find $\hat{\boldsymbol{\beta}}$ as the solution of the maximization problem

$$\arg\max_{\boldsymbol{\beta}} p\ell(\boldsymbol{\beta}; \mathbf{y})$$

The MLE of $\boldsymbol{\theta}$ is given by $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\beta}}})$; that is, step 2 yields the MLE of $\boldsymbol{\beta}$ and the MLE of $\boldsymbol{\alpha}$ is obtained by plugging $\hat{\boldsymbol{\beta}}$ into $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}$ obtained in step 1.

The profile likelihood can be used as an ordinary likelihood in several respects:

i. The maximum of $p\ell(\boldsymbol{\beta}; \mathbf{y})$ equals the overall MLE of $\boldsymbol{\beta}$.

ii. The profile log-likelihood ratio statistic for a hypothesis of the form $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$:

$$2[p\ell(\tilde{\boldsymbol{\beta}}; \mathbf{y}) - p\ell(\boldsymbol{\beta}_0; \mathbf{y})]$$

equals the log-likelihood ratio statistic; i.e.,

$$2[p\ell(\tilde{\boldsymbol{\beta}}; \mathbf{y}) - p\ell(\boldsymbol{\beta}_0; \mathbf{y})] = 2[\ell(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\alpha}}_{\tilde{\boldsymbol{\beta}}}; \mathbf{y}) - \ell(\boldsymbol{\beta}_0, \hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}_0}; \mathbf{y})]$$

$$\Rightarrow \qquad 2[p\ell(\tilde{\boldsymbol{\beta}}; \mathbf{y}) - p\ell(\boldsymbol{\beta}_0; \mathbf{y})] \overset{a}{\sim} \chi^2(t_1 - t_2)$$

where $t_1 - t_2$ = reduction in the number of independent parameters implied by H_0.

iii. A profile likelihood region can be formed by inverting a profile likelihood ratio test.

iv. The observed profile information matrix equals the portion of the full observed information matrix evaluated at $(\boldsymbol{\beta}^T, \hat{\boldsymbol{\alpha}}_{\boldsymbol{\beta}}^T)^T$ corresponding to $\boldsymbol{\beta}$.

Back to Box-Cox:

The probability density for $Z = g(Y;\lambda)$ is $N(\mu^*, \sigma^{*2})$ where $\mu^* = \mathbf{x}^T\boldsymbol{\beta}^*$, or

$$f_Z(z;\lambda,\mu^*,\sigma^*) = \frac{1}{\sigma^*\sqrt{2\pi}} \exp\left\{-\frac{(z-\mu^*)^2}{2\sigma^{*2}}\right\}$$

so by a change of variable,

$$f_Y(y;\lambda,\mu^*,\sigma^*) = \frac{1}{\sigma^*\sqrt{2\pi}} y^{\lambda-1} \exp\left\{-\frac{[(y^\lambda-1)/\lambda-\mu^*]^2}{2\sigma^{*2}}\right\}, \quad \lambda \neq 0$$

$$f_Y(y;0,\mu^*,\sigma^*) = \frac{1}{\sigma^*\sqrt{2\pi}} y^{-1} \exp\left\{-\frac{[\log y-\mu^*]^2}{2\sigma^{*2}}\right\}, \quad \lambda = 0$$

To simplify the form of the density when $\lambda \neq 0$, its useful to reparameterize. When $\lambda \neq 0$ define $\beta_1 = 1 + \lambda\beta_1^*$, $\beta_j = \lambda\beta_j^*$, $j = 2,\ldots,p$, and $\sigma = |\lambda|\sigma^*$. When $\lambda = 0$ define $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, $\sigma = \sigma^*$.

With this reparameterization, we can write the density in terms of $\boldsymbol{\beta}, \sigma$:

$$f_Y(y;\lambda,\boldsymbol{\beta},\sigma) = \frac{1}{\sigma\sqrt{2\pi}} |\lambda| y^{\lambda-1} \exp\left\{-\frac{[y^\lambda-\mathbf{x}^T\boldsymbol{\beta}]^2}{2\sigma^2}\right\}, \quad \lambda \neq 0$$

$$f_Y(y;0,\boldsymbol{\beta},\sigma) = \frac{1}{\sigma\sqrt{2\pi}} y^{-1} \exp\left\{-\frac{[\log y-\mathbf{x}^T\boldsymbol{\beta}]^2}{2\sigma^2}\right\}, \quad \lambda = 0$$

Now suppose we have a sample $(y_i, \mathbf{x}_i)$, $i = 1,\ldots,n$. For the entire data set the log-likelihood is

$$\ell(\lambda,\beta,\sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma + n\log|\lambda| + (\lambda-1)\sum_i \log y_i - \sum_i \frac{(y_i^\lambda - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2},$$
$$\text{for} \quad \lambda \neq 0 \qquad (*)$$

and

$$\ell(0,\beta,\sigma) = -\frac{n}{2}\log(2\pi) - n\log\sigma - \sum_i \log y_i - \sum_i \frac{(\log y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}, \quad \lambda = 0$$
$$(**)$$

Now form the profile log-likelihood:

For fixed λ,

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_i \mathbf{x}_i (y_i^\lambda - \mathbf{x}_i^T \boldsymbol{\beta})/\sigma^2$$
$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \sum_i \frac{(y_i^\lambda - \mathbf{x}_i^T \boldsymbol{\beta})^2}{\sigma^3}, \quad \text{for} \quad \lambda \neq 0$$

and, for $\lambda = 0$ the corresponding expressions are the same with y_i^λ replaced by $\log y_i$.

Denote the solutions of $\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{0}$ and $\frac{\partial \ell}{\partial \sigma} = 0$ as $\hat{\boldsymbol{\beta}}_\lambda$, and $\hat{\sigma}_\lambda$, which are given by

$$\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}^\lambda$$
$$\hat{\sigma}_\lambda^2 = \sum_i [y_i^\lambda - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\lambda]^2 / n = \frac{SS_E(\lambda)}{n}, \quad \text{for} \quad \lambda \neq 0$$

and again, for $\lambda = 0$ we replace y_i^λ with $\log y_i$.

Substituting these solutions into (*) and (**) we get

$$p\ell(\lambda) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{SS_E(\lambda)}{n} + n\log|\lambda| + (\lambda - 1)\sum_i \log y_i - \frac{SS_E(\lambda)}{2SS_E(\lambda)/n}, \quad \lambda \neq 0$$

and

$$p\ell(0) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{SS_E(0)}{n} - \sum_i \log y_i - \frac{SS_E(0)}{2SS_E(0)/n}, \quad \lambda = 0$$

or, simplifying,

$$p\ell(\lambda) = -\frac{n}{2}\log SS_E(\lambda) + n\log|\lambda| + (\lambda - 1)\sum_i \log y_i - \frac{n}{2}\left[1 + \log\left(\frac{2\pi}{n}\right)\right], \quad \lambda \neq 0$$

and

$$p\ell(0) = -\frac{n}{2}\log SS_E(0) - \sum_i \log y_i - \frac{n}{2}\left[1 + \log\left(\frac{2\pi}{n}\right)\right], \quad \lambda = 0$$

Likelihood-Based CIs

Consider a statistical model with a single parameter θ. The LRT statistic for $H_0 : \theta = \theta_0$ is

$$2 \log \lambda = 2[\ell(\tilde{\theta}) - \ell(\theta_0)]$$

and we reject H_0 at level α if

$$2 \log \lambda > \chi^2_{1-\alpha}(1)$$

Equivalently, we reject if

$$\ell(\theta_0) < \ell(\tilde{\theta}) - \frac{1}{2} \underbrace{\chi^2_{1-\alpha}(1)}_{=3.841, \text{when } \alpha = .05}$$

$\Rightarrow H_0$ is not rejected if $\ell(\theta_0)$ is within $\frac{1}{2}\chi^2_{1-\alpha}(1)$ units of $\ell(\tilde{\theta})$, the maximum of the log-likelihood function under the alternative hypothesis.

$\Rightarrow$ the values of θ satisfying this requirement, i.e., those θ for which

$$\ell(\theta) > \ell(\tilde{\theta}) - \frac{1}{2}\chi^2_{1-\alpha}(1)$$

form an approximate $100(1-\alpha)\%$ likelihood-based CI for θ.

- Endpoints of such an interval are typically not easily obtained analytically, but can be found numerically fairly easily by computing $\ell(\theta)$ over a grid of values for θ. We'll illustrate this procedure in the following example.

- When $\dim(\theta) > 1$ we can obtain such an interval for θ_j based on the profile log-likelihood, $p\ell(\theta_j)$.

Back to Box-Cox:

Using the technique we've just described above, we can obtain a $100(1-\alpha)\%$ CI for λ, the transformation parameter, as those values of λ so that $p\ell(\lambda)$ is within $\frac{1}{2}\chi^2_{1-\alpha}(1)$ units of its maximum.

The MASS library for R provides a boxcox() function that computes $p\ell(\lambda)$ over a user-supplied grid (min. value, max. value, step-size) for λ and plots these values against λ.

Example – Volume of Cherry Trees:

For 31 black cherry trees the following measurements were obtained:

$$V = \text{Volume of usable wood (cubic feet)}$$
$$H = \text{Height of tree (feet)}$$
$$D = \text{Diameter at breast height (inches)}$$

Goal: Predict usable wood volume from diameter and height.

See attached handout, cherry.R. Also see cherry.sas and cherry.lst on the course web site, for the same analysis using SAS.

Step 1: Fit $V = \beta_0 + \beta_1 H + \beta_2 D + e$.

- Based on initial plots of V against both explanatory variables, D and H, a multiple regression with an untransformed response V seems reasonable.
- This model is fit as lm1 and produces a fairly high R^2 value: $R^2 = 0.948$.
- Model diagnostics indicate normality is reasonable, but residual plot versus diameter shows a 'U'-shape. This lack of fit could be addressed in a variety of ways, but we will consider a transformation of the response.
  - This is an example where the effects of covariates are not additive on the original scale, and we seek a new scale on which they are additive. This is one of several common uses of transformations.

Step 2: Estimate transformation parameter and fit model of the form $g(V;\lambda) = \beta_0 + \beta_1 H + \beta_2 D + e$.

- The boxcox() function computes $p\ell(\lambda)$ over a grid of $\lambda-$values. I initially specified a grid of 100 points from from -2 to +2, but then refined the grid to find $\hat{\lambda} = 0.31$.

- It is worth noting that the boxcox() function ignores the constant $-(n/2)(1 + \log\{2\pi/n\})$ in the formula for $p\ell(\lambda)$.

- An approximate 95% CI for λ is given by the set of all λ such that $p\ell(\lambda) > -76.08 - 3.841/2 = -78.00$ which leads to $(.12, .49)$. Notice this interval excludes $\sqrt{}$, log, etc. transformations.

- SAS PROC TRANSREG will also estimate a Box Cox transformation. See cherry.sas for an illustration of how to use this procedure.

- Theory should be considered before settling on the transformation $(V^{.3}-1)/.3$, or, equivalently, $V^{.3}$. We're trying to predict a volumetric (cubic) measurement from two linear ones (in feet and inches). Therefore, it makes sense to consider the "side of the equivalent cube" as a more appropriate response than the "volume of the cube". I.e., use $V^{(1/3)}$, the cube-root of volume.

- Fitting $V^{(1/3)} = \beta_0 + \beta_1 H + \beta_2 D + e$ as model lm2 we obtain a substantially higher value of $R^2 = .9777$, and the model diagnostics look good.

Digression – Fitted Values in Transformation Models:

Notice that in a transformed-variable model

$$g(\mathbf{y};\lambda) = \boldsymbol{\eta} + \mathbf{e}$$

with normal errors, we predict the mean=median of the transformed variable.

The back-transformation to the original scale preserves the median, but not the mean. Therefore, on the original scale, the model predicts the median response, not the mean response.

Example – Lognormal Distribution:

Suppose $\log Y \sim N(\mathbf{x}^T\boldsymbol{\beta}, \sigma^2) \rightarrow Y \sim \textit{Lognormal}$.

Although $\text{median}(Y) = \exp(\mathbf{x}^T\boldsymbol{\beta}) = \exp\{\text{median}(\log Y)\}$, by the moments of the lognormal distribution

$$\text{E}(Y) = \exp\{\mathbf{x}^T\boldsymbol{\beta} + \frac{1}{2}\sigma^2\} \neq \exp\{\text{E}(\log Y)\}$$

and

$$\text{var}(Y) = [\exp(\sigma^2) - 1]\exp(2\mathbf{x}^T\boldsymbol{\beta} + \sigma^2) \neq \exp\{\text{var}(\log Y)\}$$

$\Rightarrow$ if we obtain fitted values for the original-scale response variable via

$$\exp(\hat{\mu}_i), \quad i = 1, \ldots, n$$

these are fitted values for the median response, not the mean.

For the mean, the fitted values are

$$\exp(\hat{\mu}_i + \hat{\sigma}^2/2), \quad i = 1, \ldots, n$$

In general, for power transformations Y^λ, $\lambda \neq 0$ with $\mu = \text{E}(Y^\lambda)$,

$$\begin{aligned} \text{median}(Y) &= \mu^{1/\lambda} \\ \text{E}(Y) &\approx \mu^{1/\lambda}\left(1 + \frac{\sigma^2(1-\lambda)}{2\lambda^2\mu^2}\right) \\ \text{var}(Y) &\approx \frac{\mu^{2/\lambda}\sigma^2}{\lambda^2\mu^2} \end{aligned}$$

Back to Cherry Tree Example:

- After fitting $V^{1/3} = \beta_0 + \beta_1 H + \beta_2 D + e$ we obtain fitted values for the median of V as $\widehat{V^{1/3}}^3$ and fitted values for the mean of V as

$$\hat{\mu}^{1/\hat{\lambda}} \left(1 + \frac{\hat{\sigma}^2(1-\hat{\lambda})}{2\hat{\lambda}^2\hat{\mu}^2}\right) = \widehat{V^{1/3}}^3 \left(1 + \frac{3\hat{\sigma}^2}{\widehat{V^{1/3}}^2}\right).$$

Step 3: Alternatively, consider transforming the explanatory variables and the response variable.

- Since the volume of a tree can be expected to be related to its height and diameter in a multiplicative fashion, its natural to consider an alternative model of the form $\log V = \beta_0 + \beta_1 \log H + \beta_2 \log D + e$. Next in the R script we take logs of all variables, produce scatterplots, and fit this model as lm3.

  - Note that before taking $\log(D)$ I divided by 12 to put diameter, originally in inches, on the scale of feet to match the other variables.

- For this model $R^2 = 0.9777$, for an equally good fit as compared with the cube-root model. Normality on the log scale appears to be supported.

- Once we log-transform the explanatory variables, was the log transformation appropriate for V as well? We address this model by estimating λ from

$$g(V;\lambda) = \beta_0 + \beta_1 \log H + \beta_2 \log D + e$$

  Result is $\hat{\lambda} = -0.07$ with an approximate 95% CI of $(-.24, .11)$ which easily includes 0 and excludes other possibilities such as $\lambda = 1/3$.

- Using logs of all three variables is suggested by theoretical volume equations. There are two simple shape models: a tree is shaped like 1) a cylinder, or 2) a cone.

  1. $V = \pi d^2 h/4$

  $$\Rightarrow \qquad \log V = \log \frac{\pi}{4} + \log h + 2 \log d$$

  2. $V = \pi d^2 h/12$

  $$\Rightarrow \qquad \log V = \log \frac{\pi}{12} + \log h + 2 \log d$$

  We can see which theory the data support by fitting a model (glm4) with $\beta_1 = 1$, $\beta_2 = 2$ both fixed:

  $$\log V = \beta_0 + \underbrace{1 \log H + 2 \log D}_{\text{the offset}} + e$$

- The estimated value $\hat{\beta}_0 = -1.199$ is closer to $\log(\pi/12) = -1.34$ than to $\log(\pi/4) = -.2416$, so the data supports the cone theory.

Step 4: How do GLMs, with cube-root and log link functions rather than CLMs with cube-root and log transformations of the response compare?

- Note the difference: In the GLM, we assume that the response on the original scale is normally distributed with mean that is related to covariates via $g(\mu) = \mathbf{X}\boldsymbol{\beta}$. In the CLM with transformation, we assume that $g(y; \lambda)$ is normally distributed with mean equal to $\mathbf{X}\boldsymbol{\beta}$.
  - In the former case, the response has constant variance on the original scale. In the latter case the transformed response has constant variance.
- In model glm5 we fit the GLM

$$V \sim N(\mu, \sigma^2), \qquad \mu^{1/3} = \beta_0 + \beta_1 H + \beta_2 D.$$

  We see that the model fits nearly as well as the corresponding transformation model ($R^2 = .9773$) and normality of the error distribution is supported.
- In model glm6, we fit the GLM

$$V \sim N(\mu, \sigma^2), \qquad \log \mu = \beta_0 + \beta_1 \log H + \beta_2 \log D.$$

  We again see that the model fits as well as the corresponding transformation model ($R^2 = .9778$) and normality of the error distribution is supported.

Which Model?

We now have four models which all seem to fit the data comparably well:

$$V^{1/3} \sim N(\mu, \sigma^2), \qquad \mu = \beta_0 + \beta_1 H + \beta_2 D \qquad (I)$$
$$\log V \sim N(\mu, \sigma^2), \qquad \mu = \beta_0 + \beta_1 \log H + \beta_2 \log D \qquad (II)$$
$$V \sim N(\mu, \sigma^2), \qquad \mu^{1/3} = \beta_0 + \beta_1 H + \beta_2 D \qquad (III)$$
$$V \sim N(\mu, \sigma^2), \qquad \log \mu = \beta_0 + \beta_1 \log H + \beta_2 \log D \qquad (IV)$$

- How can we compare these models? Notice *these are not nested models.* None of them can be written as a special case of one or more of the others.
- However, it is possible to express all of these models as distinct special cases of a more general model.

Models I vs. II: Both of these models are special cases of the model

$$V^* \sim N(\mu, \sigma^2), \qquad \mu = \beta_0 + \beta_1 H^* + \beta_2 D^*$$

where

$$V^* = \frac{V^{\lambda_V} - 1}{\lambda_V}, \quad H^* = \frac{H^{\lambda_H} - 1}{\lambda_H}, \quad D^* = \frac{D^{\lambda_D} - 1}{\lambda_D}$$

The comparison is between models with different values of $\boldsymbol{\lambda} = (\lambda_V, \lambda_H, \lambda_D)^T$. For model I, $\boldsymbol{\lambda} = (1/3, 1, 1)^T$, and for model II, $\boldsymbol{\lambda} = (0, 0, 0)^T$.

From the formulas at the bottom of p.123 for $p\ell(\lambda)$ ($\sum_i \log y_i = 101.4547$ for these data and the values of $SS_E(\lambda)$ are the deviance-values given by R) we have:

Model	$2p\ell(\lambda)$
I	-133.8
II	-132.2
III	-143.2
IV	-142.4

We'd like to choose between the simple null hypotheses $H_1 : \boldsymbol{\lambda} = (1/3, 1, 1)^T$ and $H_2 : \boldsymbol{\lambda} = (0, 0, 0)^T$.

From testing theory we know that the optimal test is based on the LR, or equivalently, on (twice) the difference in log-likelihoods. If this difference is "large" then we can distinguish the models.

In general, however, the distribution for such a test statistic may be difficult to obtain. We cannot appeal to Wilks theorem for the distribution of twice the difference in loglikelihoods because we do not have nested hypotheses.

- To compare models 1 and 2 we can examine

$$|2p\ell(0) - 2p\ell(1/3)| = |-132.2 + 133.8| = 1.6.$$

  This value seems small, suggesting that the models fit equally well, but we have no reference distribution.

We can go one step farther to incorporate models III and IV in our model comparisons by noting that all four models are special cases of the model

$$V^* \sim N(\mu, \sigma^2), \qquad \mu^* = \beta_0 + \beta_1 H^* + \beta_2 D^*$$

where V^*, H^*, and D^* are as before, and

$$\mu^* = \frac{\mu^{\lambda_L} - 1}{\lambda_L} \qquad \text{(L for link)}$$

Now let $\boldsymbol{\lambda} = (\lambda_V, \lambda_H, \lambda_D, \lambda_L)^T$. For the four models we have

$$\boldsymbol{\lambda} = (1/3, 1, 1, 1)^T \qquad (I)$$
$$\boldsymbol{\lambda} = (0, 0, 0, 1)^T \qquad (II)$$
$$\boldsymbol{\lambda} = (1, 1, 1, 1/3)^T \qquad (III)$$
$$\boldsymbol{\lambda} = (1, 0, 0, 0)^T \qquad (IV)$$

Therefore, an appropriate test statistic to compare model I vs model III is

$$|2p\ell(1/3) - 2p\ell(1)| = |-133.8 + 143.2| = 9.4$$

and a comparison of models II and IV can be based on

$$|2p\ell(0) - 2p\ell(1)| = |-132.2 + 142.4| = 10.2$$

These results suggest that transformation models fit the data better than the corresponding GLMs. However, this conclusion is somewhat subjective.

The problem of formally testing to choose between non-nested models is a hard one (for some ideas on this topic see F&T, sec. 4.3.3). An alternative approach not based on formal hypothesis testing procedures is to compare models in terms of **information criteria** (aka **model selection criteria**).

Intuitively, it makes sense to prefer a model under which the data are more likely than under an alternative model. That is, the maximized loglikelihood seems a reasonable criterion by which to select between competing models. This idea is consistent with test optimality results mentioned above.

However, It is always possible to increase the loglikelihood by moving toward the saturated model (adding more parameters). Therefore, we need to balance parsimony with fit. This leads to the idea of using a penalized loglikelihood criterion:

- Akaike's Information Criterion (AIC):

$$\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) - \dim(\boldsymbol{\theta})$$

  where $\hat{\boldsymbol{\theta}}$ is the MLE of the model parameter $\boldsymbol{\theta}$.

- Schwarz's Bayesian Information Criterion (BIC) (sometimes called SBC):

$$\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) - \frac{\dim(\boldsymbol{\theta})}{2} \log(n)$$

  where n is the total sample size.

- In these forms, these criteria are used as follows: models with larger values of AIC or BIC are better.

- Often (e.g., in R) you'll see both AIC and BIC defined as -2 times the definitions I gave above. In that case, smaller values of these criteria are better.

- One can use either AIC or BIC. There are arguments in favor of each, but BIC tends to be more conservative, in the sense that it tends to select simpler models. There are also some theoretical advantages to BIC: roughly, the probability that BIC chooses the right model among two nested alternatives tends to 1 as $n \to \infty$. This is not true for AIC if the smaller of the two models is the true model. However, this is a theoretical, asymptotic advantage and finite sample performace of these criteria is often similar.

- The idea of penalizing the maximized loglikelihood in these criteria is similar to that motivating adjusted R^2 in the CLM.

- Most often, information criteria are used to compare models with different numbers of estimated parameters. However, in our Cherry Tree example, the four models to be compared all have the same number of parameters and values of n, so that comparisons of information criteria for these models is equivalent to comparing maximized log-likelihood values. Thus, the use of information criteria to choose between models I–IV leads to the selection of model II, the log-transformation model.

- In truth, there is little to distinguish the four models we considered for the cherry tree data. All four models seem to fit the data well, with perhaps some preference for the Box-Cox models (models I and II) and some (slight) advantage for model (II) based on theoretical grounds.

    - The similarity in the performance of these models highlights the fact that (i) typically, no model is correct, but some models may be useful (or more useful than others), and (ii) there may be several models that are equally useful, but no model that is globally optimal.

    - This example is just one case. There are certainly many other examples where a Box-Cox transformation of y will give substantially different results (including better or worse performance) than the corresponding GLM with the same transformation (link) of the mean.

GLMs for Binary Response (Read Ch. 5 of Agresti)

Let $Y_1, \ldots, Y_n \overset{ind}{\sim}$ Bernoulli(π_i) (i.e., 0,1) random variables with

$$\begin{aligned} \pi_i &= \Pr(Y_i = 1) \qquad \text{Prob. of a "success"} \\ &= \mathrm{E}(Y_i) = \mu_i \end{aligned}$$

Suppose we want to build a GLM for μ_i (relating it to covariates or factors), $i = 1, \ldots, n$.

GLM systematic component:

$$g(\pi_i) = g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

Possible link functions:

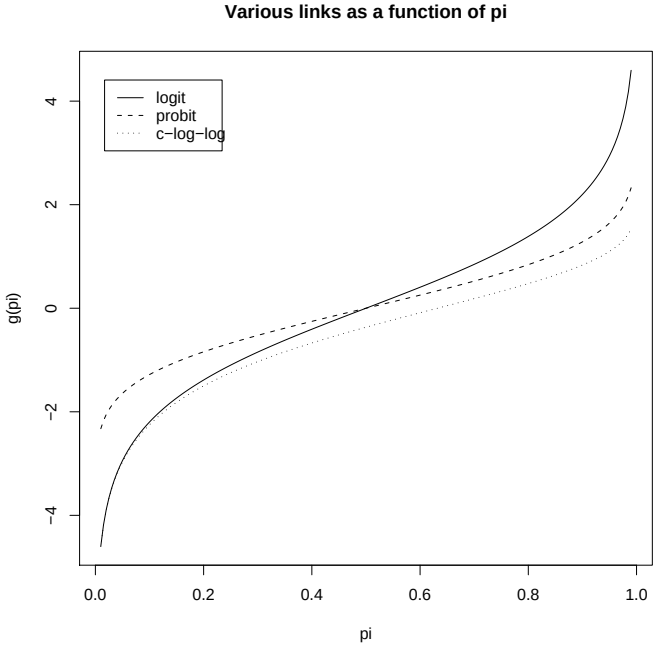
1. logit: $g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$
2. probit: $g(\pi_i) = \Phi^{-1}(\pi_i)$
3. complementary log-log: $g(\pi_i) = \log\{-\log(1-\pi_i)\}$

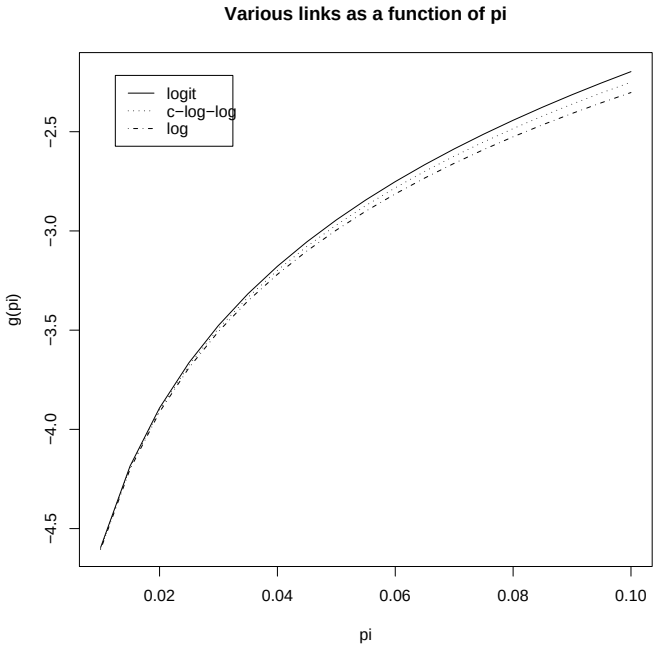
Comments:

- For $\pi \in (.1, .9)$, logit and probit functions are very similar, so these two often fit the data equally well.
- For π close to 0, logit $\approx$ log and c $-$ log $-$ log $\approx$ log, so these links give similar fits for rare events.
- logit and probit both have an **anti-symmetry property**. I.e.,

  $$g(\pi) = -g(1-\pi)$$

  $\Rightarrow$ choice of labelling doesn't matter – we get the same results no matter what we call a "success".


Various links as a function of pi
logit
probit
c−log−log
g(pi)
pi
−4
−2
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0


Various links as a function of pi
logit
c−log−log
log
g(pi)
pi
−4.5
−4.0
−3.5
−3.0
−2.5
0.02
0.04
0.06
0.08
0.10

Under anti-symmetry,

$$g(\pi_i) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$$
$$\Rightarrow \qquad g(1-\pi_i) = -\eta_i = -\mathbf{x}_i^T \boldsymbol{\beta} = \mathbf{x}_i^T(-\boldsymbol{\beta})$$

E.g., with a logit link,

If $Y_i = 1$ defined to mean subject i lived $\Rightarrow$ a unit increase in x_{ij} multiplies odds of surviving by $\exp(\beta_j)$.

If $Y_i = 1$ defined to mean subject i died $\Rightarrow$ a unit increase in x_{ij} divides odds of dying by $\exp(\beta_j)$ $\Rightarrow$ multiplies odds of living by $\exp(\beta_j)$.

c-log-log link doesn't have this property.

– As we've seen, though, c-log-log link can be well motivated by the nature of the problem (recall the dilution biassay problem a few weeks ago).

– The probit link is sometimes motivated by a *latent variable model.*

E.g., Suppose an experiment is conducted to assess the effectiveness of an insecticide by applying various doses, d_i, $i = 1, \ldots, n$, to batches of insects.

Let U_{ij} be the tolerance of the j^{th} insect in the i^{th} dose group (batch). That is U_{ij} is the level of the insecticide above which the $(i,j)^{\text{th}}$ insect will die.

Suppose $U_{ij} \sim N(\mu, \sigma^2)$. Then π_{ij} the probability of death for the $(i,j)^{\text{th}}$ insect is given by

$$\pi_{ij} = \Pr(U_{ij} < d_i) = \int_{-\infty}^{d_i} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(s-\mu)^2}{2\sigma^2}\right\} ds = \Phi\left(\frac{d_i - \mu}{\sigma}\right)$$

Now let $\beta_0 = -\mu/\sigma$, $\beta_1 = 1/\sigma$, then this model yields

$$\pi_{ij} = \Phi(\beta_0 + \beta_1 d_i) \Rightarrow \Phi^{-1}(\pi_{ij}) = \beta_0 + \beta_1 d_i,$$

the probit regression model.

Notice we never observe U_{ij}. It is a latent random variable assumed to follow a normal distribution, motivating the probit model.

Advantages of the logit link:

1. Anti-symmetry.

2. $\log\left(\frac{\pi}{1-\pi}\right)$ is the canonical parameter in the E.D. family representation of the binomial distribution. $\Rightarrow \mathbf{X}^T\mathbf{y}$ is a sufficient statistic for $\boldsymbol{\beta}$.

3. Computationally easier (than probit, especially).

4. Offers simple interpretations in terms of odds and odds ratios (a unit increase in x_j multiplies odds of success by e^{β_j}; or, if x_j is an indicator variable (e.g., $x_j = 1$ if active drug, $x_j = 0$ if placebo), then e^{β_j} is the odds ratio of success (for the active drug group compared to placebo group)).

5. With logit link, effects can be estimated in both prospective and retrospective sampling designs:

Prospective Study (Cohort Study): exposed and unexposed groups are identified and followed over time to compare incidence of disease.

Retrospective Study (Case-Control Study): diseased and disease-free subjects are identified and their exposure history investigated.

		Disease Status $\bar{D}$	D	
Exposure Status	$\bar{E}$	π_{00}	π_{01}	$\pi_{0\cdot}$
	E	π_{10}	π_{11}	$\pi_{1\cdot}$
		$\pi_{\cdot 0}$	$\pi_{\cdot 1}$	1

	Prospective	Retrospective
Row totals:	fixed	random
Col. totals:	random	fixed

The ratio of the odds of being diseased given that you were exposed to the odds of being diseased given that you were not exposed is

$$\log\left(\frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}}\right) \equiv \gamma$$

In both designs we can estimate this quantity if we use log odds models (logistic regression models) but under the two designs we get at the log odds ratio in different ways:

Prospective design:

$$\begin{aligned}\text{logodds}(D|E) - \text{logodds}(D|\bar{E}) &= \log\left(\frac{\pi_{11}}{\pi_{10}}\right) - \log\left(\frac{\pi_{01}}{\pi_{00}}\right) \\ &= \log\left(\frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}}\right) = \gamma\end{aligned}$$

Retrospective design:

$$\begin{aligned}\text{logodds}(E|D) - \text{logodds}(E|\bar{D}) &= \log\left(\frac{\pi_{11}}{\pi_{01}}\right) - \log\left(\frac{\pi_{10}}{\pi_{00}}\right) \\ &= \log\left(\frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}}\right) = \gamma\end{aligned}$$

In a prospective design, we have information appropriate to estimate

$$\pi_i = \Pr(\underbrace{Y_i = 1}_{\text{diseased}} \mid \underbrace{\mathbf{x}_i}_{\text{covariates,exp.status}})$$

Logistic regression model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \mathbf{x}_i^T\boldsymbol{\beta} \quad \Rightarrow \quad \pi_i = \frac{\exp(\alpha + \mathbf{x}_i^T\boldsymbol{\beta})}{1+\exp(\alpha + \mathbf{x}_i^T\boldsymbol{\beta})}$$

In a retrospective design, let

$$Z_i = \begin{cases} 1, & \text{if subject } i \text{ was sampled} \\ 0, & \text{otherwise} \end{cases}$$

Then the sampling proportions are

$$\begin{aligned} p_0 &= \Pr(Z_i = 1 | Y_i = 1) \quad \text{for cases} \\ p_1 &= \Pr(Z_i = 1 | Y_i = 0) \quad \text{for controls} \end{aligned}$$

By Bayes' Theorem,

$$\begin{aligned} &\Pr(Y_i = 1 | Z_i = 1, \mathbf{x}_i) \\ &= \frac{\Pr(Z_i = 1 | Y_i = 1, \mathbf{x}_i)\Pr(Y_i = 1 | \mathbf{x}_i)}{\Pr(Z_i = 1 | Y_i = 1, \mathbf{x}_i)\Pr(Y_i = 1 | \mathbf{x}_i) + \Pr(Z_i = 1 | Y_i = 0, \mathbf{x}_i)\Pr(Y_i = 0 | \mathbf{x}_i)} \\ &= \frac{\Pr(Z_i = 1 | Y_i = 1)\Pr(Y_i = 1 | \mathbf{x}_i)}{\Pr(Z_i = 1 | Y_i = 1)\Pr(Y_i = 1 | \mathbf{x}_i) + \Pr(Z_i = 1 | Y_i = 0)\Pr(Y_i = 0 | \mathbf{x}_i)} \\ &= \frac{\frac{p_0}{p_1}\exp(\alpha + \mathbf{x}_i^T\boldsymbol{\beta})}{1 + \frac{p_0}{p_1}\exp(\alpha + \mathbf{x}_i^T\boldsymbol{\beta})} \\ &= \frac{\exp(\alpha^* + \mathbf{x}_i^T\boldsymbol{\beta})}{1+\exp(\alpha^* + \mathbf{x}_i^T\boldsymbol{\beta})} \end{aligned}$$

where $\alpha^* = \alpha + \log(p_0/p_1)$.

- So, logistic regression models for prospective and retrospective designs yield the same estimates of covariate and exposure effects, but different intercepts.

6. Similarity to Discriminant/Classification Analysis:

Suppose we have a single covariate X (some blood marker for cancer, say) measured on a sample of patients and from the value of this covariate, we want to classify each patient as diseased or disease-free.

Suppose,
for nondiseased, $X \sim N(\mu_0, \sigma^2)$ (f_0), and
for diseased, $X \sim N(\mu_1, \sigma^2)$ (f_1).

Suppose we have prior probabilities p that a subject is diseased, $1-p$ that a subject is disease-free (p might be disease prevalence if sample patients selected at random from the general population).

We want $\Pr(D|X = x)$ (if $\Pr(D|X = x) > \Pr(\bar{D}|X = x)$ then classify as diseased).

By Bayes' Theorem,

$$\begin{aligned}\Pr(D|X=x) &= \frac{f_1(x)\Pr(D)}{f_1(x)\Pr(D)+f_0(x)\Pr(\bar{D})} \\ &= \frac{p\exp\left\{-\frac{(x-\mu_1)^2}{2\sigma^2}\right\}}{p\exp\left\{-\frac{(x-\mu_1)^2}{2\sigma^2}\right\}+(1-p)\exp\left\{-\frac{(x-\mu_0)^2}{2\sigma^2}\right\}} \\ &= \frac{\frac{p}{1-p}\exp\left\{\frac{\mu_0^2-\mu_1^2}{2\sigma^2}+\frac{\mu_1-\mu_0}{\sigma^2}x\right\}}{\frac{p}{1-p}\exp\left\{\frac{\mu_0^2-\mu_1^2}{2\sigma^2}+\frac{\mu_1-\mu_0}{\sigma^2}x\right\}+1}\end{aligned}$$

So, if we fit a logistic regression model to the conditional probability of disease:

$$\text{logit}\{\Pr(D|X=x)\} - \underbrace{\log\left(\frac{p}{1-p}\right)}_{\text{the ``offset''}} = \alpha+\beta x$$

this is a conditional likelihood version of the full likelihood model that assumes that X is normal in each of the two groups (classical discriminant analysis model), and we have

$$\alpha = \frac{\mu_0^2-\mu_1^2}{2\sigma^2}, \qquad \beta = \frac{\mu_1-\mu_0}{\sigma^2}.$$

- Such an approach is known as **logistic discriminant analysis**. Classical discriminant analysis is more restrictive, requiring multivariate normality of the covariates. However, when the assumptions underlying classical discriminant analysis hold, it can be substantially more powerful than logistic D.A.

ML Estimation for Binomial GLMs, Logit Link:

Data: $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$.

Error Distribution: $Y_1, \ldots, Y_n \overset{ind}{\sim} Bin(m_i, \pi_i)$ where Y_i is the sum of m_i Bernoulli(π_i) random variables ($m_i = 1$ corresponds to a binary response).

Systematic Component and Link: $g(\pi_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \boldsymbol{\eta}_i$ or $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{\eta}_i$ in the important special case of a logit link.

Loglikelihood:

$$\ell(\boldsymbol{\beta}; \mathbf{y}) = \sum_i \{y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i)\}$$

Score Equations:

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_j} &= \sum_i \frac{\partial}{\partial \pi_i}\{y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i)\}\frac{\partial \pi_i}{\partial \beta_j} \\ &= \sum_i \left\{\frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)}\right\} \frac{\partial \pi_i}{\partial \eta_i} \underbrace{\frac{\partial \eta_i}{\partial \beta_j}}_{=x_{ij}}\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \pi_i}{\partial \eta_i} &= \left(\frac{\partial \eta_i}{\partial \pi_i}\right)^{-1} = \left[\frac{\partial}{\partial \pi_i}\log\left(\frac{\pi_i}{1-\pi_i}\right)\right]^{-1} \\ &= \left(\frac{1}{\pi_i} + \frac{1}{1-\pi_i}\right)^{-1} = \pi_i(1-\pi_i)\end{aligned}$$

so

$$\frac{\partial \ell}{\partial \beta_j} = \sum_i (y_i - m_i\pi_i)x_{ij}, \qquad j = 1, \ldots, p$$

and the score equations are

$$\sum_i x_{ij} \underbrace{m_i \pi_i}_{=\mu_i} = \sum_i x_{ij} y_i, \qquad j = 1, \ldots, p$$

or, as a single vector equation,

$$\mathbf{X}^T \boldsymbol{\mu} = \underbrace{\mathbf{X}^T \mathbf{y}}_{\text{sufficient}}$$

- Notice that the score equations have a simple form:

  sufficient statistic for $\boldsymbol{\beta}$ = E{sufficient statistic for $\boldsymbol{\beta}$}

- To solve this score equation: use IRLS with starting values

$$\pi_i^{(0)} = \frac{y_i + 0.5}{m_i + 1}$$

Information Matrix:

The negative Hessian, or observed information matrix for $\boldsymbol{\beta}$ in a logistic regression has j, kth element

$$-\frac{\partial}{\partial \beta_k} \frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} m_i \frac{\partial \pi_i}{\partial \beta_k} x_{ij} = \sum_{i=1}^{n} m_i \pi_i (1 - \pi_i) x_{ik} x_{ij},$$

or, in matrix form,

$$-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \mathbf{X}^T \text{diag}\{m_i \pi_i (1 - \pi_i)\} \mathbf{X} = \mathbf{I}(\boldsymbol{\beta}).$$

- Notice that the observed and expected information matrices coincide here, since $-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}$ does not depend on $\mathbf{y}$ and therefore equals its expectation.
  - This result is general: in canonical link GLMs, the observed and expected information matrices coincide. A consequence of this is that the Fisher Scoring and Newton-Raphson algorithms become identical.
- The estimated inverse information forms an asymptotic variance-covariance matrix for $\hat{\boldsymbol{\beta}}$:

$$\hat{\text{var}}(\hat{\boldsymbol{\beta}}) = \{\mathbf{X}^T \text{diag}\{m_i \hat{\pi}_i (1 - \hat{\pi}_i)\} \mathbf{X}\}^{-1}.$$

Example – Determination of the ESR:

Collett (2003) presents data on the ESR and its dependence on certain blood plasma proteins.

- The ESR, or erythrocite sedimentation rate, is the rate at which red blood cells settle out of suspension in blood plasma. It tends to rise if the level of certain proteins in the plasma rise, which occurs in the presence of a variety of diseases. Therefore, determination of the ESR is a common blood screening test. A cut-off of ESR ≥ 20 is often taken as an indicator of a disease state.
- Of interest was to determine whether any association exists between $y = I(\text{ESR} \geq 20)$ and two proteins commonly associated with inflammatory diseases: fibrinogen (f) and $\gamma-$globulin (g). If so, measuring y could possibly be taken as a diagnostic tool for such diseases.
- See ESR.R and ESR.sas.

For now we start by considering simple logistic models without interaction of higher-order terms. In particular, we start by fitting four models:

$$\text{logit}(\pi_i) = \beta_0, \qquad (m0)$$
$$\text{logit}(\pi_i) = \beta_0 + \beta_1 f_i, \qquad (m1)$$
$$\text{logit}(\pi_i) = \beta_0 + \beta_1 g_i, \qquad (m2)$$
$$\text{logit}(\pi_i) = \beta_0 + \beta_1 f_i + \beta_2 g_i, \qquad (m3)$$

where, in each case we assume $y_i \overset{ind}{\sim} Bin(1, \pi_i)$, $i = 1, \ldots, n$ where $n = 32$.

- The anova function performs analyses of deviance for these models to reveal that f (fibrinogen) has a significant effect on y (reduction in deviance from null model is $6.04 > \chi^2_{1-.05}(1) = 3.84$) but g has a non-significant effect both when considered by itself and in addition to f.

Interpretation of coefficients: The coefficients in a logistic regression have interpretations in terms of odds ratios.

E.g., in model m3, the coefficient on f represents the ratio of the odds of $y = 1$ for a subject with a given value of f to the odds of $y = 1$ for a subject with f one unit lower.

$$\log\{\text{Odds}(y = 1|\mathbf{x} = (1, f, g)^T)\} = \beta_0 + \beta_1 f + \beta_2 g$$
$$\underline{\text{minus} \qquad \log\{\text{Odds}(y = 1|\mathbf{x} = (1, f-1, g)^T)\} = \beta_0 + \beta_1(f-1) + \beta_2 g}$$
$$\log\left\{\frac{\text{Odds}(y = 1|\mathbf{x} = (1, f, g)^T)}{\text{Odds}(y = 1|\mathbf{x} = (1, f-1, g)^T)}\right\} = \beta_1$$
$$\Rightarrow \qquad \frac{\text{Odds}(y = 1|\mathbf{x} = (1, f, g)^T)}{\text{Odds}(y = 1|\mathbf{x} = (1, f-1, g)^T)} = e^{\beta_1}$$

- In model m3, $e^{\beta_1} = e^{1.91} = 6.75$, so an increase of one unit in fibrinogen is associated with an estimated 6.75 times higher odds of $y = 1$, assuming all other predictors (in this case just g) are held constant.

- Note that this interpretation assumes all other predictors are held constant (which may, in some settings be artificial), so, like all coefficients in multiple regression models, β_1 has an interpretation in terms of *partial* association, not marginal association.

- Note that β_1 retains its interpretation regardless of the value of f. I.e., a unit increase in f has the same effect on the OR when we go from f=2 to f=3 as when we go from f=4 to f=5.
  - That is, the effect of f is assumed linear on the log-odds scale, but not on the log-probability or probability scale.
- Of course, the magnitude of β_1 depends on the scale on which f is measured. If, for example, we change units from f to $c \times f$, then $\hat{\beta}_1$ gets divided by c. E.g., a 0.1 increase in f (which is a 1 unit increase in $10f$) multiplies the odds of $y = 1$ by $\exp(\hat{\beta}_1/10) = e^{0.191} = 1.21$ (a 21% increase in odds for every 0.1 increase in f).

For many people, odds and odds ratios are not as easily understood as probabilities. The model can, of course, also be interpreted in terms of probability. However, because the model is not linear on a probability scale, a unit increase in f has an effect on $\Pr(y = 1)$ that depends upon the value of f.

- In the ESR dataset, the quartiles of f are 2.3, 2.6, and 3.2, and for g they are 32, 36, 38. Examining the effect of a 0.1 increase in f depends on where we take it:

$$\begin{aligned}
&\Pr(y = 1|\mathbf{x} = (1, 2.4, 36)^T) - \Pr(y = 1|\mathbf{x} = (1, 2.3, 36)^T) \\
&\quad = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 2.4 + \hat{\beta}_2 36) - \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 2.3 + \hat{\beta}_2 36) \\
&\quad = .069 - .058 = .011
\end{aligned}$$

whereas

$$\begin{aligned}
&\Pr(y = 1|\mathbf{x} = (1, 3.3, 36)^T) - \Pr(y = 1|\mathbf{x} = (1, 3.2, 36)^T) \\
&\quad = \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 3.3 + \hat{\beta}_2 36) - \text{logit}^{-1}(\hat{\beta}_0 + \hat{\beta}_1 3.2 + \hat{\beta}_2 36) \\
&\quad = .29 - .26 = 0.04
\end{aligned}$$

- Note that it depends not only on where we set f, but also on the values of the other covariates in the model.

We've used model m3 to illustrate these ideas, but because g is not significant in that model, we now focus on model m1.

- Model m1 has $D = 24.8$ on 30 d.f. However, the deviance is not an appropriate GOF statistic here because we have ungrouped binary data.

In the CLM, R^2 was a very useful statistic for summarizing the fit of a given model.

For logistic regression is there any analogue of R^2?

Yes, in fact several R^2-like measures have been proposed. Most are based on generalizing the idea that R^2, which, for the CLM, can be written as

$$R^2 = 1 - \frac{\text{Residual SS}}{\text{Total SS}},$$

quantifies the proportion of the total variability in the response that is explained by the fitted model.

Given this expression, a simple analogue of R^2 is

$$1 - \frac{\sum_i (y_i - \hat{\pi}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- Alternatively, the ordinary sums of squares above can be replaced by weighted least squares terms which weight by inverse estimated variances, which is more appropriate since $\text{var}(y_i)$ is not constant in a logistic regression model and the model is fit by IRLS not OLS.

Alternatively, since the objective function in fitting a logistic regression is the loglikelihood rather than a least-squares criterion, we could define R^2 as

$$1 - \frac{\log L(\hat{\boldsymbol{\beta}})}{\log \hat{L}_0}$$

where $\log L(\hat{\boldsymbol{\beta}})$ is the log-likelihood for the current model evaluated at its MLE, and $\log \hat{L}_0$ is the maximized loglikelihood for the null model that only contains a constant (or intercept) term.

- Like R^2, this quantity takes values in $[0, 1]$.

Several authors have proposed the quantity

$$\tilde{R}^2 = 1 - \left\{ \frac{\hat{L}_0}{L(\hat{\boldsymbol{\beta}})} \right\}^{2/n}$$

which reduces to the standard definition of R^2 when applied to the CLM and has a number of other good properties.

- However, Nagelkerke (1991) noted that $\tilde{R}^2$ takes values in $[0, \tilde{R}^2_{\max}]$ rather than $[0, 1]$ where

  $$\tilde{R}^2_{\max} = 1 - \left\{ \hat{L}_0 \right\}^{2/n}$$

  so he proposed instead the quantity

  $$R^2_N = \frac{\tilde{R}^2}{\tilde{R}^2_{\max}} \in [0, 1],$$

  which I will recommend for general use.

  - Both $\tilde{R}^2$ and R^2_N are given by PROC LOGISTIC if the RSQUARE option is used on the MODEL statement. Nagelkerke's R^2_N is given by the lrm() function of the Design package in R.

- For the ESR data, model m1 gives $R^2_N = 0.28$. However, R^2_N is still not a GOF statistic.

Hosmer & Lemeshow have proposed a GOF test suitable for ungroupable binary data. The idea is to artificially group the data into g groups according to a partitioning of the predicted probabilities of success. Then perform a standard X^2 goodness of fit for the resulting $g \times 2$ cross classification of these groups and the binary response.

Let y_{ij} denote the binary outcome for observation j in group i of the partition, $i = 1, \ldots, g$; $j = 1, \ldots, n_i$, and let $\hat{\pi}_{ij}$ be the corresponding estimated probability of success. Then the H-L GOF statistic is given by

$$\sum_{i=1}^{g} \frac{(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij})^2}{(\sum_j \hat{\pi}_{ij})[1 - (\sum_j \hat{\pi}_{ij})/n_i]}$$

which, according to simulation studies by H&L is approximately $\chi^2(g-2)$ under the hypothesis of model fit.

- H&L recommend using $g = 10$ groups of approximately equal size. So, for example, the data can be partitioned according to the deciles of $\hat{\pi}$. SAS uses a slight variation on this partitioning scheme.

We illustrate using the ESR data. See p.11 of ESR.lst, the output from ESR.sas. Here, we see the data and estimated fitted values, or probabilities of success, sorted by these fitted values. From these results, we can reconstruct the partitioning and resulting 10×2 table that is the basis of the H-L GOF test given on pp.9–10.

- This test statistic is 10.83 on 8 d.f., so we fail to reject the hypothesis of adequate fit ($p = .2114$).

- Note that the number of groups g and the particular partitioning scheme for a given number of groups g are somewhat arbitrary in the H-L test, so there are many ways to conduct the test and answers may vary. E.g., in my implementation in ESR.R, $g = 10$ groups are used, but with a slightly different partitioning (strictly according to the deciles). This results in a test statistic of 8.9 ($p = .3508$).

More on model interpretation:

We mentioned previously that a regression coefficient β_j on an explanatory variable x_j can be interpreted in terms of odds ratios. In addition, we can also interpret β_j in terms of the effect of a unit change in x_j on the probability of success π.

The effect on π for a given change δ in x_j as $\delta \to 0$ is the partial derivative of π with respect to x_j. Consider a model like $m1$ which depends on just one covariate — in this case x =fibinogen:

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

Then the estimated rate of change in π with respect to x is

$$\frac{\partial \pi}{\partial x} = \frac{\partial \pi}{\partial \eta} \times \frac{\partial \eta}{\partial x} = \pi(1-\pi) \times \beta_1$$

This rate of change is the tangent (or slope) of the $\pi(x)$ versus x curve. Clearly, this is steepest when $\pi(x) = 0.5$, which happends when $\eta = 0$ or, equivalently, when $x = -\beta_0/\beta_1$ in the simple logistic regression model. In that case,

$$\frac{\partial \pi}{\partial x} = \pi(1-\pi)\beta_1 = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\beta_1 = \beta_1/4.$$

- Thus, at a given value of x, $\hat{\pi}(x)[1-\hat{\pi}(x)]\hat{\beta}_1$ is the estimated rate of change in the success probability for a small change in x, and $\hat{\beta}_1/4$ is the estimated rate of change at the x-value that gives a 50% chance of success, which provides an upper bound on the rate of change in π for all x.
  - E.g., for the ESR data, the value of fibrinogen that yields $\hat{\pi} = 0.5$ is $f = 6.85/1.83 = 3.75$ at which point the rate of change in π is estimated to be $1.83/4 = 0.46$

Another technique that makes the coefficients in a logistic regression (or any regression, actually) more interpretable is to center the covariates.

- E.g., in model m1, β_0 is the log odds of $y = 1$ when fibrinogen=0. Since fibrinogen=0 is well outside the range of the observed data, this quantity is not meaningful (it involves extreme extrapolation, and fibrinogen=0 may be impossible or never observed in practice).

- However, if we replace x_i by $x_i - \bar{x}.$, then β_0 becomes the log odds of $y = 0$ when $x_i - \bar{x}. = 0$ or, equivalently, when $x_i = \bar{x}.$. Thus β_0 becomes a meaningful quantity. Such a transformation has no effect on β_1.

  - Alternatively, we could center x around other values such as the sample median of x or a known population mean of x, to give β_0 a different interpretation that might be preferable in some applications.

- Rescaling the x variable by dividing it by s_x, the sample standard deviation of the x_is gives β_1 interpretations in terms of the effect of a 1 SD change in x. E.g., $e^{\hat{\beta}_1}$ becomes the odds ratio comparing groups of subjects differing by 1 SD in their x-values.

  - Often both centering and rescaling are useful. That is, replace x_i by its z-score: $(x_i - \bar{x}.)/s_x$.

Inference:

Wald tests and confidence intervals are straight-forward to compute and are given by standard software.

- E.g., both R and SAS' PROC LOGISTIC give Wald approximate $100(1-\alpha)\%$ confidence intervals for individual regression coefficients as
$$\hat{\beta}_j \pm z_{1-\alpha/2}\text{se}(\hat{\beta}_j), \qquad j = 1, \ldots, p$$

  - E.g., a 95% CI for β_1 of $1.83 \pm 1.96(.9009) = (.0614, 3.5927)$ (use confint.default(model) in R (needs MASS package)). Exponentiating the endpoints of this interval gives a Wald interval for the odds ratio for a one-unit increase in fib: $(1.06, 36.3)$.

  - Note that PROC LOGISTIC has a UNITS statement that allows one to compute odds ratios (and corresponding confidence intervals) for changes in x other than 1 unit. E.g., to get a ratio of odds corresponding to a 0.1 unit increase in fibrinogen.

  - Likelihood ratio confidence intervals for parameters and corresponding odds ratios can be formed using the profile likelihood ratio approach we introduced when discussing Box-Cox transformations. PROC LOGISTIC will give these intervals by specifying CLPARM=pl or CLPARM=both and, for the odds ratio, CLODDS=pl or CLODDS=both.

  - R can give these likelihood intervals too by using the confint.glm() function, which is part of the MASS package.

- Wald tests of general linear hypotheses $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{b}$ can be done in R using the linear.hypothesis() function and in PROC LOGISTIC with the TEST statement or, if $\mathbf{b} = \mathbf{0}$, with the CONTRAST statement.

- Likelihood ratio tests can be done via analysis of deviance of nested models. E.g., the Wald test statistics of $H_0 : \beta_1 = 0$ in model m1 is 4.11 ($p = .0425$), whereas the LRT of this hypothesis has test statistic $2\log\lambda = 6.04$ ($p = .0139$).

Sometimes it is also of interest to make inference on $\mathrm{E}(y_i) = \pi_i$. This is most easily accomplished via Wald statistics.

- For example, to estimate $\mathrm{E}(y|\mathbf{x} = \mathbf{x}_0) = \pi(\mathbf{x}_0)$ we would use

$$\hat{\pi}(\mathbf{x}_0) = \mathrm{logit}^{-1}(\mathbf{x}_0^T \hat{\boldsymbol{\beta}})$$

- An approximate $100(1-\alpha)\%$ for $\pi(\mathbf{x}_0)$ is best obtained by forming the corresponding interval for $\eta_0 = \mathbf{x}_0^T\boldsymbol{\beta} = \mathrm{logit}(\pi(\mathbf{x}_0))$ via

$$\mathbf{x}_0^T \hat{\boldsymbol{\beta}} \pm z_{1-\alpha/2}\sqrt{\mathbf{x}_0^T \mathrm{var}(\hat{\boldsymbol{\beta}})\mathbf{x}_0} = (L, U)$$

  and then transforming the endpoints: $(\mathrm{logit}^{-1}(L), \mathrm{logit}^{-1}(U))$.

  - E.g., in R, I obtained point-wise confidence intervals for π at each point in a range of values for f=fibrinogen via

```
pred.m1 <- predict(m1,data.frame(fib=f0),se.fit=T,type="link")
L <- expit(pred.m1$fit-1.96*pred.m1$se.fit)
U <- expit(pred.m1$fit+1.96*pred.m1$se.fit)
```

  - PROC LOGISTIC uses the same approach to compute lower and upper endpoints for π (see the OUTPUT statement, LOWER and UPPER keywords).

- Note that these are point-wise intervals for $\mathrm{E}(y)$ not prediction intervals for y, nor are they simultaneous intervals for the whole curve $\pi(x)$.

Dose Response Curves:

Example – Tobacco Budworm Data

The following data were obtained from a study of the effects of an insecticide (cypermethrin) on tobacco budworm mortality:

Dose	Males Dead*	Females Dead*
1	1	0
2	4	2
4	9	6
8	13	10
16	18	12
32	20	16

* out of 20 insects of each sex.

- Of interest was the effect of dose on mortality and whether there is any difference between sexes in this effect.

- Unlike the ESR example, these data can be grouped or ungrouped. We have 240 binary responses here (dead vs alive for each insect). However, there are only two relevant covariates: dose and sex, with $6 \times 2 = 12$ *covariate classes*, so the data can be grouped as in the table above and regarded as 12 binomials (number dead out of 20) rather than 240 Bernoulli's.

  - Most software can fit models to such data in both grouped and ungrouped form. E.g., the glm() function in R and all relevant SAS PROCs (LOGISTIC, PROBIT and GENMOD) have more than one way to specify the model.

- See budworm.R. Here, we first plot the data as proportions versus dose. Then, to investigate the suitability of a logistic regression model, we plot sample logits versus dose and log(dose). The sample logits are defined with a little adjustment to avoid taking log(0) or dividing by 0. Let y_{ij} be the number dead out of $m_{ij} = m = 20$ at the jth dose level for the ith sex. Let $p_{ij} = y_{ij}/m$ be the corresponding sample proportion and define the sample logit as

$$\log\left(\frac{y_i + 0.5}{n_i - y_i + 0.5}\right).$$

- In toxicology studies such as this one, doses are often taken to be powers of 2 or 10 (a series of dilutions of the dose), so a log transformation of dose is often worth considering. Here they are powers of 2, so $\log_2$ is natural, but any base logarithm will do.
  - In fact, concentrations, in general, are often better analyzed on a log scale.
  - It appears that here, ldose $\equiv \log_2(\text{dose})$ is most appropriate.
- We begin by fitting a model in which we treat ldose as a continuous predictor and allow for different intercepts and slopes on ldose for the two sexes. Such a model can be parameterized in a variety of ways, but I used a model in which $y_{ij} \overset{ind}{\sim} Bin(m, \pi_{ij})$ where

$$\text{logit}(\pi_{ij}) = \beta_0 + \beta_1 \text{ldose}_{ij} + \beta_2 \text{male}_{ij} + \beta_3 \text{male}_{ij}\text{ldose}_{ij} \qquad (m1)$$

  - Note that two additional ways to fit this model are illustrated in models m1alt and m1alt2.
- Model m1 has deviance 4.99 on 8 d.f. In this grouped data context, the deviance is a suitable GOF test. Since 4.99 is substantially less than its d.f., the model does not exhibit significant lack of fit.
  - To illustrate that the deviance is twice the difference between the loglikelihoods of the current and saturated models, we fit model m2, which is the saturated model here.
- Based on model m1, we plot the fitted probability curves. We also plot the fitted log odss based on this model. In the summary of model m1, note that the difference in slopes between sexes ($\hat{\beta}_3$) is not significantly different from 0 ($p = .19$).

  $\Rightarrow$ insufficient evidence to conclude that dose effect differs across sexes. I.e., no linear ldose by sex interaction (on the log odds scale). Or, the lines in plot (e) are not significantly non-parallel.

- More surprisingly, the estimated sex main effect ($\hat{\beta}_2$) is also non-significant. However, note the interpretation of β_2 in this model:

$$\begin{array}{ll} & \log\{\text{Odds}(y=1|\text{male})\} = (\beta_0+\beta_2)+(\beta_1+\beta_3)\text{ldose} \\ \text{minus} & \log\{\text{Odds}(y=1|\text{female})\} = \beta_0+\beta_1\text{ldose} \\ \hline & \log(\text{OR}) = \beta_2+\beta_3\text{ldose} \end{array}$$

  - So, in this model with an interaction, the sex main effect, is the log odds ratio between sexes when ldose=0, but not at other values of ldose.

- If we wish to retain the interaction term but are more interested in the sex effect at another dose, we can shift ldose. E.g., if we are interested in the sex effect at the median value of ldose (2.5), we can replace ldose by ldose-2.5. This is done in model m3, from which we see a significant sex main effect ($p = .0032$).

- Alternatively, we may be willing to drop the non-significant linear ldose by sex interaction. This yields model m4, which fits non-significantly worse than model m1 (or equivalently, than m3).

  - Note that in model m4, the sex effect is significant, and there is no need to center ldose, since it is a parallel lines model (on the log odds scale).

  - A convenient reparameterization of model m4 is to replace the common intercept and sex effects by sex-specific intercepts. I.e., an equivalent model is m5:

$$\text{logit}(\pi_{ij}) = \alpha_i + \beta\text{ldose}_{ij} \qquad (m5)$$

Estimating the LD50 (AKA inverse regression, or calibration):

Consider now model m5. A quantity of interest in toxicology studies is the LD50, or the dose causing 50% lethality. In other contexts where death is not the event of interest, the analogous quantity is known as the ED50 (effective dose, rather than lethal dose).

- Sometimes 50% is not the level of lethality (or effectiveness) of interest. Instead one might be concerned with the LD90, for example.

Let ζ_p be the log-dose for which the probability of $y = 1$ is p. Then in a simple logistic regression on $x = \log(\text{dose})$ of the form

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

we have

$$\zeta_p = \frac{\text{logit}(p) - \beta_0}{\beta_1}$$

which we can estimate with

$$\hat{\zeta}_p = \frac{\text{logit}(p) - \hat{\beta}_0}{\hat{\beta}_1}.$$

Since this is a nonlinear function of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$, we can use the δ-method to estimate the variance of this quantity. According to the formula on the bottom of p.43,

$$\hat{\text{var}}(\hat{\zeta}_p) \approx \left[\left(\frac{\partial \zeta_p}{\partial \boldsymbol{\beta}} \right)^T \text{var}(\hat{\boldsymbol{\beta}}) \left(\frac{\partial \zeta_p}{\partial \boldsymbol{\beta}} \right) \right]_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}} \qquad (*)$$

where the necessary partial derivatives are

$$\frac{\partial \zeta_p}{\beta_0} = -\frac{1}{\beta_1}, \quad \frac{\partial \zeta_p}{\beta_1} = -\frac{\text{logit}(p) - \beta_0}{\beta_1^2} = -\frac{\zeta_p}{\beta_1}, \quad \text{or} \quad \frac{\partial \zeta_p}{\boldsymbol{\beta}} = -\beta_1^{-1} \begin{pmatrix} 1 \\ \zeta_p \end{pmatrix}.$$

Plugging into (*) and taking the square-root yields a standard error: s.e.$(\hat{\zeta}_p)$, which can be used to obtain a $100(1-\alpha)\%$ Wald interval for ζ_p:

$$\hat{\zeta}_p \pm z_{1-\alpha/2} \text{s.e.}(\hat{\zeta}_p)$$

- When using log(dose), this is an interval for the effective log-dose resulting in probability p of a success ($y = 1$). So, to get an interval on the dose scale, we need to exponentiate the endpoints, and our point estimate too.

- In budworm.R, we implement this procedure to get the ED50 and a 95% confidence interval for this quantity.
  - Note that there are two intercepts in this model, which complicates the computations, but only slightly.
  - For males, the ED50 estimate is 4.69 with a 95% interval of $(3.45, 6.37)$. Note that because we used $\log_2(x)$, the inverse transformation is 2^x not $\exp(x)$.

Fieller's Method:

An alternative method of computing a 95% CI for ζ_p is due to Fieller, whose method applies to any ratio of normal (or, in this case, approximately normal) estimators.

We illustrate using the LD50 in the simple logistic model, which is of the form $\hat{\zeta}_{0.5} = -\hat{\beta}_0/\hat{\beta}_1$, where both $\hat{\beta}_1$ and $\hat{\beta}_0$ are approximately normal.

- In what follows, we assume $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \text{var}(\hat{\boldsymbol{\beta}}))$, although in fact this is true only asymptotically (or approximately in finite samples). Because this result is only approximate, Fieller's method leads to an approximate CI for ζ_p, but it tends to be more accurate than the Wald interval.

Define

$$\psi = -\hat{\beta}_0 - \rho\hat{\beta}_1 = -(1, \rho)\hat{\boldsymbol{\beta}}, \quad \text{where} \quad \rho = -\frac{\beta_0}{\beta_1} = \zeta_{0.5}$$

Note that $\text{E}(\psi) = 0$ and $\text{var}(\psi) = (1, \rho)\text{var}(\hat{\boldsymbol{\beta}}) \begin{pmatrix} 1 \\ \rho \end{pmatrix}$, so

$$\frac{\psi}{\sqrt{\text{var}(\psi)}} \sim N(0, 1)$$

which implies

$$\begin{aligned} & \Pr\left(|\psi| \leq z_{1-\alpha/2}\sqrt{\text{var}(\psi)}\right) = 1 - \alpha \\ \Rightarrow \quad & \Pr\left(\psi^2 - z^2_{1-\alpha/2}\text{var}(\psi) \leq 0\right) = 1 - \alpha \end{aligned}$$

So approximate $100(1-\alpha)\%$ confidence interval endpoints can be found by solving the equation

$$\psi^2 - z^2_{1-\alpha/2}\hat{\text{var}}(\psi) = 0$$

which can be written as

$$(\hat{\beta}_1^2 - z^2_{1-\alpha/2}v_{11})\rho^2 + (2\hat{\beta}_0\hat{\beta}_1 - 2z^2_{1-\alpha/2}v_{01})\rho + (\hat{\beta}_0^2 - z^2_{1-\alpha/2}v_{00}) = 0,$$

a quadratic equation in ρ. Here $v_{00} = \hat{\text{var}}(\hat{\beta}_0)$, $v_{11} = \hat{\text{var}}(\hat{\beta}_1)$ and $v_{01} = \hat{\text{cov}}(\hat{\beta}_0, \hat{\beta}_1)$.

- We use Fieller's method in budworm.R to obtain a slightly different interval for the LD50 for males: $(3.41, 6.41)$.
  - Note that Fieller's method gives an interval for the LD50 for $\log_2(\text{dose})$, so we still have to take the inverse-$\log_2$ transformation of the endpoints to get an interval for the LD50 of dose on the original scale.
- PROC PROBIT in SAS will also give Fieller confidence intervals on the LD50 using the INVERSECL option on the MODEL statement. Note that SAS calls these fiducial intervals. See budworm.sas.

Cochran-Armitage Trend Test:

Consider again the linear logistic model:

$$\text{logit}(\pi_i) = \alpha + \beta x_i.$$

- For $x = \log(\text{dose})$, we might have fit such a model if we only had data from one of the two sexes in the budworm experiment.
- A hypothesis of obvious interest here is $H_0 : \beta = 0$, which addresses the question of whether there is a linear effect of x, or trend, on the log odds.
- We have learned several ways to test a linear hypothesis such as this. The Wald and likelihood ratio tests are both easily implemented. It is not as obvious how one might conduct a score test of this hypothesis, but it is certainly an option, and is asymptotically equivalent to the others.
- It turns out that the score test of this hypothesis can be shown to be equivalent to the Cochran-Armitage trend test, a test that was originally introduced outside the context of logistic regression as a test of trend in a $2 \times J$ contingency table.
- E.g., for the male budworms the data can be displayed in the following 2×6 table:

	$\log_2(\text{dose})$					
	0	1	2	3	4	5
Dead	1	4	9	13	18	20
Alive	19	16	11	7	2	0

- The Cochran-Armitage trend test examines whether there is a trend in the proportion dead across the columns. For details about the form of the test statistic, see Agresti (2002, ,§5.3.5).

- The test is implemented in PROC FREQ (see budworm.sas). But since it is asymptotically equivalent to the Wald or LR test of $H_0 : \beta = 0$ in the linear logistic model, the latter two tests are often used instead.

- The Wald and LR tests have the advantage that they can more easily be extended via more complex models. E.g., in our model m5 for the budworm data, we can test whether the common slope for males and females is equal to zero.

- In budworm.sas, PROC FREQ gives the C-A test statistic as -7.5631. Note that $-7.5631^2 = 57.2001$, the score test statistic for "overall regression" in the last call to PROC LOGISTIC in budworm.sas.

    - The nice thing about the LR or Wald test is that it can be extended to more complex situations. E.g., in model m5, we can still easily test that the slope on ldose is 0 using a Wald or LR test. E.g., the Wald test stat for this hypothesis is 65.92 in model m5.

Convergence and Existence of Finite MLEs:

Consider the saturated model m2, fit in budworm.sas. In the output for this model, there are warnings that the "validity of the model fit is questionable" and that the MLE may not exist. Furthermore, there is another warning in the log about "quasi-complete separation" of the data and non-existence of the MLE.

What's that all about?

This is not due to this being the saturated model. Rather it occurs because we are trying to use estimate probabilities for which the corresponding sample proportions are 1 and/or 0.

- E.g., in this saturated model, the linear predictor specifies a distinct log-odds for each dose by sex combination. So, the model fits probabilities for males at dose=32 where $p = 20/20 = 1$ and for females at dose=1 where $p = 0/20 = 0$.

When $p = 0$, the appropriate estimate of π is 0 too. However, obtaining an MLE $\hat{\pi} = 0$ in a model of the form

$$\text{logit}(\pi) = \eta$$

is impossible. It requires $\hat{\eta} \to -\infty$. Similarly, to get an estimate $\hat{\pi} = 1$ would require $\hat{\eta} \to \infty$.

Because we are on the logit scale, the MLEs do not exist in this situation. In fact, whenever we fit logistic regression models where covariate classes are specified for which all responses are either 0 or 1, the MLEs don't exist.

- This situation is known as *complete separation*.
- Different programs handle this problem in different ways. SAS checks for it and reports the warnings we saw in budworm.sas. Complete separation typically results in very large (in magnitude) regression coefficients and extremely large standard errors, which cannot be regarded as an appropriate basis of inference.
  - See budworm2.sas for a clearer illustration of the phenomenon. Note that the regression coefficients in this model corresponding to the two problematic covariate classes have been estimated to be very large in magnitude. If the convergence criterion for the model was set to be small enough, the model would actually continue to iterate and push these estimates further and further toward $\pm\infty$.
  - Note that a binomial linear model would have no such difficulty with non-existence of the MLEs.

Inference about conditional association in multiple 2×2 tables:

An important statistical problem with applications to epidemiology and many other fields is inference regarding two-way association between dichotomous variables in the presence of a third confounding or effect-modifying variable.

- We assume that the third variable is categorical, or a categorized continuous variable. This variable is sometimes called a stratifying variable and the set-up refered to as "stratified two-way tables".
  - In fact, the strata can be formed from the levels of a single nuisance variable, or the combinations of the levels of several nuisance variables.

Such a problem can be handled within the framework of logistic regression, but also via an (originally) non-model-based methodology due to the statisticians Cochran, Mantel and Haenszel.

A prototypical example is given by Agresti. The data below come from a study designed to compare two cream preparations, one containing an active drug and the other a placebo, for curing infection.

TABLE 6.9 Clinical Trial Relating Treatment to Response for Eight Centers

		Response				
Center	Treatment	Success	Failure	Odds Ratio	μ_{11k}	var(n_{11k})
1	Drug	11	25	1.19	10.36	3.79
	Control	10	27			
2	Drug	16	4	1.82	14.62	2.47
	Control	22	10			
3	Drug	14	5	4.80	10.50	2.41
	Control	7	12			
4	Drug	2	14	2.29	1.45	0.70
	Control	1	16			
5	Drug	6	11	∞	3.52	1.20
	Control	0	12			
6	Drug	1	10	∞	0.52	0.25
	Control	0	10			
7	Drug	1	4	2.0	0.71	0.42
	Control	1	8			
8	Drug	4	2	0.33	4.62	0.62
	Control	6	1			

Source: Beitler and Landis (1985).

- The study is an example of a *multi-center clinical trial* where patients were recruited for the experiment at several different health-care facilities.

- Of interest is whether there is an association between the cream ingredient and infection cure. However, there may be systematic differences from center to center (patients may tend to be older at some centers, the hygiene conditions may vary, etc.), so we're interested in this association conditionally on center, controlling for center differences in the analysis.
    - Here, center is essentially a blocking factor and we would like to control for block effects, while assessing treatment effects on the response.

Viewing the stratifying variable (Z) as a blocking factor and cream as a two-level treatment factor (X), then it should be no surprise that an appropriate model for such a situation is a logistic regression analogue of the randomized complete block model.

Let $y_{ik} \sim Bin(n_{ik}, \pi_{ik})$ be the binomial response under treatment i in stratum k. Then we can base inference on the model

$$\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad i = 1, 2; k = 1, \ldots, K, \qquad (*)$$

where $x_1 = 1$ (treated) and $x_2 = 0$ (control).

- Model (*) assumes that the conditional odds ratio between Y and X, which is e^β, is the same at each level of Z.

The hypothesis of conditional independence between X and Y is $H_0 : \beta = 0$, which can be tested via either a Wald test ($\hat{\beta}/\text{s.e.}(\hat{\beta}) \sim N(0, 1)$ under H_0) or via a LRT (via anaysis of deviance between the above model and the same model dropping the βx_i term).

Cochran-Mantel-Haenszel Test:

A third approach is to use the score test of H_0. It turns out that the score test of $H_0 : \beta = 0$ in this model is equivalent to a test originally proposed by Cochran (1954).

As in the logistic model (*), Cochran assumed data of the form

Data:

n_{11k}		n_{1+k}
n_{21k}		n_{2+k}

, $\quad k = 1, \ldots, K$ (K different strata).

- E.g., $n_{11k} = \#$ successes among treated, $n_{21k} = \#$ successes among untreated.
- Assume that θ_{XYk}, the odds ratio between X and Y in the kth stratum, is constant across k (a common odds ratio, call it θ_{XY}).
- And assume $n_{i1k} \overset{ind}{\sim} Bin(n_{i+k}, \pi_{ik})$, $i = 1, 2$; $k = 1, \ldots, K$.

Then the hypothesis of $H_0 : \theta_{XY} = 1$ is equivalent to $H_0 : \beta = 0$ in model (*), and the score test statistic of this hypothesis can be written as

$$CMH = \frac{[\sum_k (n_{11k} - \mu_{11k})]^2}{\sum_k \text{var}(n_{11k})}$$

where μ_{11k} and $\text{var}(n_{11k})$ are the mean and variance of n_{11k} under H_0.

In particular,

$$\mu_{11k} = n_{1+k} n_{+1k} / n_{++k}$$
$$\text{var}(n_{11k}) = n_{1+k} n_{2+k} n_{+1k} n_{+2k} / n_{++k}^3$$

Mantel and Haenszel (1959) proposed a very similar test, which conditions on the row and column totals in each 2×2 table, which makes n_{11k} hypergeometric, rather than binomial.

Under these assumptions, the Mantel-Haenszel test takes the same form as above, but where now

$$\text{var}(n_{11k}) = n_{1+k} n_{2+k} n_{+1k} n_{+2k} / \{n_{++k}^2 (n_{++k} - 1)\}.$$

Both Cochran's and M & H's test have asymptotic $\chi^2(1)$ distributions as n, the total sample size goes to ∞. This is the same limiting distribution as the Wald and LR test of $H_0 : \beta = 0$ in model (*), so all of these tests are asymptotically equivalent.

- However, all but the M-H test require $n \to \infty$ for fixed K. That is, they require increasing cell counts in each table. The M-H test, however, also has limiting $\chi^2(1)$ distribution under $K \to \infty$ as $n \to \infty$ without the assumption that the cell counts increase. Therefore, it is valid under "sparse data" configurations.

  - An important example of this is matched pairs (e.g., a matched case-control study, or a pretest-posttest situation).

  - This advantage of the M-H test makes it preferred by most statisticians over Cochran's test, though the name "Cochran-Mantel-Haenszel test" (or CMH test) is still often used when referring to the M-H test.

- The CMH test has surprisingly wide applicability. It arises in survival analysis (where it is known as the log-rank test), nonparametrics, and longitudinal data/repeated measures analysis.

- Mantel and Haenszel originally proposed their test with a continuity correction, but the correction results in a slightly conservative test so we omit it in this presentation.

- The CMH test can be generalized to K stratified $I \times J$ tables. There are three such generalizations depending on whether neither, one, or both of the row and column variables in the $I \times J$ tables are ordinal (as opposed to nominal).

The Mantel-Haenszel test can be viewed as a test directed at the alternative hypothesis that a weighted average of the stratum-specific odds ratios, say $\bar{\theta}_{XY\cdot} = \sum_k w_h \theta_{XYk} / \sum_k w_k$, differs from 1.

Estimation:

- In addition to a test statistic, Mantel and Haenszel ('59) also proposed an estimator for $\bar{\theta}_{XY.}$, the "average" odds ratio that takes the form of a weighted average of stratum-specific ORs:

$$\hat{\theta}_{\text{MH}} = \frac{\sum_k \frac{n_{12k}n_{21k}}{n_{++k}}\hat{\theta}_{XYk}}{\sum_k \frac{n_{12k}n_{21k}}{n_{++k}}} = \frac{\sum_k R_k}{\sum_k S_k},$$

where

$$R_k = \frac{n_{11k}n_{22k}}{n_{++k}}, \quad S_k = \frac{n_{12k}n_{21k}}{n_{++k}}.$$

  - Confidence intervals for $\bar{\theta}_{XY.}$ are usually obtained via exponentiating the Wald interval for $\log \bar{\theta}_{XY.}$ which uses

$$\hat{\text{var}}\{\log \hat{\theta}_{\text{MH}}\} = \frac{\sum_k (n_{11k}+n_{22k})R_k/n_{++k}}{2(\sum_k R_k)^2} + \frac{\sum_k (n_{12k}+n_{21k})S_k/n_{++k}}{2(\sum_k S_k)^2} + \frac{\sum_k \{(n_{11k}+n_{22k})S_k + (n_{12k}+n_{21k})R_k\}/n_{++k}}{2(\sum_k R_k)(\sum_k S_k)}$$

  - This **Mantel-Haenzel odds ratio estimator** does not assume homogeneity of the stratum-specific odds ratios (although it only makes sense to estimate the average OR when the stratum-specific OR are similar and, at least mostly, in a consistent direction).

  - Under the assumption of homogeneous odds ratios, $\hat{\theta}_{\text{MH}}$ can be shown to be the conditional MLE of the common OR, conditional on the stratum-specific table row and column margins.

  - Other estimators exist that assume homogeneity of the stratum-specific ORs (e.g., the unconditional MLE).

Other estimators of "average partial association" have been proposed and given the Mantel-Haenszel name:

Mantel-Haenszel Risk and Rate Difference Estimators:

- Under Binomial assumption for cell counts we talk of *risks.*
- Under Poisson assumption for cell counts we talk of *rates.*
- Let $\delta_k = \pi_{Y=1|X=1,k} - \pi_{Y=1|X=2,k}$ be the difference between the proportion of diseased subjects in the exposed and unexposed populations (risk difference). Assume $\delta = \delta_k$, $k = 1, \ldots, K$.
- Let $\gamma_k = \mu_{11k}/\mu_{1+k} - \mu_{21k}/\mu_{2+k}$ be the difference between the disease rates in the exposed and unexposed populations (rate difference). Assume $\gamma = \gamma_k$, $k = 1, \ldots, K$.
- M-H risk and rate difference estimators and their asymptotic variances coincide.

$$\begin{aligned}\hat{\delta}_{\text{MH}} = \hat{\gamma}_{\text{MH}} &= \frac{\sum_k \frac{n_{1+k}n_{2+k}}{n_{++k}}(\hat{p}_{h1} - \hat{p}_{h2})}{\sum_k \frac{n_{1+k}n_{2+k}}{n_{++k}}} \\ &= \frac{\sum_k (n_{11k}n_{2+k}/n_{++k} - n_{21k}n_{1+k}/n_{++k})}{\sum_k n_{1+k}n_{2+k}/n_{++k}}.\end{aligned}$$

Mantel-Haenszel Risk and Rate Ratio Estimators:

- Let $\phi_k = \pi_{Y=1|X=1,k}/\pi_{Y=1|X=2,k}$ be the ratio of the proportions of diseased subjects in the exposed and unexposed populations (risk ratio). Assume $\phi = \phi_k$, $k = 1, \ldots, K$.

- Let $\omega_k = (\mu_{11k}\mu_{2+k})/(\mu_{1+k}\mu_{21k})$ be the ratio of the disease rates in the exposed and unexposed populations (rate ratio). Assume $\omega = \omega_k$, $k = 1, \ldots, K$.

- M-H risk and rate ratio estimators coincide.

$$\hat{\phi}_{\mathrm{MH}} = \hat{\omega}_{\mathrm{MH}} = \frac{\sum_k n_{11k}n_{2+k}/n_{++k}}{\sum_k n_{21k}n_{1+k}/n_{++k}}.$$

- Variance estimators for M-H risk and rate ratio estimators differ.

- See Sato (1990, *Environmental Health Perspectives*) for asymptotic variance formulas for $\hat{\delta}_{\mathrm{MH}}, \hat{\gamma}_{\mathrm{MH}}, \hat{\phi}_{\mathrm{MH}}$, and $\hat{\omega}_{\mathrm{MH}}$.

Mantel-Haenszel estimators have several advantages over competitors.

- Simplicity and ease of computation.

- Optimality properties for M-H odds ratio estimator (Birch '64, Breslow and Day '80).

- M-H estimators are "dually consistent" (consistent in both "large-stratum" and "sparse-data" asymptotic situations). Not true for competitors.

- M-H estimators are still meaningful when stratum-specific association is not exactly constant across strata. However, the hypothesis of constant association across strata is sometimes of interest and tests of such hypotheses exist.

  - The most well-known is the *Breslow-Day test* of common odds ratio, but the hypothesis can also be tested by a GOF test of model (*).

- For additional references on CMH tests and M-H estimators see Hall *et al.* (2000, *Handbook of Statistics, Vol. 18*).

Example — Anti-infection Cream Data

- See CMHExample.sas and its output. Here we use PROC FREQ to recreate the the three-way table, Table 6.9 of Agresti. The CMH option on the TABLES statement generates the CMH test ($CMH = 6.38$, $p = .0115$) and M-H estimates $\hat{\theta}_{\text{MH}} = 2.13$ and $\hat{\phi}_{\text{MH}} = 1.42$.
  - At level $\alpha = 0.05$ we reject the hypothesis of no association between the cream ingredient and infection cure, controlling for differences across centers.
  - Since $\hat{\theta}_{\text{MH}} = 2.13 > 1$ there is a positive association between the active incgredient and cure. We estimate the odds of cure to be 2.13 times higher when the active ingredient is used than when it is not, on average across the centers used in this clinical trial.
  - The fact that the 95% CI for the common odds ratio $(1.18, 3.87)$ does not include 1 is consistent with the result of the CMH test.
- In addition, we fit model (*) using PROC LOGISTIC. The Wald test of $H_0 : \beta = 0$ has test statistic 6.42 ($p = .0113$), a result very close to that of the CMH test. The LRT statistic is 283.69-277.02=6.67 ($p = .0098$) which also agrees closely.
- The Breslow-Day test of constant odds ratios across the 8 centers has test statistic 8.00 which is asymptotically $\chi^2(7)$ under H_0 ($p = .3330$). Alternatively, the deviance or Pearson GOF tests of model (*) can be used, which yield a similar results: $D = 9.75, p = .2034$, $X^2 = 8.03, p = .3303$.
  - So there is not sufficient evidence to conclude that the odds ratio differs across strata. This lends support to using the CMH test, although the CMH test is still appropriate even under heterogeneous odds ratios, as long as the degree of heterogeneity is not extreme (e.g., inconsistent direction of association).
- See CMHExample.xlsx for the "hand calculations" of the M-H estimators.

Overdispersion

Suppose $Y \sim Bin(m, \pi)$. Then we know that $\mathrm{E}(Y) = m\pi$ and $\mathrm{var}(Y) = m\pi(1-\pi)$.

Suppose we have a sample $y_i, \ldots, y_n$ (iid copies of Y). We can estimate $\mathrm{var}(Y)$ by the sample variance

$$s^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$$

or we can use the estimated binomial variance:

$$m\left(\frac{y_\cdot}{mn}\right)\left(1 - \frac{y_\cdot}{mn}\right).$$

- If our assumption that $y_1, \ldots, y_n \overset{iid}{\sim} Bin(m, \pi)$ is correct, these two estimates should agree (be close to one another).

However, this is often not the case. Frequently, we find that

$$s^2 > m\left(\frac{y_\cdot}{mn}\right)\left(1 - \frac{y_\cdot}{mn}\right).$$

- This phenomenon is known as **overdispersion** or extra-binomial variation.
  - can only occur for $m > 1$
  - occurs frequently for bounded count (seemingly binomial) and unbounded count (seemingly Poisson) data. (In the Poisson case, $s^2 > \bar{y}$.)
  - under-dispersion is also possible, but is much more rare.

Genesis:

Over- (and under-) dispersion cannot happen if the model is correct. Its presence is an indication of a failure in the model assumptions:

The most common mechanism is clustering of the response. That is, Y is not really the sum of m independent Bernoulli's each with the same success probability, but instead, Y is actually the sum of ℓ binomials with possibly different success probabilities:

$$Y = Z_1 + \cdots + Z_\ell \qquad (\ell = \text{number of clusters})$$

where

$$Z_i \sim Bin(m_i, \pi_i), \qquad \text{and } m_1 + \cdots + m_\ell = m$$

- Differences among the π_i's $\Rightarrow$ overdispersion.

To see how this works let's simplify slightly and suppose that the cluster size is the same in all clusters; i.e., $m_i = k$ for all i ($\Rightarrow m = k\ell$).

Suppose π_i's are indep. random variables with $\mathrm{E}(\pi_i) = \pi$ and $\mathrm{var}(\pi_i) = \tau^2\pi(1-\pi)$. It follows that

$$\begin{aligned}\mathrm{E}(Y) &= \mathrm{E}\{\mathrm{E}(Y|\boldsymbol{\pi})\}, \quad \text{where } \boldsymbol{\pi} = (\pi_1, \ldots, \pi_\ell)^T \\ &= \mathrm{E}(\sum_{i=1}^{\ell} k\pi_i) = k\ell\pi = m\pi\end{aligned}$$

and

$$\begin{aligned}\mathrm{var}(Y) &= \mathrm{E}\{\mathrm{var}(Y|\boldsymbol{\pi})\} + \mathrm{var}\{\mathrm{E}(Y|\boldsymbol{\pi})\} \\ &= \mathrm{E}\{\sum_{i=1}^{\ell} k\pi_i(1-\pi_i)\} + \mathrm{var}(k\sum_{i=1}^{\ell}\pi_i) \\ &= \underbrace{m\pi(1-\pi)}_{\text{binom.var.}} \quad \underbrace{[1+(k-1)\tau^2]}_{\text{overdispersion } (>1)}\end{aligned}$$

If we let ϕ represent the overdispersion factor, we have shown that clustering of the responses yields

$$\mathrm{var}(Y) = \phi m\pi(1-\pi).$$

This argument suggests one of the models that is commonly used to handle overdispersion: the **beta-binomial model.**

Suppose $Y|\pi \sim Bin(m, \pi)$ where π is drawn from a beta distribution with density

$$f_\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

Let

$$B(\alpha,\beta) = \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^{-1},$$

then

$$f_\pi(p) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}$$

The unconditional distribution of Y has density f_Y, say, given by

$$\begin{aligned} f_Y(y) = \Pr(Y=y) &= \int_0^1 \Pr(Y=y|\pi=p) f_\pi(p) dp \\ &= \int_0^1 \binom{m}{y} p^y (1-p)^{m-y} f_\pi(p) dp \\ &= \binom{m}{y} \frac{B(\alpha+y, m+\beta-y)}{B(\alpha,\beta)} \end{aligned}$$

which is known as the beta-binomial distribution.

- Under the beta-binomial,

$$\mathrm{E}(y) = m\pi, \qquad \text{and} \qquad \mathrm{var}(Y) = \underbrace{\{1+(m-1)\tau^2\}}_{=\phi} m\pi(1-\pi)$$

  where $\pi = \alpha/(\alpha+\beta)$ and τ^2 is a function of α and β.

Another common cause of overdispersion is the presence of **correlated binary responses.**

Suppose $Y_i = X_{i1} + X_{i2}$ where $X_{i1}, X_{i2} \sim Bernoulli(\pi_i)$.

Regardless of whether or not X_{i1}, X_{i2} are uncorrelated,

$$\mathrm{E}(Y_i) = \mathrm{E}(X_{i1}) + \mathrm{E}(X_{i2}) = \pi_i + \pi_i = 2\pi_i$$

or $\mathrm{E}(Y_i) = m_i\pi_i$ where $m_i = 2$, the number of "trials".

If X_{i1}, X_{i2} are independent,

$$\mathrm{var}(Y_i) = \mathrm{var}(X_{i1}) + \mathrm{var}(X_{i2}) = \pi_i(1-\pi_i) + \pi_i(1-\pi_i) = 2\pi_i(1-\pi_i)$$

However, if X_{i1}, X_{i2} are correlated with corr. coef. ρ,

$$\begin{aligned}\mathrm{var}(Y_i) &= \mathrm{var}(X_{i1}) + \mathrm{var}(X_{i2}) + 2\mathrm{cov}(X_{i1}, X_{i2}) = 2\pi_i(1-\pi_i) + 2\rho\pi_i(1-\pi_i) \\ &= 2\pi_i(1-\pi_i)[1+\rho]\end{aligned}$$

- Thus, positively correlated binary responses $\Rightarrow$ overdispersion, negatively correlated binary responses (less common) $\Rightarrow$ underdispersion.

More generally, for Y_i the sum of m_i binary variables which are correlated with constant pairwise correlation ρ,

$$\mathrm{var}(Y_i) = m_i\pi_i(1-\pi_i)[1+(m_i-1)\rho]$$

- Correlation among the binary components of a pseudo-binomial response will be present when the response represents the sum of clustered binomials with varying success probabilities. Thus, correlation among the binary components can be either a cause of, or a symptom of overdispersion.

Overdispersion will typically be detected from lack of fit in a GLM based on a binomial model. When the deviance and Pearson X^2 statistics are appropriate measures of L.O.F. (number of parameters in the saturated model doesn't increase as $m_i \to \infty$ and none of the m_i's are very small), then a deviance (or X^2 value) greatly in excess of the residual d.f. *can indicate* overdispersion.

- The deviance can also be inflated due to outliers, an inadequately specified linear predictor (e.g., omitted or inappropriately scaled covariates) or an inappropriate link function.

- However, it is often the case that even after addressing these model inadequacies there is still substantial lack of fit (overdispersion).

How to handle it:

- Over the last 20 years there has been an explosion of research on how to handle overdispersion. Many of the techniques can be distinguished as either **empirical** or **mechanistic** models.

An empirical model is one consisting of a set of (possibly parametric) assumptions about the family of distributions from which the data are drawn that are sufficiently flexible so that a member of the family fits the data well.

A mechanistic model is one consisting of a set of assumptions about the family of distributions from which the data are drawn that are deduced from the mathematics of the (hypothesized or known) mechanism generating the data.

- The beta-binomial model is an example of a mechanistic model. Such models often allow stronger conclusions to be drawn because of stronger assumptions about the genesis of the data.

- Rather than model the source of extra-binomial variation, an empirical approach to overdispersion is to model the observable characteristics of the data.

In particular we can specify a **quasi-likelihood** model that makes only first and second assumptions rather than assuming the entire distribution of the data. A simple approach is just to assume

$$\mathrm{E}(Y_i) = m_i\pi_i, \qquad \text{and} \qquad \mathrm{var}(Y_i) = \phi m_i\pi_i(1-\pi_i)$$

where ϕ is now no longer fixed at 1, but instead is an unknown dispersion parameter to be estimated.

Quasi-likelihood

- Read the chapter by McCullagh handed out in class.

Quasi-likelihood may be thought of as an extension of least squares to nonlinear models in which the variance may depend upon the mean.

- As in LS, in QL estimation, we make only first and second moment assumptions on the response, rather than specifying the full likelihood.
- In some situations, working backward from our assumptions about first and second moments, we can infer a log-likelihood-like quantity, the quasi-likelihood function, which can be used as the basis for inference on the parameters, very much as we would do in a fully specified likelihood analysis.

Suppose $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ is a response vector consisting of independent components with mean and variance

$$\mathrm{E}(\mathbf{Y}) = \boldsymbol{\mu}(\boldsymbol{\beta}), \qquad \mathrm{var}(\mathbf{Y}) = \phi\mathbf{V}(\boldsymbol{\mu}) = \phi\mathrm{diag}\{v_1(\mu_1), \ldots, v_n(\mu_n)\}$$

Here, $v_i(\mu_i)$ is the variance function for Y_i. Typically, $v_1, \ldots, v_n$ are all the same variance function, call it $v(\mu)$.

We define the **quasi-score function** for the i^{th} observation as

$$U_i = \frac{Y_i - \mu_i}{\phi v(\mu_i)}$$

U_i has the following score-like properties:

$$\begin{aligned} \mathrm{E}(U_i) &= 0 \\ \mathrm{var}(U_i) &= \frac{1}{\phi v(\mu_i)} \\ -\mathrm{E}\left(\frac{\partial U_i}{\partial \mu_i}\right) &= \frac{1}{\phi v(\mu_i)} \end{aligned}$$

Since these properties underly much of the first-order asymptotic theory of likelihood-based inference, it is reasonable to estimate parameters based on U_i and to perform inference based on the quasi-(log)likelihood function

$$Q_i(\mu_i; y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi v(t)} dt$$

provided that this integral exists.

- By construction, $\frac{\partial Q_i}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi v(\mu_i)}$.

By the independence of $Y_1, \ldots, Y_n$,

$$Q(\boldsymbol{\mu}; \mathbf{y}) = \sum_{i=1}^{n} Q_i(\mu_i; y_i)$$

is the joint quasi-likelihood.

We may also define analogues to the deviance:

$$D_i(y_i; \mu_i) = -2\phi Q_i(\mu_i; y_i) = 2 \int_{\mu_i}^{y_i} \frac{y_i - t}{v(t)} dt$$

is the quasi-deviance for a single observations, and

$$D(\mathbf{y}; \boldsymbol{\mu}) = \sum_i D_i(y_i; \mu_i)$$

is the total quasi-deviance.

- Notice that D_i and D do not depend on ϕ.

Quasi-likelihood Estimation: By differentiating $Q(\boldsymbol{\mu}; \mathbf{y}_i)$ with respect to $\boldsymbol{\beta}$, we obtain the QL estimating equations,

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$$

where

$$\begin{aligned}\mathbf{U}(\boldsymbol{\beta}) &= \phi^{-1} \sum_i \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}^T} \right)^T (y_i - \mu_i)/v(\mu_i) \\ &= \phi^{-1} \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T} \right)^T \mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu})\end{aligned}$$

which we solve to yield the QLE of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$.

Let $\mathbf{D} = \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T}$. It may be shown that

$$\text{var}(\mathbf{U}(\boldsymbol{\beta})) = -\text{E}\left(\frac{\partial \mathbf{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \right) = \phi^{-1}\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D} \equiv \mathbf{I}_n(\boldsymbol{\beta})$$

the quasi-information matrix for $\boldsymbol{\beta}$, and

$$\text{avar}(\hat{\boldsymbol{\beta}}) = \mathbf{I}_n^{-1}(\boldsymbol{\beta}) = \phi(\mathbf{D}^T\mathbf{V}^{-1}\mathbf{D})^{-1}$$

which we can estimate by evaluating $\mathbf{D}$ and $\mathbf{V}$ at $\hat{\boldsymbol{\beta}}$ and replacing ϕ by an estimate, $\hat{\phi}$.

- Notice that in the quasi-score equation, ϕ cancels, so just as in ML, the QL estimation of $\boldsymbol{\beta}$ doesn't depend on ϕ, and ϕ can be estimated at convergence.

Solving the quasi-score equation can be done by Fisher scoring. Based on some initial value $\boldsymbol{\beta}^{(0)}$ sufficiently close to $\boldsymbol{\beta}$, the Fisher scoring update is given by

$$\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(0)} + (\mathbf{D}^{(0)T}(\mathbf{V}^{(0)})^{-1}\mathbf{D}^{(0)})^{-1}\mathbf{D}^{(0)T}(\mathbf{V}^{(0)})^{-1}(\mathbf{y} - \boldsymbol{\mu}^{(0)})$$

Notice that this exactly the same update as in ML, so QL estimation can be done by IRLS exactly as before. The only difference being that at convergence, we estimate ϕ rather than taking it to be equal to the fixed value 1.

Like the MLE, the quasi-likelihood estimator is asymptotically normal under mild regularity conditions:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{a}{\sim} N(\mathbf{0}, n\mathbf{I}_n^{-1}(\boldsymbol{\beta}))$$

Estimation of ϕ:

At convergence, ϕ can be estimated in one of several ways. The simplest approach is to use either the MOM or the deviance estimator.

MOM estimator:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{g} \frac{(y_i - \hat{\mu}_i)^2}{v_i(\hat{\mu}_i)} = \frac{X^2}{n-p}$$

Deviance estimator:

$$\hat{\phi} = \frac{D(\mathbf{y}; \hat{\boldsymbol{\mu}})}{n-p}$$

- The MOM estimator is consistent in the sparse-data situation, whereas the deviance estimator is not.
- Other estimators are available (extended-QL, pseudo-likelihood, etc.) which we may talk about if we have time.
- It is important to note that once we have allowed ϕ to be unknown and estimated it by a measure of the lack-of-fit from the fixed-ϕ model, the scaled deviance is no longer an appropriate measure of goodness of fit!
- The bottom line of the QL analysis in which we allow $\phi \neq 1$, is that regression parameter estimation is unchanged, but the standard errors of the regression parameters are multiplied by $\sqrt{\hat{\phi}}$. Failing to account for overdispersion results in underestimated s.e.s.

Example - Toxoplasmosis & Rainfall:

Residents of 34 cities in El Salvador were tested for toxoplasmosis and the percent testing positive recorded. In addition, amount of rainfall was measured for each city.

Question of interest: Is amount of rainfall related to incidence of toxoplasmosis? and if so, what's the nature of this relationship?

- See handout.
- First we plot the response, ppos=proportion testing positive, against rainfall. The plot shows no clear simple relationship, but perhaps a relatively high order polynomial model may fit the data reasonably well.
- First we fit a 5^{th}=order polynomial. According to the (scaled) deviance this model does not fit well. However, it is important to note that there are several small values of m_i, the binomial denominator, in this data set. For this reason, the chi-square approximation for D^* may not be adequate.
- Ignoring G.O.F. for now, we consider reducing the order of the polynomial. In comparison to their s.e.'s the order 4 and 5 terms appear to be nonsignificant, but the order 3 term may be significant. Therefore, we fit the cubic model.
- The reduction in deviance from 5^{th} to 3^{rd} order is $\Delta D^* = 1.438$ on Δd.f. $= 2$ ($p = .487$). The reduction from 3^{rd} to 0^{th} order is $\Delta D^* = 11.577$ on Δd.f. $= 3$ ($p = .009$). Based on these results we settle on a cubic model.
- Comment: In general, non-hierarchical models should be used rarely. That is, in a cubic model, it rarely makes sense to set linear and/or quadratic terms (or intercepts, for that matter) to 0. Similarly, in models with interactions, it is almost always appropriate to retain all main effects corresponding to covariates or factors involved in the interactions. Such non-hierarchical models should only be used when some theory exists to support such constraints.

- Notice that according to the Hosmer-Lemeshow GOF test, the cubic model doesn't fit adequately:

$$X^2 = 33.34 \qquad \text{d.f.} = 8 \qquad p < .0001.$$

- Because these data are not groupable and the model involves a continuous covariate, the deviance and Pearson X^2 statistics cannot be referred to chi-square reference distributions and are not appropriate GOF tests. (Although see the chapter by McCullagh for a discussion of the appropriate reference distribution for these statistics in this context and an argument as to why they also indicate lack of fit here.)

  So, our cubic model does not fit well and cannot be significantly improved by simply adding higher order terms. Since no additional covariates are available and we have carefully selected the linear predictor, there is little we can do to improve the model, so we decide to adopt it and adjust for the overdispersion that is present.

- To provide valid inferences, we refit the model allowing $\phi \neq 1$ and we estimate ϕ with the MOM estimator to yield $\hat{\phi} = 1.94$.

- The scaled deviance is no longer an appropriate GOF criterion.

- Notice that the regression parameter estimates are exactly the same as they were with ϕ fixed at 1, but their standard errors have been multiplied by $\sqrt{\hat{\phi}} = 1.393$.

- The change in the inference is apparent from the 95% pointwise confidence bands on the regression function.

- See the GLIM manual, sec. 12.2.1.4, for a more complete analysis of this example.

We have seen that when we observed grouped binary responses that are correlated or when we observe the sum of independent clustered binary responses with varying success probabilities we end up with data that satisfy

$$\mathrm{E}(Y_i) = m_i\pi_i \qquad \mathrm{var}(Y_i) = m_i\pi_i(1-\pi_i)[1+(m_i-1)\gamma]$$

- When $m_i = m$ for all i in our sample, then the overdispersion parameter $\phi = [1+(m_i-1)\gamma]$ is constant. However, when m_i varies, $\phi_i = [1+(m_i-1)\gamma]$ varies too.
- The quasi-likelihood model doesn't assume anything about how the overdispersion in the data were generated. However, if either the clustering mechanism or the correlation mechanism is the cause of the overdispersion then the overdispersion is not constant. In that case the QL approach estimates the "typical" overdispersion, roughly speaking.

Alternatively, we could parameterize the overdispersion in terms of γ which is constant. Williams (1982) has shown how γ can be estimated by equating X^2 to its approximate expected value.

Recall that IRLS depends only upon the first two moments of the distribution. Therefore it can be used to fit models in which a full likelihood specification implies certain mean and variance, or, more generally, any GLM-type model defined by the specification of the mean and variance directly.

That is, IRLS can handle models specified only by the following assumptions

$$\mathrm{E}(Y_i) = \mu_i = g^{-1}(\eta_i), \quad \eta_i = \mathbf{x}_i^T\boldsymbol{\beta}, \quad \mathrm{var}(Y_i) = \phi v_i(\mu_i)/w_i$$

where g is an assumed link, $\mathbf{x}_i$ is an assumed set of covariates, v_i is an assumed variance function, and w_i are known weights.

In the varying m_i case, the overdispersion model that we'd like to fit is of this form, with Y_i the number of "successes" out of m_i trials, g = logit, $v_i(\mu_i) = m_i\pi_i(1-\pi_i)$ and $w_i = 1/[1+(m_i-1)\gamma]$.

γ can be estimated by fitting this model and then using the method of moments, setting X^2 from the fitted model equal to its approximate expected value.

It can be shown that

$$X^2 = \sum_i \frac{w_i(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

has approximate expected value

$$\mathrm{E}(X^2) \approx \sum_i w_i \left(1 - \frac{u_i^2 d_i w_i}{v_i(\pi_i)}\right)[1 + (m_i - 1)\gamma]$$

where d_i is the i^{th} diagonal element of $\mathrm{var}(\hat{\boldsymbol{\eta}})$ and $u_i = \partial\mu_i/\partial\eta_i$.

Solving for γ we get

$$\hat{\gamma} = \frac{X^2 - \sum_i w_i \left(1 - \frac{u_i^2 d_i w_i}{v_i(\pi_i)}\right)}{\sum_i w_i(m_i - 1)\left(1 - \frac{u_i^2 d_i w_i}{v_i(\pi_i)}\right)}$$

- Note that the rhs above depends upon γ, so this estimation procedure must be iterated until convergence.
- Since γ measures the overdispersion not explainable by *any* binomial model, its best to estimate γ based on the maximal model.
- Once $\hat{\gamma}$ has been obtained, models can be fitted with iterative weights $w_i = 1/[1 + (m_i - 1)\hat{\gamma}]$ and these models can be compared in the usual way via deviance or X^2 comparisons.
- This **Williams procedure** can be thought of as a less restrictive version of fitting the beta-binomial model.
- PROC LOGISTIC in SAS implements the Williams procedure if you use the SCALE=WILLIAMS option in the MODEL statement (see trout.sas).

Example – Trout Eggs:

An experiment was conducted to investigate the effects of time on the survival of Brown trout eggs. Four boxes of trout eggs were placed in each of five locations in a stream. At weeks 4, 7, 8 and 11 after placing the boxes, a randomly selected box from each location was examined for egg survival. The data are as follows (table entries are s/m where s = number of eggs surviving out of m eggs in the box):

Location	Survival Period (weeks)			
in stream	4	7	8	11
1	89/94	94/98	77/86	141/155
2	106/108	91/106	87/96	104/122
3	119/123	100/130	88/119	91/125
4	104/104	80/97	67/99	111/132
5	49/93	11/113	18/88	0/138

- Note that this is not a repeated measures design. Different boxes of eggs were examined at each time point.
- See the handout trout.sas. Here we implement both the simple QL approach with constant overdispersion parameter and the Williams procedure. The latter is well motivated here because the m_i's vary greatly.

Poisson Regression

Example – Bicycle Traffic:

Gelman et al. (1995) report the results of a survey of bicycle and other traffic in the neighborhood of the UC-Berkeley campus conducted in 1993. Sixty streets were selected at random, with a stratification into 3 levels of activity and whether the street had a marked bicycle lane. The counts observed during one hour are shown below. Note that for two streets, the data were lost.

Type of street	Bike lane?		Counts									
Residential	yes	bikes	16	9	10	13	19	20	18	17	35	55
		other	58	90	48	57	103	57	86	112	273	64
Residential	no	bikes	12	1	2	4	9	7	9	8		
		other	113	18	14	44	208	67	29	154		
Side	yes	bikes	8	35	31	19	38	47	44	44	29	18
		other	29	415	425	42	180	675	620	437	47	462
Side	no	bikes	10	43	5	14	58	15	0	47	51	32
		other	557	1258	499	601	1163	700	90	1093	1459	1086
Main	yes	bikes	60	51	58	59	53	68	68	60	71	63
		other	1545	1499	1598	503	407	1494	1558	1706	476	752
Main	no	bikes	8	9	6	9	19	61	31	75	14	25
		other	1248	1246	1596	1765	1290	2498	2346	3101	1918	2318

- Of interest here is the effect of bike lanes on bike traffic and whether that effect differs by type of street (residential, side, main). Note that these data do not come from a designed experiment, but rather are observational. So, causal conclusions are not possible.

The standard distribution for modelling count data is the Poisson distribution. If events occur independently at a constant rate μ (events per unit time), then the number of events that occur over a fixed time interval T will follow a Poisson distribution with mean $T\mu$.

- Such a model seems ideally suited to the bike traffic data and is used as the distribution of first consideration for most statistical analyses of unbounded count data.
    - Note however that unbounded counts aren't automatically Poisson. E.g., event rates are often not constant during the time interval of observation, and some counts occur which are not event counts in the Poisson sense (e.g., the number of psychotherapy visits a patient attends before abandoning therapy).

Let Y_{ijk} = the number of bicycles counted on the k^{th} street of the ith type ($i = 1, 2, 3$) with jth level of bike lanes ($j = 1$ if lanes present, $j = 2$ otherwise).

If bike traffic on the $(i, j, k)^{\text{th}}$ street occurs according to a Poisson process with rate μ_{ijk}, then

$$\mathrm{E}(Y_{ijk}) = \mu_{ijk}, \qquad \mathrm{var}(Y_{ijk}) = \mu_{ijk}$$

and

$$\Pr(Y_{ijk} = y_{ijk}) = \frac{e^{-\mu_{ijk}} \mu_{ijk}^{y_{ijk}}}{y_{ijk}!}$$

ML Estimation in Poisson GLMs:

Data: $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$.

Error Distribution: $Y_1, \ldots, Y_n \overset{ind}{\sim} \text{Poisson}(\mu_i)$ where Y_i is a non-negative integer-valued random variable.

Systematic Component and Link: $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} = \boldsymbol{\eta}_i$ or $\log(\mu_i) = \boldsymbol{\eta}_i$ in the important special case of a log link.

- In Poisson and other unbounded count data regression models, the log link, which is canonical, is used almost exclusively.

Loglikelihood:

$$\begin{aligned}\ell(\boldsymbol{\beta}; \mathbf{y}) &= \sum_i [y_i \log\{\mu_i(\boldsymbol{\beta})\} - \mu_i(\boldsymbol{\beta}) - \log(y_i!)] \\ &\overset{*}{=} \sum_i [y_i \mathbf{x}_i^T \boldsymbol{\beta} - \exp(\mathbf{x}_i^T \boldsymbol{\beta}) - \log(y_i!)]\end{aligned}$$

$*-$ under a log link.

Score Equations:

$$\begin{aligned}\frac{\partial \ell}{\partial \beta_j} &= \sum_i \left\{ \frac{y_i - \mu_i(\boldsymbol{\beta})}{\mu_i(\boldsymbol{\beta})} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} \right\} \\ &\overset{*}{=} \sum_i \left\{ \frac{y_i - \mu_i(\boldsymbol{\beta})}{\mu_i(\boldsymbol{\beta})} \mu_i(\boldsymbol{\beta}) x_{ij} \right\} = \sum_i (y_i - \mu_i) x_{ij}\end{aligned}$$

so, for the log-linear model, the score equations are

$$\sum_i x_{ij} y_i = \sum_i x_{ij} \mu_i, \qquad j = 1, \ldots, p$$

or, as a single vector equation,

$$\underbrace{\mathbf{X}^T \mathbf{y}}_{\text{sufficient}} = \mathbf{X}^T \boldsymbol{\mu}$$

- Notice that again, for the canonical link, the score equations have a simple form:

$$\text{sufficient statistic for } \boldsymbol{\beta} = \text{E}\{\text{sufficient statistic for } \boldsymbol{\beta}\}$$

- To solve this score equation: use IRLS with starting values

$$\mu_i^{(0)} = y_i + 0.5$$

Information Matrix:

The negative Hessian, or observed information matrix, for $\boldsymbol{\beta}$ in a Poisson log-linear model has j, kth element

$$-\frac{\partial}{\partial \beta_k}\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_k} x_{ij} = \sum_{i=1}^{n} \mu_i x_{ik} x_{ij},$$

or, in matrix form,

$$-\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \mathbf{X}^T \text{diag}\{\mu_i\}\mathbf{X} = \mathbf{I}(\boldsymbol{\beta}).$$

- Notice that, as in all canonical link GLMs, the observed and expected information matrices coincide here.
- The estimated inverse information forms an asymptotic variance-covariance matrix for $\hat{\boldsymbol{\beta}}$:

$$\hat{\text{var}}(\hat{\boldsymbol{\beta}}) = \{\mathbf{X}^T \text{diag}\{\hat{\mu}_i\}\mathbf{X}\}^{-1}.$$

Example – Bicycle Traffic:

Recall we observe y_{ijk} = the number of bikes observed on the kth street on the ith street type, jth level of bike lanes. Though it was not a designed experiment, the stratified sampling of 10 streets at each of two levels of the factor of interest (bike lanes) within each of three street types mimics a randomized complete block design with replication, where street type is playing the role of the blocking factor, and presence/absence of bike lanes is playing the role of the treatment factor.

Therefore, we consider a Poisson loglinear analog of the model we would consider for a RCBD with replication:

$$\log(\mu_{ijk}) = \mu + \beta_i + \alpha_j + \gamma_{ij},$$

where β_i is the effect of the ith street type, α_j an effect of the jth bike lane condition, and γ_{ij} an interaction term for these two factors.

- To avoid an overparameterized model we need to add restrictions to our model.
  - The classical sum-to-zero restrictions are $\sum_i \beta_i = \sum_j \alpha_i = 0$, $\sum_i \gamma_{ij} = 0$ for each j, and $\sum_j \gamma_{ij} = 0$ for each i.
  - Alternatively, we can set $\beta_1 = \alpha_1 = \gamma_{11} = \gamma_{21} = \gamma_{31} = 0$.
  - Other equivalent non-overparameterized versions of the model are
    $$\log(\mu_{ijk}) = \beta_i + \alpha_i I(j = 1),$$
    and
    $$\log(\mu_{ijk}) = \beta_{ij}.$$
- See bikes.R.
- Wald tests yield a significant interaction between bike lanes and street type, and separate tests of the bike lane effect in each street type suggest that bike lanes increase bike traffic on residential and main streets, but not on side streets (street type 2).

Parameter Interpretation:

With a log link, a parameter estimate $\hat{\beta}_j$ represent the effect of a unit increase in x_j on the log mean. Or, equivalently, $e^{\hat{\beta}_j}$ is the estimated multiplicative effect on μ.

- E.g., for residential streets, we would estimate that the presence of bike lanes is associated with a $\exp(1.18) = 3.26$ times higher mean number of bikes/hour.

- However, we need to be a bit careful before drawing these conclusions. Does our model fit adequately and, therefore, is it an appropriate model on which to base inference?

Overdispersion in Poisson Regression:

Overdispersion in Poisson regression is typical. In fact, it's difficult to find a true Poisson regression situation (as opposed to a Poisson loglinear model for a contingency table, say) where it is not exhibited.

- In the bikes example, the model has deviance $D = 501.82$ and Pearson statistic $X^2 = 503.45$ on 52 degrees of freedom, both indicating serious lack of fit.

  - One needs to be a little bit careful about using D and X^2 as formal χ^2 GOF tests in this setting as these data are not groupable into a contingency of fixed dimension. However, even in this context, large values of $D/d.f.$ and/or $X^2/d.f.$ are good indicators of overdispersion as long as the underlying Poisson means are large.

- In this example there are not additional covariates to consider or a more complex linear predictor that can be sensibly formulated to try to capture this lack of fit. Instead it makes sense to consider whether the data are not Poisson distributed but perhaps follow a distribution with a less restrictive mean-variance relationship.

Another example:

Example – Pump Failures:

Gaver and O'Muircheartaigh (1987) present data on the number of failures y_i and the period of operation t_i (measured in 1,000s of hours) for 10 pumps from a nuclear plant. The pumps were operated in two different modes; four being run continuously (C) and the others kept on standby (S) and only run intermittently. The data are as follows:

Mode=C		Mode=S	
y_i	t_i	y_i	t_i
5	94.320	1	15.720
5	62.880	3	5.240
14	125.760	1	1.048
19	31.440	1	1.048
		4	2.096
		22	10.480

- Unlike the bikes example, here we are counting events over differing periods of time (or *exposure* to the risk of failure). Let μ_i be the failure rate, or mean number of failures for a single unit of time. Then we assume y_i, the number of failures over t_i periods of time, follows a Poisson distribution $y_i \sim \text{Poisson}(t_i\mu_i)$ where

$$\text{E}(y_i) = t_i\mu_i$$

  and we assume μ_i follows a loglinear model that depends upon the explanatory variables. That is,

$$\log\{\text{E}(y_i)\} = \log(t_i\mu_i) = \log(t_i) + \log(\mu_i) = \log(t_i) + \mathbf{x}_i^T\boldsymbol{\beta}$$

  Here, $\log(t_i)$ is a fixed, known part of the linear predictor for $\log\{\text{E}(y_i)\}$ which we call an offset.

- See pump.sas.

- In pump.sas we first fit a simple Poisson regression model (model m1) that just allows for different rates in the two modes of operation:

$$\log(\mathrm{E}(y_i)) = \log(t_i) + \beta_0 + \beta_1 \mathrm{modeS}$$

  Here, $\exp(\beta_1)$ represents the multiplicative effect on the failure rate when pumps are on standby.

- The Wald statistic for the mode effect is $(1.882/.234)^2 = (8.08)^2 = 64.98$ indicating that there is a significant difference in the failure rates across modes ($p < .0001$). However, again the model fits poorly: $D = 71.43$ and $X^2 = 89.23$ on 8 residual df.

- The data are substantially overdispersed in comparison to a Poisson assumption, invalidating our inferences.

As in the binomial case, overdispersion can arise in several different ways. A simple approach to handling overdispersion that does not involve any assumptions about the mechanism causing the extra-Poisson variability is to simply assume

$$\mathrm{E}(Y_i) = \mu_i, \qquad \mathrm{var}(Y_i) = \phi\mu_i$$

where ϕ is an unknown positive-valued dispersion parameter not necessarily equal to 1.

- This is the QL approach, which is implemented in model m1q1. As in the binomial case, parameter estimates do not change, but we must multiply standard errors by $\sqrt{\hat{\phi}} = \sqrt{11.15} = 3.34$. Here, ϕ is estimated by the MOM estimator: $\hat{\phi} = X^2/d.f.$.

- Note that a quasi-likelihood ratio test of modeS is given by $2(ql_A - ql_0) = 2(6.73 - 4.35) = 4.76$ which is asymptotically $\chi^2(r)$ where $r = 1$ is the number of restrictions placed by the null hypothesis and ql_0 and ql_A are the maximized quasilikelihoods under the null and alternative, respectively.

  - For this test it is crucial that the overdispersion parameter remain fixed for the two qls being compared (fixed at the value estimated under H_A).

  - SAS also computes an approximate F test in addition to the chi-square statistic. This quantity is given by $2(ql_A - ql_0)/r$ and follows an approximate $F(r, \text{d.f.}_E)$ distribution. In essence this statistic accounts for the additional uncertainty introduced by having to estimate ϕ and can be expected to perform better than the chi-square test in small to medium-sized samples.

Negative Binomial Model:

To deal with overdispersion in the binomial model, we mixed a binomial with its conjugate distribution, the beta, to obtain the beta-binomial distribution.

For the Poisson distribution, its conjugate is the gamma distribution.

Suppose $Y|\lambda \sim \text{Poisson}(\lambda\mu)$ where λ is drawn from a gamma$(\alpha, 1/\alpha)$ distribution with density

$$f(\lambda) = \frac{\alpha^\alpha \lambda^{\alpha-1} e^{-\alpha\lambda}}{\Gamma(\alpha)}$$

- $\lambda \sim \text{gamma}(\alpha, 1/\alpha)$ implies

$$\text{E}(\lambda) = 1, \qquad \text{var}(\lambda) = 1/\alpha$$

Under this mixture, marginally we have

$$\Pr(Y = y) = \frac{\Gamma(y+\alpha)}{y!\Gamma(\alpha)} \left(\frac{\mu}{\mu+\alpha}\right)^y \left(\frac{\alpha}{\mu+\alpha}\right)^\alpha, \quad y = 0, 1, \ldots$$

- This is the frequency function of the negative binomial distribution and we write $Y \sim NB(\mu, \alpha)$.

- For $Y \sim NB(\mu, \alpha)$, it can be shown that

$$\mathrm{E}(Y) = \mu, \qquad \mathrm{var}(Y) = \mu + \mu^2/\alpha \qquad (*)$$

- In the negative binomial regression model, we assume a loglinear model for μ:

$$\mu = \mu(\mathbf{x}) = \exp(\mathbf{x}^T \boldsymbol{\beta})$$

- Note that the NB distribution generalizes the Poisson. For $\alpha \to \infty$ we recover the Poisson, so a test of $\theta = 0$ versus $\theta \neq 0$ where $\theta = 1/\alpha$ provides a test of the adequacy of the Poisson distribution versus a negative binomial alternative.

  - A LRT of this hypothesis can be performed by computing twice the loglikelihood for the Poisson model with twice the loglikelihood for the NB model with same linear predictor.

  - Because the null hypothesis places the parameter of interest at the boundary of the parameter space, standard asymptotics do not apply. Instead, this LRT statistic is asymptotically distributed as a random variable which is 0 half the time and $\chi^2(1)$ half of the time. $\Rightarrow$ p-value is .5 times the p-value obtained by comparing against a $\chi^2(1)$.

- More generally, it can be shown that if λ is a random variable with mean 1 and variance $1/\alpha$ (not necessarily gamma), then marginally Y has first two moments as given in (*).

- Note that the NB dispersion parameter in PROC GENMOD is parameterized in terms of $\theta = 1/\alpha$ rather than α (and the online documentation calls uses the symbol k rather than θ).

- In pump.sas we use PROC GENMOD to fit a NB version of model m1 (m1NB). The NB scale parameter $k = 1/\alpha$ is estimated via ML to yield $\hat{k} = 0.77$ or $\hat{\alpha} = 1/.77 = 1.30$.

- The LR test for differences between modes gives quite similar results to that obtained in the overdispersed Poisson (QL) analysis. Clearly, ignoring the overdispersion in the Poisson model drastically overestimates the significance of the mode effect.

- Note that by fitting a NB model in which you we fix $k = 0$, we recover the Poisson model. Fitting this model (model m1Alt in pump.sas) gives a Lagrange multiplier test (aka score test) of $H_0 : k = 0$. This provides an asymptotically $\chi^2(1)$ test of overdispersion relative to a Poisson.

- This test involves a NB alternative and therefore is most powerful relative to such an alternative, but can be regarded as a general test of overdispersion. It typically will be more powerful than a Pearson or deviance GOF test of the Poisson model and doesn't involve the non-standard asymptotics of the LRT.

Other models for overdispersed counts:

In practice, it is rare for count data to be Poisson distributed. Overdispersion is the rule rather than the exception. Overdispersed Poisson models (QL models) and NB models essentially add a dispersion parameter to account for extra-Poisson variability.

However, in some cases the nature of the overdispersion can be modeled much more specifically. One common problem in count data is the presence of many zeros. In such data sets the non-zero observations may not be especially dispersed but many more zero counts may occur than is consistent with a Poisson distribution.

In such situations a **zero-inflated Poisson** (ZIP) model may be useful.

For example, consider a study of cockroach infestation in city apartments. In this study traps were set out for t_i days in the ith of n apartments around the city. Explanatory variables including the income and ethnicity of the apartment dwellers, indicators for neighborhood, and measures of the quality of the apartment were available.

However, after building a model of the form $\{y_i\} \overset{ind}{\sim} \text{Poisson}(t_i\mu_i)$ where

$$\log \mathrm{E}(y_i) = \log t_i + \log \mu_i = \log t_i + \mathbf{x}_i^T \boldsymbol{\beta},$$

the observed number of responses with $y_i = 0$ was greatly in excess of $\sum_{i=1}^{n} \hat{\Pr}(Y_i = 0)$ according to the model.

In this context it is natural to consider a ZIP model, which assumes

$$y_i \sim \begin{cases} 0, & \text{with probability } p_i, \\ \text{Poisson}(t_i\mu_i), & \text{with probability } 1 - p_i. \end{cases}$$

Here, the idea is that some apartments are completely free of cockroaches and for these apartments the response is always 0, and other apartments have roaches which are caught at a Poisson rate of μ_i.

- Here, p_i can be thought of as the probability that the apartment is roach-free (not the probability that 0 roaches are caught since traps in apartments with roaches can happen to catch 0 by chance).

- Such a model typically assumes a loglinear model for μ_i and can assume p_i is constant or, possibly, assume a model for possible dependence of p_i on covariates as well (e.g., through a logistic model).

- Often the Poisson component is replaced by a NB, resulting in a zero-inflated negative binomial model which allows zero-inflation as well as extra-Poisson variability in the non-zero counts.

- These models are not GLMs, but can be thought of as extensions of GLMs and they can be fit using GLM tools (e.g., IRLS) and inference in these models is done similarly as in GLMs (e.g., Wald, LR and score tests, etc.).

Log-linear Models for Cross-classified Data (Read Chs.8–9 of Agresti)

Recall the three sampling models for contingency tables:

1. Poisson sampling.
   - All margins random.

   Likelihood function (two-way table):

$$L(\boldsymbol{\mu}; \mathbf{n}) = \prod_{i,j} \frac{e^{\mu_{ij}} \mu_{ij}^{n_{ij}}}{n_{ij}!}$$

2. Multinomial Sampling.
   - grand total (table margin) n is fixed.

   Likelihood:

$$L(\boldsymbol{\pi}; \mathbf{n}) = n! \prod_{i,j} \frac{\pi_{ij}^{n_{ij}}}{n_{ij}!}$$

3. Product Multinomial Sampling.
   - E.g., in a two-way table, each row corresponds to an independent multinomial distribution $\Rightarrow n_i$ (i^{th} row margin) is fixed by design for each i.

   Likelihood:

$$L(\boldsymbol{\pi}; \mathbf{n}) = \prod_{i=1}^{I} \left\{ n_i! \prod_{j=1}^{J} \frac{\pi_{j|i}^{n_{ij}}}{n_{ij}!} \right\}$$

In all three sampling models, the log-linear model can be used as the systematic component of a GLM for expected cell counts.