CHAPTER 5

LOGISTIC AND POISSON REGRESSION

5.1 LOGISTIC REGRESSION IN EPIDEMIOLOGY

Logistic regression is a special case of the binomial/Bernoulli GLM described previously where, for binary response random Y_i with canonical **logistic** (or logit) link

$$Y_i|\theta_{0i} \sim Bernoulli\left(\theta_{0i}\right) \qquad f_{Y_i|\theta_{0i}}\left(y_i;\theta_{0i}\right) = \theta_{0i}^{y_i}\left(1-\theta_{0i}\right)^{1-y_i} \qquad \text{logit}\left(\theta_{0i}\right) = \log\left(\frac{\theta_{0i}}{1-\theta_{0i}}\right) = x_i^T\beta$$

For this model, the expected response μ_i is

$$\mu_i = \theta_{0i} = \frac{\exp\left\{x_i^T\beta\right\}}{1+\exp\left\{x_i^T\beta\right\}}. \tag{5.1.1}$$

The deviance for the logistic regression GLM (with dispersion parameters are given by $\phi = \omega_i = 1$) is

$$D = -2\sum_{i=1}^{n} y_i \log\left(\widehat{\theta}_i\right) + \left(1-y_i\right)\log\left(1-\widehat{\theta}_i\right) = 2\sum_{i=1}^{n}\left[\log\left(1+\exp\left\{x_i^T\widehat{\beta}\right\}\right) - y_i x_i^T\widehat{\beta}\right]$$

5.1.1 RELATIVE RISKS, ODDS AND ODDS RATIOS FOR 2×2 TABLES

For the logistic regression GLM, results relating to odds and odds ratios are straightforward. From (5.1.1), we have the **odds on occurrence** or **odds on incidence**

$$\frac{P\left[Y_i=1\right]}{P\left[Y_i=0\right]} = \frac{\theta_i}{1-\theta_i} = \exp\left\{x_i^T\beta\right\}$$

and for two different predictor vectors x_i and x_j, the **odds ratio**

$$\psi = \frac{P\left[Y_i=1\right]/P\left[Y_i=0\right]}{P\left[Y_j=1\right]/P\left[Y_j=0\right]} = \frac{\theta_i/\left(1-\theta_i\right)}{\theta_j/\left(1-\theta_j\right)} = \frac{\exp\left\{x_i^T\beta\right\}}{\exp\left\{x_j^T\beta\right\}} = \exp\left\{\left(x_i-x_j\right)^T\beta\right\} \tag{5.1.2}$$

$$\log\psi = \left(x_i-x_j\right)^T\beta$$

Now, suppose that x_i and x_j (both vectors of $K+1$ elements) differ only in element k; then

$$\psi = \exp\left\{\left(x_i-x_j\right)^T\beta\right\} = \exp\left\{\left(x_{ik}-x_{jk}\right)\beta_k\right\} \qquad \Longrightarrow \qquad \log\psi = \left(x_{ik}-x_{jk}\right)\beta_k$$

and thus the odds ratios for different predictor values is available in a straightforward form. Thus the log-odds ratio is a linear function of the predictor and the coefficient. The odds ratio for **discrete (factor) predictors** in the logistic GLM is particularly straightforward; suppose that a factor has I levels, parameterized in the linear predictor using parameters $\alpha_1,...,\alpha_I$. Then the odds ratio between different levels, i_1 and i_2 of this factor is, from (5.1.2)

$$\psi = \exp\left\{\alpha_{i_1}-\alpha_{i_2}\right\} \qquad \text{so that} \qquad \log\psi = \alpha_{i_1}-\alpha_{i_2}$$

COHORT STUDY EXAMPLE : The data from such a typical cohort study can be represented in the usual way by a 2×2 table with entries $(n_{11}, n_{12}, n_{21}, n_{22})$, and our representation has been to say that $P(F|E) = \pi_1$ and $P(F|E') = \pi_0$, with

$$\text{ODDS ON INCIDENCE} \quad \omega_1 = \frac{\pi_1}{1-\pi_1} \qquad \omega_0 = \frac{\pi_0}{1-\pi_0}$$

$$\text{ODDS ON EXPOSURE} \quad \Omega_1 = \frac{P(E|F)}{P(E'|F)} \qquad \Omega_0 = \frac{P(E|F')}{P(E'|F')} \qquad \Longrightarrow \text{ODDS RATIO } \psi = \frac{\omega_1}{\omega_0} = \frac{\Omega_1}{\Omega_0}$$

Here, then, EXPOSURE is a predictor factor with $I = 2$ levels ($E = 0, 1$) say, and the response is Binomial/Bernoulli, with $Y_i = 1$ corresponding to a "case". The typical model used is a **product binomial**, that is, the exposed and unexposed groups are independent with

$$\begin{aligned} Y_i \mid (E=1) &\sim Bernoulli(\pi_1) \quad \Longrightarrow \quad N_{11} \mid (E=1) \sim Binomial(n_{\cdot 1}, \pi_1) \\ Y_i \mid (E=0) &\sim Bernoulli(\pi_0) \quad \Longrightarrow \quad N_{12} \mid (E=0) \sim Binomial(n_{\cdot 2}, \pi_0) \end{aligned}$$

that yields the usual mles $\widehat{\pi}_1 = \dfrac{n_{11}}{n_{\cdot 1}}$ and $\widehat{\pi}_0 = \dfrac{n_{12}}{n_{\cdot 2}}$, so that

$$\begin{aligned} \text{ODDS ON INCIDENCE :} \quad \widehat{\omega}_1 &= \frac{n_{11}}{n_{21}} \qquad \widehat{\omega}_0 = \frac{n_{12}}{n_{22}} \\ &\Longrightarrow \quad \text{ODDS RATIO :} \quad \widehat{\psi} = \frac{n_{11} n_{22}}{n_{12} n_{21}} \\ \text{ODDS ON EXPOSURE :} \quad \widehat{\Omega}_1 &= \frac{n_{11}}{n_{12}} \qquad \widehat{\Omega}_0 = \frac{n_{21}}{n_{22}} \end{aligned}$$

In a logistic GLM setting, a natural model would (in the conventional notation)

$$\text{For } E = 0 : \log\left(\frac{\pi_0}{1-\pi_0}\right) = \mu \qquad \text{For } E = 1 : \log\left(\frac{\pi_1}{1-\pi_1}\right) = \mu + \alpha$$

Hence the **relative risk** or **rate ratio** is

$$\frac{\pi_1}{\pi_0} = \frac{\exp\{\mu+\alpha\}}{1+\exp\{\mu+\alpha\}} \times \frac{1+\exp\{\mu\}}{\exp\{\mu\}} = \frac{\exp\{\alpha\}+\exp\{\mu+\alpha\}}{1+\exp\{\mu+\alpha\}}$$

and so that the **odds on incidence** in the unexposed/exposed groups and **odds ratio** and **log odds ratio** are

$$\omega_0 = \frac{\pi_0}{1-\pi_0} = \exp\{\mu\} \qquad \omega_1 = \frac{\pi_1}{1-\pi_1} = \exp\{\mu+\alpha\} \qquad \psi = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)} = \exp\{\alpha\} \therefore \log\psi = \alpha$$

NOTE: Remember that, in **cohort** studies, all of these quantities are estimable. In **case control** studies, the **odds on exposure** in the cases and in the controls, that is, (Ω_1, Ω_0) are estimable in the usual way, but that the the parameters relating to **odds on incidence** in the cases and in the controls (ω_1, ω_0), as well as **relative risk** are not; in practice the **rare disease hypothesis** is applied, where it is assumed that, for example

$$\frac{\pi_1}{\pi_0} \approx \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}$$

as $(1-\pi_1)$ and $(1-\pi_0)$ are approximately 1. In addition, if the exposure rate in the population is **known** and equal (to θ_E say) in the cases and in the controls, we have that

$$\frac{P(F|E)}{P(F|E')} = \frac{P(E)}{P(E')}\frac{P(E|F)}{P(E|F')} \Longrightarrow \frac{\pi_1}{\pi_0} = \left(\frac{1-\theta_E}{\theta_E}\right)\Omega_1$$

and the estimate of the relative risk is

$$\left(\frac{1-\theta_E}{\theta_E}\right)\widehat{\Omega}_1$$

5.1.2 EXTENSION TO MORE THAN 2 EXPOSURE LEVELS

In the GLM framework, the extension to models for more than two exposure levels is straightforward; if exposure has I_E levels (labelled $0, 1, 2, ...J_E - 1$ say), with incidence rate π_e for exposure level $e = 0, 1, 2, ..., J_E - 1$, the models can be extended

- **FACTOR PREDICTORS:** The simplest logistic regression GLM has

$$\begin{aligned} e &= 0 : \log\left(\frac{\pi_0}{1-\pi_0}\right) = \mu \\ e &> 0 : \log\left(\frac{\pi_e}{1-\pi_e}\right) = \mu + \alpha_e \end{aligned}$$

  a model with J_E parameters, so that the odds ratios for level e against level 0 is

$$\psi_e = \frac{\pi_e/(1-\pi_e)}{\pi_0/(1-\pi_0)} = \exp\{\alpha_e\}$$

  and for level e_1 against level e_2 is

$$\psi_{e_1,e_2} = \frac{\pi_{e_1}/(1-\pi_{e_1})}{\pi_{e_2}/(1-\pi_{e_2})} = \exp\{\alpha_{e_1} - \alpha_{e_2}\}.$$

- **CONTINUOUS PREDICTORS**: If exposure can be represented as a **continuous** predictor, assuming a common amount of exposure, x_e, within an exposure level, then the model becomes

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

  that is, a model with only 2 parameters. For this model,

$$\psi_{e_1,e_2} = \exp\{(x_{e_1} - x_{e_2})\beta_1\}.$$

  Finally, if the exposure level is known individually to be x_i for subject i, then the model becomes

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

  and we have the standard regression model.

- **ORDINAL PREDICTORS: THE CONSTANT ADJACENT ODDS RATIO MODEL** A common modelling situation arises when the levels of the discrete predictor are **ordered** (the variable is termed an **ordinal** predictor variable); in this case, using the above models we retain the structure of the continuous model,

$$\log\left(\frac{\pi_e}{1-\pi_e}\right) = \alpha_0 + \beta_0 e$$

  and thus, for $e > 0$

$$\psi_{e,e-1} = \frac{\pi_e/(1-\pi_e)}{\pi_{e-1}/(1-\pi_{e-1})} = \frac{\exp\{\alpha_0 + \beta_0 e\}}{\exp\{\alpha_0 + \beta_0(e-1)\}} = \exp\{\beta_0\}.$$

  that is, identical for all e. This is the **constant adjacent odds ratio model**

5.1.3 EXTENSION TO STRATIFIED EXPOSURE

Suppose now that we wish to study the influence of exposure in the presence of a potential **confounding** factor; suppose that factor A with levels $a = 0, 1, ..., K_A - 1$ is also believe to modify risk in the presence or absence of exposure. In standard notation we might fit the following two factor models:

(I) E : a common incidence rate for all levels of the confounder in exposed and unexposed groups

(II) A: a common incidence rate for all **exposure** levels modified by different levels of the **confounder**

(III) $E + A$: a common, additive modification of incidence rate in the exposed and unexposed groups by the **confounder**

(IV) $E * A$: an interaction between **exposure** and **confounder**

These models can be represented as follows:

(I) K_E parameters

$$\log\left(\frac{\pi_e}{1-\pi_e}\right) = \begin{cases} \mu & e = 0 \\ \mu + \alpha_e^E & e > 0 \end{cases}$$

(II) K_A parameters

$$\log\left(\frac{\pi_a}{1-\pi_a}\right) = \begin{cases} \mu & a = 0 \\ \mu + \alpha_a^A & a > 0 \end{cases}$$

(III) $K_E + K_A - 1$ parameters

$$\log\left(\frac{\pi_{ea}}{1-\pi_{ea}}\right) = \begin{cases} \mu & e = 0, a = 0 \\ \mu + \alpha_e^E & e = 1, ..., K_E - 1, a = 0 \\ \mu + \alpha_a^A & e = 0, a = 1, ..., K_A - 1 \\ \mu + \alpha_e^E + \alpha_a^A & e = 1, ..., K_E - 1, a = 1, ..., K_A - 1 \end{cases}$$

(IV) $K_E K_A$ parameters

$$\log\left(\frac{\pi_{ea}}{1-\pi_{ea}}\right) = \begin{cases} \mu & e = 0, a = 0 \\ \mu + \alpha_e^E & e = 1, ..., K_E - 1, a = 0 \\ \mu + \alpha_a^A & e = 0, a = 1, ..., K_A - 1 \\ \mu + \alpha_e^E + \alpha_a^A + \gamma_{ea}^{EA} & e = 1, ..., K_E - 1, a = 1, ..., K_A - 1 \end{cases}$$

Taking Model 4 as the most general model, we can inspect odds ratios for any combination of two levels of exposure and any two levels of the confounder; if $K_E = 2$ (exposed and unexposed) we have that for $E = 0$, the log odds ratio for confounder level a_1 against confounder level $a_2 > 0$ is

$$\log \psi_{a_1,a_2}^{(0)} = \left(\mu + \alpha_{a_1}^A\right) - \left(\mu + \alpha_{a_2}^A\right) = \alpha_{a_1}^A - \alpha_{a_2}^A$$

and for $E = 1$

$$\log \psi_{a_1,a_2}^{(1)} = \left(\mu + \alpha_1^E + \alpha_{a_1}^A + \gamma_{1a_1}^{EA}\right) - \left(\mu + \alpha_1^E + \alpha_{a_2}^A + \gamma_{1a_2}^{EA}\right) = \left(\alpha_{a_1}^A - \alpha_{a_2}^A\right) + \left(\gamma_{1a_1}^{EA} - \gamma_{1a_2}^{EA}\right)$$

and thus if

$$\left(\gamma_{1a_1}^{EA} - \gamma_{1a_2}^{EA}\right) = 0$$

then there is **no interaction** between exposure and confounder for factor levels a_1 and a_2. Broadly, if Model 3 fits as adequately as Model 4, then there is no prospect of confounding.

5.1.4 EXTENSION TO POLYTOMOUS RESPONSE

The data in the logistic regression models above can be presented in the form of a 2×2 table, or more generally $2 \times J$ table, with the two rows corresponding to the two outcomes F and F' $(0/1)$. Thus the response is **dichotomous.** A more general form of data has a **polytomous response**, where the outcomes are class labels $0, 1, 2, ..., K$; the responses themselves can be

- **ordinal** or ordered, that is, where $0 < 1 < 2 < ... < K$ in some sense specific to the response concerned. For example, in a study measuring pain severity as a response, we may have responses 0,1,2,3 corresponding to NONE, MILD, MODERATE, SEVERE.
- **nominal,** or unordered, where the labels are simply that, and have no intrinsic meaning.

Maintaining the J levels of the exposure factor as the most general case, we thus have a data array which is $(K+1) \times J$, for example for $K = 3$ and $J = 6$

	EXPOSURE LEVEL					
RESPONSE	1	2	3	4	5	6
0	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	n_{16}
1	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	n_{26}
2	n_{31}	n_{32}	n_{24}	n_{25}	n_{35}	n_{36}
3	n_{41}	n_{42}	n_{43}	n_{44}	n_{45}	n_{46}

and typically we assume that the data are independent within the columns of the table. Let

$$\pi_{kj} = P\left[\text{Response is } k | \text{Exposure level is } J\right] \qquad k = 0, 1, ..., K$$

be the conditional probability of response k for exposure level j. Then $\sum_{j=0}^{K} \pi_{kj} = 1$ for each j, and the counts in column j of the table have a **multinomial** rather than **binomial** distribution, with joint mass function

$$\binom{n_{\cdot j}}{n_{1j}, x_{2j}, ..., x_{I+1,j}} \pi_{0j}^{n_{1j}} \pi_{1j}^{n_{2j}} ... \pi_{Kj}^{n_{K+1,j}}.$$

Many GLMs can be fitted to these data, each with different form of the linear predictor. For a single exposure level

- **RELATIVE RISK MODEL** with reference category 0

$$\frac{\pi_k}{\pi_0} = \exp\left\{x^T \beta_k\right\}$$

- **ADJACENT CATEGORY RESPONSE MODEL:**

$$\frac{\pi_k}{\pi_{k-1}} = \exp\left\{x^T \beta_k\right\}$$

- **PROPORTIONAL ODDS MODEL**

$$\frac{P\left[Y > k\right]}{P\left[Y \leq k\right]} = \frac{\sum_{j=k+1}^{K} \pi_j}{\sum_{j=0}^{k} \pi_j} = \exp\left\{x^T \beta_k\right\}$$

- **CONTINUATION MODEL**

$$\frac{P\left[Y > k\right]}{P\left[Y = k\right]} = \frac{\sum_{j=k+1}^{K} \pi_j}{\pi_k} = \exp\left\{x^T \beta_k\right\}$$

5.2 POISSON REGRESSION

Poisson regression is the term applied to the Poisson GLM described previously where, for non-negative integer response variable Y_i where, typically, the canonical **log** link is used

$$Y_i|\lambda_{0i} \sim Poisson\left(\lambda_{0i}\right) \qquad f_{Y_i|\lambda_{0i}}\left(y_i;\lambda_{0i}\right) = \frac{\exp\left\{-\lambda_{0i}\right\}\lambda_{0i}^{y_i}}{y_i!} \qquad \log\left(\lambda_{0i}\right) = x_i^T\beta$$

The expected response μ_i is $\mu_i = \lambda_{0i} = \exp\left\{x_i^T\beta\right\}$. The deviance for the Poisson regression GLM (dispersion parameters $\phi = \omega_i = 1$).is

$$D = -2\sum_{i=1}^{n}\left[y_i\left(x_i^T\widehat{\beta} - \log y_i\right) - \left(\exp\left\{x_i^T\widehat{\beta}\right\} + y_i\right)\right]$$

5.2.1 OFFSET MODELS FOR POISSON DISTRIBUTED DATA

An underlying model for Poisson count data is the Poisson process model; we assume that, conditional on predictor variables x_i, the **incidence rate** for occurrence of the disease is $\lambda_i = \exp\left\{x_i^T\beta\right\}$, and thus in a total exposure time T_i for that subset of individuals, we have that the random variable recording the number of incidences, Y_i, has a Poisson distribution

$$Y_i|\lambda_i, T_i \sim Poisson\left(T_i\lambda_i\right).$$

This is part motivation for the reporting of incidence rates

$$\widehat{\lambda}_i = \frac{y_i}{T_i}$$

where T_i is the number of person years (total time on study) for category i. In this case, T_i is a **constant**, so in the model for λ_i, the linear predictor should be combined with an **offset**, that is

$$\log\lambda_i = x_i^T\beta + \log T_i$$

and again we explain the systematic variation in the rate via the link and covariate. In SPLUS, the offset model is fitted routinely as follows;

$$\texttt{glm(y}\tilde{}\texttt{factor(exposure)} + \texttt{offset(log(T))}, \texttt{family} = \texttt{poisson(link} = \texttt{log))}$$

5.2.2 POISSON APPROXIMATIONS TO BINOMIAL SAMPLING MODELS

The offset can also be used to **standardize** for other factors, or rather to reconstruct a hypothetical **standard population** within which the incidences are to be observed. Suppose that, after appropriate time/category standardization, a binomial model for incidences is deemed appropriate

$$Y_i|\theta_i, N_i \sim Binomial\left(N_i, \theta_i\right)$$

which we approximate by

$$Y_i|\lambda_i, T_i \sim Poisson\left(N_i\theta_i\right).$$

and again, if $\lambda_i = N_i\theta_i$, so that $\log\lambda_i = \log\theta_i + \log N_i$

$$\log\lambda_i = x_i^T\beta + \log N_i$$

5.3 MODEL SELECTION AND VALIDATION

There are three general issues/techniques that can be used to motivate the choice of a particular model in a linear/generalized linear model setting. The first is the general principal of **parsimony**, the second is the use of **deviance** as a model validation measure, and the third is **information criteria**, where maximized log-likelihood values are used as the basis for model comparison, after appropriate adjustment for **model complexity** is made.

5.3.1 PARSIMONY

The principal of parsimony encapsulates the idea that simpler models (models with fewer parameters) are preferable to more complex models (with larger numbers of parameters); there are several reasons why this is so

- simpler models may be easier to fit
- simpler models are easier to interpret; for example, main effects only models are more straightforward to interpret than models with interaction terms.
- simpler models may have better "**out-of-sample**" prediction properties. This is because complex models may **overfit** the observed data, to the detriment of the models' prediction ability. In an extreme case, when the **saturated** model is used to fit the observed data exactly, the model has no prediction ability at all.

5.3.2 DEVIANCE-BASED MEASURES

Recall that, for the scaled deviance, we have stated that

$$D^*\left(y;\widehat{\theta},\phi\right)=\frac{D\left(y;\widehat{\theta}\right)}{\phi}\sim\chi^2_{n-d} \tag{5.3.3}$$

if the linear predictor has d parameters (for a model with K predictors and an intercept, $d=K+1$) This result is **exact** in the linear model, but only ever **approximate** in the GLM, with the approximation being occasionally quite poor (see Note 1 below). Nevertheless, it motivates the use of deviance as a model adequacy assessment, as it implies that, as the properties of the χ^2_{n-d} distribution include an expectation of $n-d$, we would expect a adequately fitting model to have

$$D^*\left(y;\widehat{\theta},\phi\right)\approx n-d \qquad \text{or} \qquad \frac{D^*\left(y;\widehat{\theta},\phi\right)}{n-d}\approx 1$$

For the binomial and Poisson models, $\phi=1$.

NOTES:

1. The approximate result presented in (5.3.3) breaks down in two ways in some GLMs. First, the Chi-squared approximation itself becomes compromised. Secondly, the distribution **conditional on true parameter** θ of random variable

   $$D^*\left(Y;\widehat{\theta},\phi\right)$$

   does depend on the θ, rather than being **independent** of θ as implied by (5.3.3). For some models the distribution of D^* can be investigated using simulation.

2. The value of the deviance presented can depend on the way the data are analyzed. Consider a Bernoulli/binomial example for an exposure factor having two levels ($i = 0, 1$); suppose the data in the two (unexposed and exposed) cohorts are represented as $\{y_{ij} : i = 0, 1, j = 1, 2, ..., n_i\}$, and let $s_i = \sum_{j=1}^{n_i} y_{ij}$, $s = s_0 + s_1$, and $n = n_0 + n_1$. The **individualized** likelihood, and the three possible models are

$$\prod_{i=0}^{1} \left\{ \prod_{j=1}^{n_i} \pi_{ij}^{y_{ij}} \left(1 - \pi_{ij}\right)^{1-y_{ij}} \right\}.$$

NULL	$\widehat{\pi} = \frac{s}{n}$	$d = 1$	$D = -2\left[s \log \widehat{\pi} + (n-s) \log\left(1 - \widehat{\pi}\right)\right]$
EXPOSURE FITTED	$\widehat{\pi}_i = \frac{s_i}{n_i}$	$d = 2$	$D = -2\left[\sum_{i=0}^{1} s_i \log \widehat{\pi}_i + (n_i - s_i) \log\left(1 - \widehat{\pi}_i\right)\right]$
SATURATED	$\widehat{\pi}_{ij} = y_{ij}$	$d = n$	$D = 0$

Now suppose that the individual information is to be disregarded and a **pooled** analysis is to be carried out. We now have only two data points, the pair (s_0, s_1), and a likelihood

$$\prod_{i=0}^{1} \left\{ \binom{n_i}{s_i} \pi_i^{s_i} \left(1 - \pi_i\right)^{n_i - s_i} \right\}$$

The only model available now are the NULL, with $\widehat{\pi} = s/n$, $d = 1$, and deviance $-2\left[s \log \widehat{\pi} + (n-s) \log\left(1\right.\right.$ and the SATURATED where $\widehat{\pi}_i = s_i/n_i$, $d = 2$, and likelihood

$$\prod_{i=0}^{1} \left\{ \binom{n_i}{s_i} \left(\frac{s_i}{n_i}\right)^{s_i} \left(1 - \frac{s_i}{n_i}\right)^{n_i - y_{ij}} \right\} \neq 1 \qquad \therefore \qquad l_S\left(\widehat{\pi}_0, \widehat{\pi}_1\right) \neq 0$$

Note that the estimates for the pooled and individual analyses are **identical.** Note also, however, that the two null deviances, $D_0^{(I)}$ and $D_0^{(P)}$ for the individual and pooled analyses respectively are different from each other; in fact

$$D_0^{(P)} = D_0^{(I)} + 2 \sum_{i=0}^{1} \left[s_i \log \widehat{\pi}_i + (n_i - s_i) \log\left(1 - \widehat{\pi}_i\right)\right]$$

3. The **Pearson-type** deviance/goodness of fit measure

$$\chi^2\left(y, \widehat{\theta}\right) = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\mu}_i\right)^2}{V\left(\hat{\mu}_i\right)/w_i}$$

can also be used as a model assessment quantity, as again, approximately

$$\chi^2\left(Y, \widehat{\theta}\right) \sim \chi^2_{n-d}$$

although, as before, the chi-squared approximate distribution and the independence of the true value of θ may not hold

5.3.3 INFORMATION CRITERIA

A common method for model assessment utilizes **Information Criteria**. The central idea is that the maximized log-likelihood value should be a good indicator of model fit, but needs to be **penalized** with respect to the model complexity, that is, the number of parameters fitted.

Suppose that the fit of model with canonical parameters θ_i (or β, where $g(h(\theta_i)) = x_i^T\beta$) is to be assessed. Writing the maximized log likelihood as $l(\widehat{\beta}) = \log f_{Y|\beta}(y;\widehat{\beta})$, the better the fit of the model, the larger $l(\widehat{\beta})$ will be. However, the number of parameters fitted will contribute to the magnitude of $l(\widehat{\beta})$ (if more parameters are fitted, the larger we might expect $l(\widehat{\beta})$ to be), and so to obtain a fair model assessment criterion, we need to penalize $l(\widehat{\beta})$ in some way. We define the **Information Criterion (IC)** for model M with d-dimensional parameter vector β_M , by

$$\mathbf{IC}_M = -2l_M\left(\widehat{\beta}_M\right) + c(n,d)$$

for some function $c(n,d)$. There are two common choices for $c(n,d)$;

1. **AKAIKE INFORMATION CRITERION (AIC)**: $c(n,d) = 2d$, so that

$$\mathbf{AIC}_M = 2\left[-l_M\left(\widehat{\beta}_M\right) + d\right]$$

For example, in the **Normal Linear Model** with design matrix X,

$$l_M\left(\widehat{\beta}_M\right) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(y - X\widehat{\beta}_M\right)^T\left(y - X\widehat{\beta}_M\right)$$

so that if σ^2 is **known**

$$\mathbf{AIC}_M = \frac{\text{RSS}}{\sigma^2} + 2d + const$$

where RSS$= \left(y - X\widehat{\beta}_M\right)^T\left(y - X\widehat{\beta}_M\right)$. If σ^2 is **unknown**, the ML estimate

$$\widehat{\sigma}^2 = \frac{1}{n}\left(y - X\widehat{\beta}_M\right)^T\left(y - X\widehat{\beta}_M\right)$$

can be used, so that

$$\mathbf{AIC}_M = n\log\left(\frac{\text{RSS}}{n}\right) + 2d + const$$

2. **BAYES INFORMATION CRITERION (BIC)**: $c(n,d) = 2d\log n$, so that

$$\mathbf{BIC}_M = 2\left[-l_M\left(\widehat{\beta}_M\right) + d\log n\right]$$

In both cases, the model with the smallest **IC** is to be selected. The advantage of the BIC is that it is a **consistent** model selection procedure, in that it selects the correct model (when a correct model is amongst the models available).

5.4 OVERDISPERSION IN GLMS

Overdispersion is a common phenomenon in the GLM analysis of observed data. We concentrate here on Binomial and Poisson distributed data for illustration; deviance formulae have been established for these models previously, and in both models, we have the dispersion parameter $\phi = 1$. We have also established that two estimates of ϕ exist for the exponential-dispersion family with K parameters

$$f_{Y_i|\theta,\phi}(y_i;\theta_i,\phi) = \exp\left\{\frac{w_i\left(y_i\theta_i + c(\theta_i)\right)}{\phi} + d\left(y_i, \frac{\phi}{w_i}\right)\right\}.$$

(where the coefficients or **weights** w_i are fixed constants), namely

$$\widehat{\phi}_D = \frac{D(y,\widehat{\theta})}{n-K} \qquad \widehat{\phi}_P = \frac{1}{n-K}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{V\left(\hat{\mu}_i\right)/w_i}.$$

We have established previously that, for the Binomial and Poisson models, $\widehat{\phi}_D$ and $\widehat{\phi}_P$ should be 1 approximately. This is not always the case however. If the estimate of ϕ is much greater than 1, we might conclude that the model does not fit the data very well. We now examine possible sources of overdispersion, and possible inferential solutions.

5.4.1 POSSIBLE SOURCES OF OVERDISPERSION

Overdispersion (relative to a fitted model) in Binomial and Poisson data can arise in several ways. In fact, it may be that apparent overdispersion can be eliminated with more careful analysis. Apparent overdispersion, diagnosed with reference to $\widehat{\phi}_D, \widehat{\phi}_P \gg 1$ may be eliminated in the following ways:

- There may be an available **systematic** component (exposure/predictor/confounder, main effect or interaction) that has been omitted from the model.
- The **link function** may be inappropriately chosen, and another link may reduce the dispersion estimate
- There may be **atypical** or **outlying** observations that are present, for which the prediction model is inadequate despite the fact that the model might be perfectly adequate elsewhere.
- It might be that the **chisquared approximation** assumed for deviance in GLMs may not be appropriate, and the estimators **biased,** so that apparent overdispersion is not actually present

5.4.2 MODELS FOR OVERDISPERSED BINOMIAL DATA

If all of the possibilities above have been eliminated, indicating genuine overdispersion, it may be necessary to develop extensions to the models used We do this here in the binomial case; suppose that we have a model with a total of I exposure/confounder cross-classification categories with incidence probabilities $\pi_0, \pi_1, ..., \pi_{I-1}$,observed counts $y_0, ..., y_{I-1}$ and total samples of size $n_0, ..., n_{I-1}$, so that a basic model is

$$Y_i \sim Binomial\left(n_i, \pi_i\right) \qquad i = 0, 1, ..., I-1$$

and a canonical logit link to log-odds parameters, say

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = x_i^T\beta$$

with d dimensional parameter vector β might be used. To represent overdispersion, we might have instead that, π_i is itself taken as a **random variable** with fixed first and second moments, for example

$$E_{\pi_i|\xi_i}[\pi_i] = \xi_i \qquad Var_{\pi_i|\xi_i}[\pi_i] = \kappa\xi_i(1-\xi_i)$$

where now constants ξ_i are related to the log-odds

$$\log\left(\frac{\xi_i}{1-\xi_i}\right) = x_i^T\beta.$$

The model implied for observable count Y_i, **conditional** on π_i remains

$$E_{Y_i|\pi_i}[Y_i|\pi_i] = n_i\pi_i \qquad Var_{Y_i|\pi_i}[Y_i|\pi_i] = n_i\pi_i(1-\pi_i)$$

but now, **unconditional of π_i but conditional on ξ_i** we have, by iterated expectation

$$E_{Y_i|\xi_i}[Y_i|\xi_i] = E_{\pi_i|\xi_i}\left[E_{Y_i|\pi_i}[Y_i|\pi_i]|\xi_i\right] = E_{\pi_i|\xi_i}[n_i\pi_i|\xi_i] = n_i\xi_i$$

and

$$\begin{aligned}
Var_{Y_i|\xi_i}[Y_i|\xi_i] &= E_{\pi_i|\xi_i}\left[Var_{Y_i|\pi_i}[Y_i|\pi_i]|\xi_i\right] + Var_{\pi_i|\xi_i}\left[E_{Y_i|\pi_i}[Y_i|\pi_i]|\xi_i\right] \\
&= E_{\pi_i|\xi_i}[n_i\pi_i(1-\pi_i)|\xi_i] + Var_{\pi_i|\xi_i}[n_i\pi_i|\xi_i] \\
&= n_i\left[E_{\pi_i|\xi_i}[\pi_i|\xi_i] - E_{\pi_i|\xi_i}\left[\pi_i^2|\xi_i\right]\right] + n_i^2\kappa\xi_i(1-\xi_i) \\
&= n_i\left[\xi_i - \left\{Var_{\pi_i|\xi_i}[\pi_i|\xi_i] + \left\{E_{\pi_i|\xi_i}[\pi_i|\xi_i]\right\}^2\right\}\right] + n_i^2\kappa\xi_i(1-\xi_i) \\
&= n_i\left[\xi_i - \left\{Var_{\pi_i|\xi_i}[\pi_i|\xi_i] + \left\{E_{\pi_i|\xi_i}[\pi_i|\xi_i]\right\}^2\right\}\right] + n_i^2\kappa\xi_i(1-\xi_i) \\
&= n_i\left[\xi_i - \left\{\kappa\xi_i(1-\xi_i) + \xi_i^2\right\}\right] + n_i^2\kappa\xi_i(1-\xi_i) \\
&= n_i\xi_i(1-\xi_i)\left\{1 + (n_i-1)\kappa\right\}
\end{aligned}$$

and thus the implied variance in response is effectively **inflated** by a factor of $\{1+(n_i-1)\kappa\}$.

This derivation implies that a plausible model for overdispersed data might be

$$Var_{Y_i|\pi_i}[Y_i|\pi_i] = \phi n_i\pi_i(1-\pi_i)$$

for some dispersion parameter ϕ in the special case $n_i = n$, that is, we should set

$$\phi = \{1+(n-1)\kappa\}.$$

In this model, with fitted parameters we have

$$\widehat{\mu}_i = n_i\widehat{\pi}_i \qquad V(\widehat{\mu}_i) = \phi n_i\widehat{\pi}_i(1-\widehat{\pi}_i)$$

the Pearson-type goodness of fit measure is

$$\chi^2\left(y,\widehat{\theta}\right) = \frac{1}{\phi}\sum_{i=0}^{I-1}\frac{(y_i - n_i\widehat{\pi}_i)^2}{n_i\widehat{\pi}_i(1-\widehat{\pi}_i)} \tag{5.4.4}$$

with $n_i = n$, where approximately $\chi^2\left(Y,\widehat{\theta}\right) \sim \chi^2_{I-d}$. In the case where the n_is are **not** identical, the formula in (5.4.4) can also be used, this is, essentially, the argument used previously to justify the estimators ϕ_D and ϕ_P.

5.4.3 MODEL SELECTION FOR OVERDISPERSED DATA

Previously we have seen how to choose between two models on the basis of **deviance** comparison; for models M_1 and M_2 with numbers of parameters $d_1 < d_2$, we suggested that

$$\frac{D_{M_1}(y, \theta_{M_1}) - D_{M_2}(y, \theta_{M_2})}{\phi} \sim \chi^2_{d_2 - d_1}$$

indicating an estimator of dispersion parameter of the form

$$\widehat{\phi} = \frac{D_{M_1}\left(Y, \widehat{\theta}_{M_1}\right) - D_{M_2}\left(y, \widehat{\theta}_{M_2}\right)}{d_2 - d_1} \tag{5.4.5}$$

may be appropriate, where D_{M_i} is either ordinary of Pearson-type deviance. In fact, our assessment of the presence of overdispersion may be more appropriately done via (5.4.5); we are only concerned about overdispersion relative to our model selection task.

A different approach inspects the **ratio** of scaled deviances: To compare the (relative) fits of the two models we might inspect

$$\frac{\left[D_{M_1}\left(Y, \widehat{\theta}_{M_1}\right) - D_{M_2}\left(y, \widehat{\theta}_{M_2}\right)\right] / \left[\phi\left(d_2 - d_1\right)\right]}{D_{M_2}\left(y, \widehat{\theta}_{M_2}\right) / \left[\phi\left(n - d_2\right)\right]} = \frac{\left[D_{M_1}\left(Y, \widehat{\theta}_{M_1}\right) - D_{M_2}\left(y, \widehat{\theta}_{M_2}\right)\right] / \left(d_2 - d_1\right)}{D_{M_2}\left(y, \widehat{\theta}_{M_2}\right) / \left(n - d_2\right)}$$

which, approximately has a $Fisher\left(d_2 - d_1, n - d_2\right)$ distribution; note that the dispersion parameter has **cancelled**. This result is directly analogous to the ANOVA F-test results described earlier.

5.4.4 FITTING THE OVERDISPERSION MODEL: QUASI-LIKELIHOOD

The estimates $\widehat{\pi}_i = g\left(x_i^T \widehat{\beta}\right)$ that appear in, for example, (5.4.4) cannot be computed by the usual iterative (IRLS) algorithms (see handout); effectively there is no likelihood function, as we have only specified **moments** (expectation and variance) of the distribution of observables Y_i. Inference mechanisms for such models are referred to as **quasi-likelihood** approaches, and a brief introduction is given below.

The **score equations** for a GLM in the exponential dispersion family with

- expectation μ_i and variance $\phi V\left(\mu_i\right) / w_i$
- link function g
- linear predictor $x_i^T \beta$

(so that $g\left(\mu_i\right) = x_i^T \beta$) take the form

$$\sum_{i=1}^{n} \frac{w_i\left(y_i - \mu_i\right)}{\phi V\left(\mu_i\right)} \frac{x_{ik}}{g'\left(\mu_i\right)} = 0 \qquad\qquad k = 1, ..., K. \tag{5.4.6}$$

Note that these are a form of **unbiased estimating equation,** that is

$$\sum_{i=1}^{n} G_i\left(Y_i, \beta\right) = 0$$

with

$$E_{Y_i|\beta}\left[G_i\left(Y_i,\beta\right)\right]=0 \qquad \text{if} \qquad E_{Y_i|\beta}\left[Y_i\right]=\mu_i=g^{-1}\left(x_i^T\beta\right)$$

Suppose now that we retain the essential moment components of the model,

$$E\left[Y_i\right]=\mu_i \qquad\qquad Var\left[Y_i\right]=\phi_i V\left(\mu_i\right) \tag{5.4.7}$$

where $\phi_i=\phi/w_i$, but **no additional distributional assumptions are to be made.**

In the exponential family, if

$$l(\theta,\phi)=\sum_{i=1}^{n}\left[\frac{\left(y_i\theta_i+c(\theta_i)\right)}{\phi_i}+d(y_i,\phi)\right]$$

so that

$$\frac{\partial l(\theta,\phi)}{\partial\mu_i}=\frac{\partial}{\partial\mu_i}\left\{\frac{\left(y_i\theta_i+c(\theta_i)\right)}{\phi_i}\right\}=\frac{y_i}{\phi_i}\frac{\partial\theta_i}{\partial\mu_i}+\frac{1}{\phi_i}\frac{\partial c(\theta_i)}{\partial\theta_i}\frac{\partial\theta_i}{\partial\mu_i}=\frac{\left(y_i-\mu_i\right)}{\phi_i}\frac{1}{V\left(\mu_i\right)}$$

it still follows that the usual expectation/variance results for the score function hold, that is

$$E\left[\frac{\partial l(\theta,\phi)}{\partial\mu_i}\right]=E\left[\frac{Y_i-\mu_i}{\phi_i V\left(\mu_i\right)}\right]=\frac{1}{\phi_i V\left(\mu_i\right)}\left(E\left[Y_i\right]-\mu_i\right)=0$$

and

$$\begin{aligned}
E\left[-\frac{\partial^2 l(\theta,\phi)}{\partial\mu_i^2}\right] &= E\left[-\frac{1}{\phi_i}\frac{\partial}{\partial\mu_i}\left\{\frac{Y_i-\mu_i}{V\left(\mu_i\right)}\right\}\right] \\
&= -\frac{1}{\phi_i}E\left[\left\{\frac{-V\left(\mu_i\right)-\left(Y_i-\mu_i\right)V'\left(\mu_i\right)}{\left\{V\left(\mu_i\right)\right\}^2}\right\}\right] \\
&= \left\{\frac{V\left(\mu_i\right)+\left(E[Y_i]-\mu_i\right)V'\left(\mu_i\right)}{\phi_i\left\{V\left(\mu_i\right)\right\}^2}\right\} \\
&= \frac{1}{\phi_i V\left(\mu_i\right)}=Var\left[\frac{Y_i-\mu_i}{\phi_i V\left(\mu_i\right)}\right]=Var\left[\frac{\partial l(\theta,\phi)}{\partial\mu_i}\right]
\end{aligned} \tag{5.4.8}$$

Recall that these results give the moments of the asymptotic normal distribution of ML estimates $\widehat{\beta}_{ML}$.

Thus if we **define** (rather than deduce)

$$\frac{\partial l(\theta,\phi)}{\partial\mu_i}=\frac{Y_i-\mu_i}{\phi_i V\left(\mu_i\right)}$$

then (5.4.8) **also hold given only** (5.4.7). Effectively, we define the **quasi-likelihood estimates** as the parameters β that provide the solution to the set of equations

$$\sum_{i=1}^{n}\frac{w_i\left(y_i-g^{-1}\left(x_i^T\beta\right)\right)}{\phi V\left(g^{-1}\left(x_i^T\beta\right)\right)}\frac{x_{ik}}{g'\left(g^{-1}\left(x_i^T\beta\right)\right)}=0 \qquad\qquad k=1,...,K.$$

Specifically, the **quasi-score** or **quasi-likelihood** equations are identical to (5.4.6)

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\phi_i V(\mu_i)} \frac{x_{ik}}{g'(\mu_i)} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{w_i (y_i - \mu_i)}{V(\mu_i)} \frac{x_{ik}}{g'(\mu_i)} = 0 \qquad k = 1, ..., K. \tag{5.4.9}$$

but where variance function $V(\mu_i)$ is replaced by a general variance function given by, for example,

- $Var[Y_i] = \mu_i^2$, the **constant coefficient of variation model** where

$$\frac{E[Y_i]}{\sqrt{Var[Y_i]}} = \frac{\mu_i}{\mu_i} = 1$$

- $Var[Y_i] = \phi_i \mu_i (1 - \mu_i) / n_i$ (an **overdispersed binomial**-type model)
- $Var[Y_i] = \phi_i \mu_i$ (an **overdispersed Poisson**-type model)
- $Var[Y_i] = \phi_i \mu_i^2$ (an **overdispersed Exponential**-type model).

The model can be fitted using the IRLS algorithm to yield **quasi maximum likelihood (QML)** estimates $\widehat{\beta}_{QML}$, **quasi (Pearson-type) deviance** and **dispersion** estimate

$$\chi^2\left(y, \widehat{\theta}\right) = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)/w_i} \qquad \widehat{\phi} = \frac{\chi^2\left(y, \widehat{\theta}\right)}{n - K}$$

and estimated variance covariance (and hence standard-errors) of $\widehat{\beta}_{QML}$