1 Review of the linear model
   1.1 Estimation ................................................. 5
   1.2 Statistical properties .................................. 7
   1.3 Hypothesis testing ...................................... 7
      1.3.1 Tests for linear restrictions ..................... 8
   1.4 Geometric interpretation ............................... 8
   1.5 Aliasing and multicollinearity ...................... 11
      1.5.1 Multicollinearity ................................. 13
   1.6 Interactions ............................................. 14
      1.6.1 Higher order interactions ....................... 15
      1.6.2 Interactions without main effects ............. 15
   1.7 Residuals .................................................. 16

2 Generalized Linear Models ................................. 18
   2.1 The Exponential Family of Distributions .............. 18
   2.2 Link functions ........................................... 21
   2.3 Estimation ................................................. 24
   2.4 Iteratively reweighted least-squares ................ 28
   2.5 Quadratic algorithms ................................... 31
   2.6 Asymptotic properties .................................. 35
      2.6.1 Likelihood ratio tests ............................ 36
      2.6.2 Model comparison .................................. 37
      2.6.3 Pearson $X^2$ statistic ........................... 39
   2.7 Residuals .................................................. 40

3 Models for count data ....................................... 43
   3.1 Poisson regression and log-linear models ............ 43
Section 1
Review of the linear model

The objective of this course is to understand and extend the theory and methodology of linear (regression) modelling to more general settings.

In regression modelling, we attempt to devise a model for the response data formed by combining information from predictor data.

Notation
For an independent sample, indexed, \(i = 1, \ldots, n\) denote by \(Y_i, i = 1, \ldots, n\) the response random variable (or dependent variable or outcome) and by \(X_i, i = 1, \ldots, n\) the predictor random variables (or independent variables or covariates). \(X_i = (X_{i1}, \ldots, X_{ip})\) is in general a \((1 \times p)\) row vector.

Note
1. In the regression model, the \(X_i\)'s will be treated as fixed constants. Further, the \(X_{ij}, j = 1, \ldots, p\) need not be derived from separate variables. \(i.e.\) can include \(X_{i1}, X_{i2}, \ldots\). We could also have a discrete predictor \(e.g.\) \(X \in \{1, 2, \ldots, L\}\), then
   \[
   X_{ij} = \begin{cases} 
   1 & \text{if } X_i = j \\
   0 & \text{if } X_i \neq j 
   \end{cases} = 1_j(X)
   \]
   for \(j = 1, \ldots, L\). Discrete predictors are termed factors.
2. \(X_{ij}\) could be a transformed of another variable, for example \(X_{ij} = \log X_i\) or \(X_{ij} = \cos(2\pi X_i)\) (harmonic regression), etc.
3. \(X_{ij}\) could be a transform of more than one variable, \(e.g.\) \(X_{ij}X_{i1}^{(1)}X_{i2}^{(2)}\), etc.

We will denote observed data by lowercase variables; \(y_i, x_i = (x_{i1}, \ldots, x_{ip})\) a \((1 \times p)\) for the same individual, and by \(x_j = (x_{ij}, \ldots, x_{nj})^T\) the \((n \times 1)\) vector for \(j\) across individuals.

The (general) linear model is defined by considering the conditional expectation of \(Y_i\), given \(X_i = x_i\). We specify that
   \[
   \mathbb{E}(Y_i | X_i = x_i) = x_i\beta
   \]
   where \(\beta = (\beta_1, \ldots, \beta_p)^T\) is a \((p \times 1)\) column vector of parameters. Considering all data simultaneously we have
   \[
   \mathbb{E}(Y | X = x) = X\beta
   \]

\(^1\)We will treat \(X_i\) as fixed even if we consider the tuple \((Y_i, X_i)\) as arising from a stochastic process. We will model the stochastic variation in \(Y_i\) conditional on the variation of known \(X_i\). We don’t have to worry about the probability distribution of the \(X_i\)'s and we can manipulate them however we like numerically.
respectively \((n \times 1)\), where \(X\) is the \((n \times p)\) design matrix with \((i, j)\)th element \(x_{ij}\) and where \(\beta\) is \((p \times 1)\).

A probabilistic (or statistical) model is formed by considering the presence of residual error random variables. Specifically, we write that

\[
Y = X\beta + \epsilon
\]

where \(\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top\) is a \((n \times 1)\) vector of random variables with

\[
\mathbb{E}(\epsilon_i) = 0, \quad i = 1, \ldots, n \quad \text{Var}(\epsilon) = \mathbb{E}(\epsilon\epsilon^\top) = \Sigma
\]

where \(\Sigma\) is an \((n \times n)\) symmetric positive definite matrix. In the simplest case, \(\Sigma = \sigma^2 I_n\), where \(I_n\) is the \((n \times n)\) identity matrix, that is \(\Sigma_1, \ldots, \Sigma_n\) are identical in terms of moments and are uncorrelated. More commonly,

\[
\epsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)
\]

so that \(Y | X = x \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)\).

1.1 Estimation

In the linear model, we need to estimate \(\beta\) and \(\sigma^2\). \(\beta\) is estimated typically using the least-squares criterion; we choose \(\beta\) to minimize

\[
S(\beta) = (Y - X\beta)^\top \Sigma^{-1} (Y - X\beta)
\]

We minimize this quadratic form analytically by solving

\[
\frac{\partial S(\beta)}{\partial \beta} = 0
\]

a \((p \times 1)\) vector of equations (since \(S(\beta)\) is a smoothly varying function).
We have
\[ \frac{\partial S(\beta)}{\partial \beta} = -2X^\top \Sigma^{-1}Y + 2X^\top \Sigma^{-1}X\beta \]
so equating to zero, we obtain the normal equations that \( \hat{\beta} \) solves
\[ (X^\top \Sigma^{-1}X)\hat{\beta} = X^\top \Sigma^{-1}Y \]
and if \( (X^\top \Sigma^{-1}X) \) is non-singular we have
\[ \hat{\beta} = (X^\top \Sigma^{-1}X)^{-1}X^\top \Sigma^{-1}Y. \] (1.1)
If we make the simplifying assumption that \( \Sigma = \sigma^2 I_n \), then this simplifies to
\[ \hat{\beta} = (X^\top X)^{-1}X^\top Y, \]
the ordinary least square (OLS) formula.\(^5\) The estimator
\[ \hat{\beta} = (\Sigma^{-1}X)^{-1}X^\top \Sigma^{-1}Y \]
- for general \( \Sigma \) – is termed Generalized Least Squares
- if \( \Sigma \) is diagonal, say, \( \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2) \), then we often rewrite \( W = \Sigma^{-1} = \text{diag}(w_1, \ldots, w_n) \) so that
\[ \hat{\beta} = (X^\top W X)^{-1}X^\top WY \]
- termed Weighted Least Squares
- if \( \Sigma = \sigma^2 I_n \) – termed Ordinary Least Squares.
If \( \Sigma \) is parametrized via a “low” dimensional parameter vector, we may hope to estimate these parameters from the data; in the general case, we must treat \( \Sigma \) as a known quantity.
In the OLS case, no estimation of \( \sigma^2 \) is necessary to estimate \( \beta \).

Note
\[ Y = X\beta + \varepsilon. \]
If we do a Cholesky decomposition, namely if \( \Sigma = MM^\top \), then
\[ M^{-1}Y = M^{-1}X\beta + M^{-1}\varepsilon \]
or
\[ Y^* = X^*\beta + \varepsilon^* \]
\(^5\)In the GLS case, weighted least-squares or generalized least-squares, we use usually a two-step procedure, estimating the matrix \( \Sigma \) assumed to be function of few parameters, usually by some MLE equivalent method.
where $\text{Var} (\varepsilon^*) = I_n$. A generalized least-squares problem can be converted into an ordinary least-squares problem.\(^6\)

### 1.2 Statistical properties

In the linear model case, properties of $\hat{\beta}$ are easy to derive using standard results for expectation and variance. We have unbiasedness of our estimator and

$$E(\hat{\beta}) = \beta \quad \text{Var}(\hat{\beta}) = (X^\top \Sigma^{-1} X)^{-1}$$

which of course simplifies in the OLS case; if $\Sigma = \sigma^2 I_n$, then $\text{Var}(\hat{\beta}) = \sigma^2 (X^\top X)^{-1}$. These results hold in generality without any normality assumption. If $\varepsilon \sim N_n(0, \sigma^2 I_n)$, then $\hat{\beta} \sim N_p(\beta, \sigma^2 (X^\top X)^{-1})$.

Typically, $\sigma^2$ is unknown: it is usually estimated by

$$\hat{\sigma}^2 = \frac{1}{n-p} (Y - X\hat{\beta})^\top (Y - X\hat{\beta})$$

### 1.3 Hypothesis testing

In the linear setting, we can perform two types of test: test on individual parameters or for linear restrictions.

Hypothesis tests for individual $\beta$’s can be carried out under the assumption of normal residual errors. To test the null hypothesis

$$H_0 : \beta_j = b_j \quad H_1 : \beta_j \neq b_j$$

we have the usual result that

$$\frac{\hat{\beta}_j - b_j}{\sqrt{v_{jj}}} \sim T(n-p)$$

denoting the Student-$t$ distribution with $(n-p)$ degrees of freedom and where $v_{jj}$ is the $j$th diagonal element of the matrix $\hat{\sigma}^2 (X^\top X)^{-1}$. $v_{jj}$ is the estimated standard error of $\hat{\beta}_j$.\(^7\)

**Note**

Outside of the normal residual case, these results hold approximately (for finite $n$ and

\(^6\)The inversion of $(X^\top W X)$ is usually a trivial problem. We can turn the problem of the generalized least-squares into a recursive weighted least-squares. In R, there is a function for weighted least-squares, using weights provided as an argument to the `lm` function.

\(^7\)It turns out to be a natural statistic (pivotal quantity); moreover it is also the Likelihood ratio test and the Wald-test statistic for this model.
1.3.1. Tests for linear restrictions

Suppose we have the null hypothesis

\[ H_0 : A\beta = C \]

which imposes \( q \) linear constraints on \( \beta \), e.g. \( \beta_1 = \beta_2 = \beta_3 = 0 \) or \( \beta_1 = 2\beta_2 \), etc.

The test of such null hypotheses (against the general alternative) is based on an \( F \)-statistic (comparing the fit across the two models). In this case, let \( \hat{\beta}_0 \) be the parameter estimates under \( H_0 \) and \( \hat{\beta}_1 \) be the parameter estimates under \( H_0 \cup H_1 \). That is \( \hat{\beta}_0 \) is computed under the restriction imposed by \( H_0 \) and \( \hat{\beta}_1 \) is computed under no restrictions.

Let

\[
S(\hat{\beta}_0) = (Y - X\hat{\beta}_0)\top(Y - X\hat{\beta}_0) \\
S(\hat{\beta}_1) = (Y - X\hat{\beta}_1)\top(Y - X\hat{\beta}_1)
\]

and note that \( S(\hat{\beta}_0) \geq S(\hat{\beta}_1) \). Then

\[
F = \frac{[S(\hat{\beta}_0) - S(\hat{\beta}_1)]/q}{S(\hat{\beta}_1)/(n - p)}
\]

where \( S \) is the sum of squared residuals.\(^9\) If \( H_0 \) is true, then \( F \sim F(q, n - p) \) (Snedecor-Fisher distribution).\(^10\) These results are derived directly from normal distribution theory.\(^11\)

1.4 Geometric interpretation

Let \( y = (y_1, \ldots, y_n)\top \in \mathbb{R}^n \) an object in the \( n \)-dimensional data space.

The model space is defined by thinking about linear combinations of the (observed) param-

\(^8\)These exactness properties will hold in the GLM case asymptotically and approximately. We can look at the estimates and the interpretation for the \( F \)-value will be the same in the GLM as in the LM case.

\(^9\)This can be regarded as the increase in unit misfit from unrestricted to restricted model and the denominator is like a \( \Sigma \) quantity – the best possible estimate of \( \Sigma \) assuming the general model is appropriate.

\(^10\)\texttt{anova} or \texttt{aov} can be used in \texttt{R} to test for nested models. Most statistical procedures, like the \( F \)-test above, only work for nested models, similarly for goodness-of-fit criterion (such as information criterion as AIC or BIC).

\(^11\)It turns out that in the GLM case, no similar derivation exist – reliance on asymptotic result will be necessary.
eters \((X)\). We can also think of \(X\beta \in \mathbb{R}^n\), but restricted to the \(p\)-dimensional subspace

\[
\mathcal{X}_p = \left\{ X : X = \sum_{j=1}^{p} \lambda_j \mathbf{x}_j \right\},
\]

here \(\lambda_1, \ldots, \lambda_n\) are real constants and \(\mathbf{x}_1, \ldots, \mathbf{x}_p\) are the \(p\)-columns of \(X\). We can take a completely geometric viewpoint and prioritize \(\beta\) thinking about where we can get to in \(\mathcal{X}_p\). Now evidently,

\[
y = X\beta + (y - X\beta)
\]

where all observations are non-stochastic, \(y\) is the observed response, \(X\beta\) is the fitted value and \(y - X\beta\) is the residual component. The least-squares criterion is to choose \(\hat{\beta}\) as

\[
\hat{\beta} = \arg \min_{\beta} (y - X\beta)^\top (y - X\beta)
= \arg \min_{\beta} \|y - X\beta\|^2
\]

\(i.e.\) we choose \(\hat{\beta}\) to minimize the “length” of the residual vector \((y - X\beta)\). Clearly, \(\|y - X\beta\|^2\) is minimized when \(y - X\tilde{\beta}\) is perpendicular to \(\mathcal{X}_p\) and \(X\hat{\beta}\) is the projection of \(y\) onto \(\mathcal{X}_p\), the \(p\)-dimensional subspace.

Figure 1: Projection matrix

![Projection matrix](image)

therefore the vector \(y - X\hat{\beta}\) should be orthogonal to \(X\hat{\beta}\) \(i.e.\)

\[
(X\hat{\beta})(y - X\hat{\beta}) = 0 \quad \Rightarrow \quad X^\top y = (X^\top X)\hat{\beta}
\]

which means that

\[
\hat{\beta} = (X^\top X)^{-1}X^\top y
\]
assuming that $X^\top X$ is non-singular. By Pythagoras' theorem, we have

$$\|y\|^2 = \|X\hat{\beta}\|^2 + \|y - X\hat{\beta}\|^2$$

which is a familiar sum of square decomposition. That is, $\|y\|^2$ is the observed sum of squares, $\|X\hat{\beta}\|^2$ the fitted sum of squares and $\|y - X\hat{\beta}\|^2$ the residual sum of squares.

The optimal estimate of $\beta$, denoted by $\hat{\beta}$, defines the point in the ‘model space’, $X\hat{\beta}$, that is nearest to $y$. Note that also

$$X^\top (y - X\hat{\beta}) = X^\top y - (X^\top X)\hat{\beta}$$

$$= X^\top y - (X^\top X)(X^\top X)^{-1}X^\top y$$

$$= 0_p,$$

the $(p \times 1)$ vector of zeros; therefore $(y - X^\top \hat{\beta})$ is orthogonal to $\hat{y} = X\hat{\beta}$ and $(y - X^\top \hat{\beta})$ is also orthogonal to $X$ (i.e. the columns of $X$).

For random variable $Y$ and estimator $\hat{\beta}$

$$E \left( (X\hat{\beta})^\top (Y - X\hat{\beta}) \right) = 0$$

and also

$$E \left( X^\top (Y - X\hat{\beta}) \right) = 0_p$$

Finally, if $\beta_0$ is the true value of $\beta$, then

$$E \left( (X\beta_0)^\top (Y - X\beta_0) \right) = 0$$

$$E \left( X^\top (Y - X\beta_0) \right) = 0_p$$

Note

The expectation $E(\cdot)$ is done with respect to the conditional distribution of $Y$ given $X = X$ at the true value $\beta_0$.\(^{12}\)

Summary

In geometric terms, we can define $\hat{\beta}$ as the solution to the equations

$$\begin{align*}
(X\beta_0)^\top (y - X\beta_0) &= 0 \\
X^\top (y - X\beta_0) &= 0_p
\end{align*}$$

\(^{12}\)In general and unless stated otherwise, the expectations are with respect with respect to $Y|X$ under the true data generating mechanism.
respectively a $(1 \times 1)$ and $(p \times 1)$ systems, that is
\[
\sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} \beta_j (y_i - x_i \beta) = 0
\]
\[
\sum_{i=1}^{n} x_{ij} (y_i - x_i \beta) = 0,
\]
for $j = 1, \ldots, p$.

As $X \beta = E(Y|X, \beta) = \mu(X; \beta)$, we may rewrite the previous expressions
\[
(\mu(X; \beta))^\top (y - \mu(X; \beta)) = 0
\]
\[
(\hat{\mu}(X; \beta))^\top (y - \mu(X; \beta)) = 0_p
\]
where $\hat{\mu}$ is the $(n \times p)$ matrix with $(i,j)^{th}$ element
\[
\frac{\partial \mu_i(X; \beta)}{\partial \beta_j} \mu_i(X; \beta) = E(Y_i|X, \beta) = x_i \beta.
\]

Finally, note that
\[
\hat{Y} = X\hat{\beta} = X(X^\top X)^{-1}X^\top Y = HY
\]
where $H$ is the $(n \times n)$ ‘hat’ matrix that defined the projection of $Y$ onto the model space. Note that $H$ is an idempotent matrix, i.e. $H^\top H = H$.

1.5 Aliasing and multicollinearity

In the formulation of the linear model, the design matrix $X$ is presumed to be full rank, so that $X$ is a $p$-dimensional, the $p$ columns of $X$ are linearly independent, etc. If $X$ is not full-rank, this causes problems in inference.

Example 1.1 (Simple factor predictor)
If $A$ denotes levels $1, 2, \ldots, a$ so that
\[
E(Y_i|A_i = j) = \mu + \alpha_j, \quad j = 1, 2, \ldots, a
\]
–this model is often written
\[
1 + A
\]
with parameters $\mu, \alpha_1, \ldots, \alpha_a$; the design matrix $\mathbf{X} = (\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_a)$ where

$$\mathbf{x}_0 = \mathbf{1}_n^\top$$
$$\mathbf{x}_j = (x_{ij}, \ldots, x_{nj})^\top, \quad j = 1, \ldots, a$$

where

$$x_{ij} = \begin{cases} 1 & A_i = j \\ 0 & \text{otherwise} \end{cases}$$

However, $\mathbf{x}_0 = \sum_{j=1}^a \mathbf{x}_j$, the columns are linearly dependent, and a parameter is aliased.

A constraint is required, for example (1) $\mu = 0$ or (2) $\alpha_1 = 0$ (or indeed $\alpha_j = 0$, precisely one $j$).\(^{13}\) Also, (3) $\sum_{j=1}^a \alpha_j = 0$, a centering parametrization, is a valid constraint. In this course, we say “under the usual constraints”.

(1) $\widehat{\alpha}_j = \bar{y}_j = \frac{1}{n_j} \sum_{i=1}^n y_i x_{ij}$ for $j = 1, \ldots, a$ and where $n_j = \sum_{i=1}^n x_{ij}$, the count of observations that have factor level $j$.
(2) $\bar{\mu} = \bar{y}_1$, $\widehat{\alpha}_j = \bar{y}_j - \bar{y}_1$, for $j = 2, \ldots, a$.
(3) The third looks at the center tendency in the data and look at the differences from this as

$$\widehat{\mu} = \frac{1}{a} \sum_{j=1}^a \bar{y}_j \quad \widehat{\alpha}_j = \bar{y}_j - \widehat{\mu}$$

**Example 1.2 (Two factor predictors)**

Consider a model with two factor predictors, with Factor $A$ has $a$ levels and Factor $B$ has $b$ levels. We often write this as

$$\mathbb{E}(Y_i | A_i = j, B_i = k) = \mu + \alpha_j + \beta_k + \gamma_{jk} = \mu_{jk}$$

written

$$1 + A + B + A.B$$

—the main effect plus interaction models.\(^{14}\) The indices run from $j = 1, \ldots, a$ and $k = 1, \ldots, b$, therefore it appears that we have $1 + a + b + ab$ parameters. Thus some parameters are aliased as we only can identify $a \times b$ distinct $\mu$ values.

\(^{13}\)In R, in case of aliasing, the software will use a baseline category and constraint $\alpha_1 = 0$ and report NA in the report.

\(^{14}\)Where $A.B$ denotes the interaction the effect of $A$ is modified by the presence of $B$ and depending on which level we are at. A more formal and precise definition will be given shortly.
Again, constraints must be applied. For example,

\[
\mu_{jk} = \begin{cases} 
\mu & \text{if } j = k = 1 \\
\mu + \alpha_j & \text{if } j = 2, \ldots, a, k = 1 \\
\mu + \beta_k & \text{if } j = 1, k = 2, \ldots, b \\
\mu + \alpha_j + \beta_k + \gamma_{jk} & \text{if } j = 2, \ldots, a, k = 2, \ldots, b 
\end{cases}
\]

a full factorial model.\textsuperscript{15}

**Note**

These constraints are **identifiability** constraints, not constraints formulated for modelling purposes (such as \( \gamma_{jk} = 0 \), etc.)

In **incomplete** designs, non-identifiability/aliasing can occur due to data sparsity. For example, if \( A \) and \( B \) take respectively 3 levels, we could get the following contingency table with two factors:

\[
\begin{array}{ccc}
\hline 
B & 1 & 2 & 3 \\
\hline 
1 & x & x & o \\
2 & x & x & o \\
3 & o & o & x \\
\hline 
\end{array}
\]

for \( x \) denoting cells with observations and \( o \) empty cells.

**1.5.1. Multicollinearity**

Multicollinearity exists if there is an approximate linear dependence between predictors \( i.e. \) there exist \( \lambda_1, \ldots, \lambda_p \) such that

\[
\sum_{j=1}^{p} \lambda_j x_j \approx 0.
\]

If there is a perfect multicollinearity such that

\[
\sum_{j=1}^{p} \lambda_j x_j = 0,
\]

\textsuperscript{15}In the linear case, it is easy how the constraints fit with restrictions and relations with mean levels. However, in the GLM case, we will get the relation in terms of transforms of the mean. You should be comfortable with the interpretation of the parameters \( \alpha_j, \gamma_{jk} \) which you should be familiar with from previous course. R uses the previous as a default when attempting to fit models of the above form.
then $X$ is not full rank, $X^\top X$ is singular, even if there is “approximate” linear dependence, $X^\top X$ may be numerically singular, or numerically unstable on inversion.\footnote{\(\text{This can be diagnosed; in practice, one will see massive standard errors reported, or small changes in values (for e.g. new observations), leading to drastic changes in parameter estimates.}\)}

**Note**

OLS estimates are **model-specific**, i.e. the models

$$Y = X_1\beta_1 + \varepsilon$$

$$Y = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \varepsilon$$

may return different estimates for $\beta_1$. This will occur if $X_1^\top X_2 \neq 0$. This persists for the GLM case.

See picture from handout for geometric interpretation.

We could think of sequential fit – there is then no unique decomposition for the models. Thus, it is not possible to talk about the influence of a variable when more than 2 covariates are present. One could fit using regressors $X_1$, then $X_2$, or $X_2$, then fit $X_1$ on residuals on the second model, or fit the two together resulting estimate – see handout for the linear algebra results. Only when $X_1, X_2$ are orthogonal will the least square estimates remain unchanged when doing sequential fitting versus fitting joint model – thus estimates of $\beta$ are conditional on the model fitted. The same result apply in the GLM case: in a model with main effects plus interactions – one will often be ommiting a variable, and so we can’t pin down the influence of a variable, say $X_1$.

### 1.6 Interactions

An interaction between predictors $X_1$ and $X_2$ in a model indicates that the expected response is **modified** in a **different** way by changing $x_2$ at different values of $x_1$, i.e.

$$E(Y|X_1 = x_1, X_2 = x_2) - E(Y|X_1 = x_1, X_2 = x'_2)$$

$$\neq E(Y|X_1 = x'_1, X_2 = x_2) - E(Y|X_1 = x'_1, X_2 = x'_2)$$

for at least one setting of $(x_1, x'_1, x_2, x'_2)$. This holds if $X_1, X_2$ are covariates, factors or one is a factor, one a covariate.

For example, if $X_1$ is a factor with $L$ levels and $X_2$ is a covariate, then the model

$$1 + X_1 + X_2 + X_1.X_2$$
implies

\[ E(Y|X_1 = L, X_2 = x_2) = (\beta_0 + \beta_0L) + (\beta_1 + \beta_1L)x_2 \]

for \( l = 2, \ldots, L \) and for \( L = 1 \),

\[ E(Y|X_1 = 1, X_2 = x_2) = \beta_0 + \beta_1x_2 \]

i.e. \( \beta_0L \) is the modification to the intercept \( \beta_0 \) when \( X_1 = L \), \( \beta_1L \) is the modification to slope \( \beta_1 \).

1.6.1. Higher order interactions

For three predictors \( X_1, X_2, X_3 \), we may consider the three-way interaction \( X_1.X_2.X_3 \). The presence of such an interaction implies that two-way interactions of \( X_1.X_2, X_1.X_3 \) and \( X_2.X_3 \) have different effects at different levels of the omitted variable. For example, the combined effect of \( X_1 \) and \( X_2 \) evidenced via the interaction \( X_1.X_2 \) is different when \( X_3 = x_3 \) compared with when \( X_3 = x'_3 \).

1.6.2. Interactions without main effects

The common convention is to consider only models in which interactions are included only in conjunction with the corresponding main effects, i.e. the model

\[ 1 + X_1 + X_2 + X_1.X_2 \]  \hspace{1cm} (1.2)

is “legitimate”, whereas

\[ 1 + X_1 + X_1.X_2 \]  \hspace{1cm} (1.3)

is not “legitimate” as the implication of (1.3) is that \( X_2 \) is influential in the model only for certain specific values of \( X_1 \).

---

17 In R, interactions are fitted using the colon symbol :. In this course, we will always fit interactions only when already fitting the corresponding main effects, otherwise we get an insensitive model and it put constraints on the X – this is common in the LM and GLM settings.

18 While two-way interaction indicates modifications to main effect by adding the other variable; in a similar fashion, the three-way interaction will mean that there is an interaction of the omitted variable on the two-way interaction.
Example 1.3
For the two factor predictors, model (1.3) might imply

\[
E(Y|X_1 = j, X_2 = k) = \begin{cases} 
\beta_0, & \text{if } j = 1; \; k = 1, \ldots, L \\
\beta_0 + \beta_{1j}, & \text{if } j = 2, \ldots, J; \; k = 1 \\
\beta_0 + \beta_{1j} + \gamma_{jk}, & \text{if } j = 2, \ldots, J; \; k = 2, \ldots, L 
\end{cases}
\]

so that at \( j = 1 \), changing \( X_2 \) has no effect, but for \( j = 2, \ldots, J \), changing \( X_2 \) does have an effect.

Example 1.4
For covariates, consider

\[
E(Y|X = x) = \beta_0 + \beta_2 x^2. \tag{1.4}
\]

If we location-scale transform \( x \) to \( z \), i.e. set \( x = \mu + \sigma z \). \(^{19}\) The model becomes

\[
E(Y|Z = z) = \beta_0 + \beta_2 (\mu + \sigma z)^2 \\
= \beta_0 + \beta_2 (\mu^2 + 2\mu \sigma z + \sigma^2 z^2) \\
= \beta'_0 + \beta'_1 z + \beta'_2 z^2
\]

which has a linear term in it. The polynomial regression without linear term in it is after transformation involving a linear term. This means that if we want to fit model (1.4), we can’t use such transformation. \(^{20}\)

1.7 Residuals
The vector of fitted values \( \hat{y} = X\hat{\beta} \) allows us to form the vector of residuals

\[
R = Y - \hat{Y} = Y - X\hat{\beta} \\
= (I_n - X(X^TX)^{-1}X^T)Y \\
= (I_n - H)Y
\]

with observed values

\[
r = (I_n - H)y.
\]

\(^{19}\)Related to the \( \gamma_{jk} \) coefficient.

\(^{20}\)In the case of covariates, putting interactions, but omitting a main effect yields a specific restriction on the location on X-axis of the variable; i.e. we can’t location-shift the axis. For such an example, consider the transformation from Fahrenheit to Celsius. See Nelder and McCullagh on this for more extensive discussion.
As we have seen, by construction

\[ \tilde{y}^\top \mathbf{r} = \tilde{y}^\top (\mathbf{y} - X\hat{\beta}) = \tilde{y}^\top (\mathbf{Y} - \tilde{\mathbf{y}}) = 0 \]

and

\[ \tilde{x}_j^\top \mathbf{r} = \tilde{x}_j^\top (\mathbf{y} - X\hat{\beta}) = 0, \quad j = 1, \ldots, p \]

However, also, the model assumption that the \( \varepsilon \) components have constant variance and are uncorrelated can be checked by inspection of the observed residuals. Patterns in the residual plots of \( \tilde{x}_j \) versus \( \mathbf{r} \) indicate violation of the zero mean assumption made for \( \varepsilon \). \(^{21}\)

The residuals are constructed as

\[ \mathbf{R} = \mathbf{Y} - \tilde{\mathbf{Y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y} \]

which imply

\[ \text{Var} (\mathbf{R}) = (\mathbf{I}_n - \mathbf{H})\text{Var} (\mathbf{Y}) (\mathbf{I}_n - \mathbf{H})^\top = \sigma^2 (\mathbf{I}_n - \mathbf{H}); \]

elements of \( \mathbf{H} \) are terms of order \( 1/n \), (i.e. decay very quickly), therefore \( \text{Var} (\mathbf{R}) \approx \sigma^2 \mathbf{I}_n. \)

\(^{21}\) \( \mathbf{r} \) will impose heteroscedasticity and induce correlation in the residuals, so we can only capture mild violations of the assumptions.
Section 2
Generalized Linear Models

We now aim to extend the idea of “regression” modelling to other types of response data; we seek a model for $E(Y|X = x)$ but wish to relax the assumptions of additive, zero mean residual errors.

2.1 The Exponential Family of Distributions

Suppose first that the probability density (or mass function) of $Y$ given model parameters $(\theta, \phi)$ takes the form

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

for specific functions $a, b, c$. If $\phi$ is known, this model is referred to as the exponential family, whereas if $\phi$ is unknown, it is referred to as the exponential-dispersion family.$^{22}$

Note

The support of $f_Y$ cannot depend on $(\theta, \phi)$ [and $a(\phi) > 0$].

$\theta$ is the canonical parameter, $\phi$ is the dispersion parameter.

If $\ell(\theta, \phi; y) = \log f_Y(y; \theta, \phi)$, it is straightforward to verify

$$E_{f_Y}(\hat{\ell}_\theta(\theta, \phi; Y)) = 0$$

where

$$\hat{\ell}(\theta, \phi; y) = \frac{\partial}{\partial \theta} \ell(\theta, \phi; y)$$

and

$$E_{f_Y}(\hat{\ell}_{\theta\theta}(\theta, \phi; y)) + E\left( (\hat{\ell}_\theta(\theta, \phi; y))^2 \right) = 0$$

where

$$\hat{\ell}(\theta, \phi; y) = \frac{\partial^2}{\partial \theta^2} \ell(\theta, \phi; y).$$

$^{22}$We are restricting our attention with two unknown parameters models, namely the case with $\theta$ unknown and $\phi$ known (one unknown parameter distribution) and the case where $\theta, \phi$ are unknown. This covers many one-dimensional distributions of interest.
Here, $\theta$ is 1-dimensional, $\ell$ is suitably differentiable with respect to $\theta$ and all expectations are taken with respect to the “true” data generating mechanism (DGM), all expectations finite.

In the exponential family, we have

$$E(Y) = b(\theta) = \frac{\partial}{\partial \theta} (b(\theta)) = \mu, \text{ say.}$$

$$\text{Var}(Y) = a(\phi) \dot{b}(\theta) = a(\phi) \frac{\partial^2}{\partial \theta^2} b(\theta).$$

This implies that

$$\text{Var}(Y) = a(\phi) V(\mu)$$

i.e. we have an explicit mean-variance relationship.\textsuperscript{23}

**Example 2.1 (Poisson model)**

Let $Y \sim \mathcal{P}(\lambda)$, a Poisson random variable. We write

$$f_Y(y; \lambda) = \frac{\lambda^y e^{-\lambda}}{y!} = \exp \left\{ y \log(\lambda) - \lambda - \log(y!) \right\} 1_{y \geq 0, y \in \mathbb{Z}^+}$$

In this example, the canonical parameter is $\theta = \log(\lambda)$, the dispersion parameter $\phi = 1$, therefore we can rewrite the density as

$$f_Y(y; \theta) = \exp \left\{ \frac{y\theta - e^\theta}{\phi} - \log(y!) \right\}$$

therefore $b(\theta) = e^\theta$, $\dot{b}(\theta) = e^\theta$, $\ddot{b}(\theta) = e^\theta$ and $E(Y) = e^\theta = \mu$, $\text{Var}(Y) = e^\theta = \mu$ and $V(\mu) = \mu$.

**Example 2.2 (Gamma model)**

Let $X \sim \mathcal{G}(\alpha, \beta)$, a Gamma random variable with density

$$f_Y(y; \alpha, \beta) = \frac{\beta^\alpha \gamma^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)} 1_{y > 0} = \exp \left\{ (\alpha - 1) \log(y) - \beta y + \alpha \log(\beta) - \log(\Gamma(\alpha)) \right\}$$

Consider a reparametrization that explicitly involves $E(Y) = \alpha/\beta = \mu$.

\textsuperscript{23} Examples of continuous random variables that are not exponential family include uniform, Weibull, Student, discrete uniform, notably. In certain cases, we can get separation and $V(\mu)$ may be the identity.
e.g. set $\mu = \alpha/\beta, \nu = \alpha$ if and only if $\alpha = \nu, \beta = \nu/\mu$ so that

$$f_Y(y; \mu, \nu) = \frac{1}{\Gamma(\nu)} \left( \frac{\nu y}{\mu} \right)^\nu \exp \left\{ -\frac{\nu y}{\mu} \right\} \frac{1}{y} 1_{y>0}.$$ 

Set $\theta = -1/\mu, a(\phi) = 1/\nu$ which imply $b(\theta) = -\log(-\theta), \dot{b}(\theta) = -1/\theta = \mu$ and $\ddot{b}(\theta) = 1/\theta^2 = \mu^2$, thus

$$\text{Var}(Y) = \left( \frac{1}{\nu} \right) \mu^2.$$ 

so there is a quadratic relationship between the mean and the variance, $V(\mu) = \mu^2$. More details are given in the handout.

**Example 2.3 (Bernoulli model)**

Let $Y \sim B(\pi)$, a Bernoulli random variable with mass function

$$f_Y(y; \pi) = \pi^y (1 - \pi)^{1-y} 1_{y \in \{0,1\}} = \exp \left\{ y \log \left( \frac{\pi}{1 - \pi} \right) + \log(1 - \pi) \right\}.$$ 

Set

$$\theta = \log \left( \frac{\pi}{1 - \pi} \right) \ i.e. \ \pi = \frac{e^\theta}{1 + e^\theta}, \ 1 - \pi = \frac{1}{1 + e^\theta}, \ \phi = 1$$

therefore

$$b(\theta) = \log(1 + e^\theta), \quad \dot{b}(\theta) = \frac{e^\theta}{1 + e^\theta} (= \pi) = \mu \quad \text{and} \quad \ddot{b}(\theta) = \frac{e^\theta}{(1 + e^\theta)^2} (= \pi(1 - \pi))$$

implies that $V(\mu) = \mu(1 - \mu)$.

**Example 2.4 (Binomial model)**

Let $Y \sim B(m, \pi)$ a binomial random variable with $m > 0$, fixed integer with mass function

$$f_Y(y; \pi) = \binom{m}{y} \pi^y (1 - \pi)^{m-y} 1_{0 \leq y \leq m} = \exp \left\{ y \log \left( \frac{\pi}{1 - \pi} \right) + m \log(1 - \pi) + \log \left( \binom{m}{y} \right) \right\}.$$ 

In this form, $\mathbb{E}(Y) = m\pi$ and $\text{Var}(Y) = m\pi(1 - \pi)$. This form is less appealing as $\mathbb{E}(Y)$ depends on $m$. So consider the data transformation from $y \mapsto y/m$. For the “new” data, we
have
\[
f_Y(y, \pi) = \exp \left\{ m y \log \left( \frac{\pi}{1 - \pi} \right) + m \log(1 - \pi) + \log \left( \frac{m}{m y} \right) \right\} 1_{y \in \{0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{m}{m} \}}
\]
\[
= \exp \left\{ \frac{y \log \left( \frac{\pi}{1 - \pi} \right) + \log(1 - \pi)}{1/m} + \log \left( \frac{m}{m y} \right) \right\} 1_{y \in \{0, \frac{1}{m}, \frac{2}{m}, \ldots, \frac{m}{m} \}}
\]

Set
\[
\theta = \log \left( \frac{\pi}{1 - \pi} \right) \quad b(\theta) = \log(1 + e^\theta) \quad \phi = \frac{1}{m}
\]

and
\[
\mathbb{E}(Y) = \pi = \frac{e^\theta}{1 + e^\theta} = \mu
\]
\[
\text{Var}(Y) = \frac{\pi(1 - \pi)}{m} = \frac{\mu(1 - \mu)}{m} = \phi V(\mu)
\]

where \( V(\mu) \) is again equal to \( \mu(1 - \mu) \).

### 2.2 Link functions

We seek to connect the expected response to the set of predictors. Let
\[
\mu(x) = \mathbb{E}(Y | X = x)
\]
and suppose that \( \mu(x) \) depends on a finite dimensional parameter \( \beta \), in particular \( \mu(x; \beta) = \mu(\eta) \).\(^{24}\) where
\[
\eta = x\beta = \sum_{j=1}^{p} x_j \beta_j
\]
\( x \) is \( 1 \times p \) vector and \( \beta \) a \( p \times 1 \) vector. \( \eta \) is termed the **linear predictor** (linear in the parameters \( \beta_1, \ldots, \beta_p \)).

In the linear model, \( \mu(x; \beta) = x\beta \) and typically we do not constrain \( \mu \), i.e. \( \mu(x; \beta) \) takes values on the whole real line.

However, for some response data, we should seek to impose constraints. For example, if \( Y \sim \mathcal{B}(\pi) \), then \( \mathbb{E}(Y) = \pi \), but we have that \( 0 < \pi < 1 \). Thus, if we have \( Y \sim \mathcal{B}(\pi(x)) \), then we should constrain \( \mu(x) \) to the same range.

\(^{24}\)Thus depends on the inner product.
In this case, if $\beta = x\beta$ takes values on $\mathbb{R}$ (i.e. with $x, \beta$ unconstrained in general), the transformation

$$
\mu(\eta) = \frac{e^\eta}{1 + e^\eta}, \quad -\infty < \eta < \infty
$$

ensures that $0 < \mu(\eta) < 1$. In this case,

$$
\eta + \log\left(\frac{\mu}{1 - \mu}\right) = \log\left(\frac{\pi}{1 - \pi}\right).
$$

The relationship between $\mu$ and $\eta$ is specified by a link function – in the above example, the link function is given by

$$
g(t) = \log\left(\frac{t}{1 - t}\right)
$$

termed the logistic (logit) link, while the inverse of the link function

$$
g^{-1}(t) = \frac{e^t}{1 + e^t},
$$

is termed expit function.

Other possible link functions that may be used here include

1. the probit link, where $\eta = \Phi^{-1}(\mu)$, where $\Phi$ denotes the inverse standard normal CDF.$^{25}$ In this case

$$
\mu = \Phi(\eta).
$$

2. the complementary log-log, where

$$
\eta = \log(-\log(1 - \mu))
\Rightarrow \quad \mu = 1 - e^{-e^\eta}
$$

The link function is usually chosen to be a monotonic increasing function.

**Note**

Here, the link function may be based on any cumulative distribution function.$^{26}$

The choice of a link function is a modelling choice.$^{27}$

---

$^{25}$Maps the probabilities back to the quantiles scale.

$^{26}$For example, in the above we had the logistic distribution, the normal distribution and the Gumbel distribution CDFs. These are the ones used.

$^{27}$Standard comparison methods for model selection won’t generally be appropriate, residuals could be looked at, but it is relatively tricky. More often is a sensitivity analysis used.
In some situations, an obvious link function is implied.

Example 2.5 (Binomial (rescaled))
We rescale the data so that \( Y \in \{0, 1/m, \ldots, 1\} \). Recall that

\[
f_Y(y, \pi) = \exp \left\{ y \log \left( \frac{\pi}{1-\pi} \right) + \log(1-\pi) \right\} + \log \left( \frac{m}{my} \right)
\]

so that the canonical parameter is

\[
\theta = \log \left( \frac{\pi}{1-\pi} \right).
\]

But \( \pi \equiv \mu \) in this case, so the logit link is perhaps the most natural choice.

If the link function is chosen such that \( \theta = \eta \), the link function is termed the canonical link function.

It is easy to derive. In the Poisson case, the canonical link is

\[
g(t) = \log t,
\]
while for the Gamma, the canonical link is the reciprocal

\[
g(t) = \frac{1}{t}
\]
and the right hand side is positive.\(^{28}\)

Note
If \( \eta = \theta \) using the canonical link,

\[
f_Y(y_i, \theta_i, \phi) = \exp \left( \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(\phi, y_i) \right) = \exp \left( \frac{y_i x_i \beta - b(x_i \beta)}{a(\phi)} + c(\phi, y_i) \right)
\]

For data \( y_1, \ldots, y_n \) (independent, but not identically distributed), the likelihood is

\[
L_n(\beta, \phi, y, x) = \exp \left( \frac{\sum_{i=1}^n y_i x_i \beta - \sum_{i=1}^n b(x_i \beta)}{a(\phi)} + \sum_{i=1}^n c(y_i, \phi) \right)
\]

\(^{28}\) The canonical link may not be suitable, since \( g(t) \) may not be positive; it may thus not be a sensible and a link function as the log may be more appropriate.
therefore $\sum_{i=1}^{n} x_{ij} Y_i$ is a sufficient statistic for $\beta_j$, for $j = 1, \ldots, p$.

To recapitulate, for the exponential-dispersion family,

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

where

$$\mathbb{E}(Y) = \hat{b}(\theta) = \mu$$

$$\text{Var}(Y) = a(\phi) \hat{b}(\theta) = a(\phi) V(\mu)$$

with linear predictor

$$\eta_i = x_i \beta = \sum_{j=1}^{p} x_{ij} \beta_j.$$

We model $\mu = \mathbb{E}(Y_i | X_i = x_i)$ by setting $\mu = g^{-1}(\eta)$.

**Example 2.6 (Failure of “O”-rings, Challenger catastrophe)**

See handout.

### 2.3 Estimation

Statistical inference for GLMs is usually based on the maximum likelihood principle.

For independent data $y_1, \ldots, y_n$, the likelihood is given by

$$L(\theta, y) = \prod_{i=1}^{n} f_Y(y_i; \theta) = L_n(\theta)$$

The log-likelihood is

$$\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log f_Y(y_i; \theta)$$

The maximum likelihood estimate, $\hat{\theta}_n$, is defined by

$$\hat{\theta}_n = \arg \max_{\theta} \ell_n(\theta)$$

and $\hat{\theta}_n$ is typically compute by differentiating $\ell_n(\theta)$ with respect to $\theta$, and equating to zero.

$$\frac{\partial \ell_n(\theta)}{\partial \theta} = 0_p$$

---

$^{29}$ Nowadays, the dimension reduction is less important as forty years ago, unless sample size are very large.
generally a \((p \times 1)\) system. Write \(\ell_n(\theta)\) for \(\partial \ell_n(\theta)/\partial \theta\); the system of equations to be solved has components

\[
\frac{\partial \ell_n(\theta)}{\partial \theta_j} = 0, \quad \text{for } j = 1, \ldots, p
\]

In general, \(\hat{\theta}_n\) is such that \(\ell_n(\hat{\theta}_n) = 0_p\) must be found numerically.\(^{30}\)

Note

By a Taylor approximation, for \(\theta\) in the parameter space \(\Theta \subseteq \mathbb{R}^p\) in a neighborhood of \(\hat{\theta}_n\), we have

\[
\ell_n(\theta) = \ell_n(\hat{\theta}_n) + (\theta - \hat{\theta}_n)^\top \hat{\ell}_n(\hat{\theta}_n) + \frac{1}{2} (\theta - \hat{\theta}_n)^\top \hat{\ell}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) + \text{remainder}
\]

i.e.

\[
\ell_n(\theta) \simeq \ell_n(\hat{\theta}_n) + (\theta - \hat{\theta}_n)^\top \hat{\ell}_n(\hat{\theta}_n) + \frac{1}{2} (\theta - \hat{\theta}_n)^\top \hat{\ell}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n)
\]

where \(\hat{\ell}_n(\theta)\) is the \((p \times p)\) matrix of second derivatives

\[
\frac{\partial^2 \ell(\theta)}{\partial \theta_j \partial \theta_k}, \quad \text{for } j, k = 1, \ldots, p.
\]

But \(\hat{\ell}_n(\hat{\theta}_n) = 0\), so on rearrangement we have that

\[
\ell_n(\hat{\theta}_n) - \ell_n(\theta) \simeq -\frac{1}{2} (\theta - \hat{\theta}_n)^\top \hat{\ell}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n)
\]

and thus near the mode of \(\ell(\theta)\), the function behaves like a quadratic function in \(\theta\). Thus

\[
\exp \left( \ell_n(\hat{\theta}_n) - \ell_n(\theta) \right) = \frac{\mathcal{L}_n(\hat{\theta}_n)}{\mathcal{L}_n(\theta)} = \exp \left( \frac{1}{2} (\theta - \hat{\theta}_n)^\top \hat{\ell}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) \right)
\]

In a GLM, \(\theta\) is replaced by \(\beta = (\beta_1, \ldots, \beta_p)^\top\).

Example 2.7 (Binomial response with factors (discrete predictors))

Suppose we have \(Y_1, \ldots, Y_p\) with \(Y_j \sim \mathcal{B}(m_j, \pi_j)\) for \(j = 1, \ldots, p\). Then by analytical calculations, we have

\[
\hat{\pi}_j = \frac{y_j}{m_j}, \quad \text{for } j = 1, \ldots, p
\]

We may write this as a GLM, by considering a factor predictor \(X\) taking \(p\) levels. In the

\(^{30}\)Recall that maximum likelihood estimator has lowest variance asymptotically among the class of unbiased parametric estimators. In \(\mathbb{R}\), there are many algorithms implemented to do routinely the optimization; we will see some examples of that later on in the course.
canonical parametrization, with the canonical link, we set

$$β_j = η_j = \log \left( \frac{π_j}{1 - π_j} \right)$$

so that

$$π_j = \frac{e^{β_j}}{1 + e^{β_j}} = \expit(β_j),$$

for $j = 1, \ldots, p$. By invariance of ML estimation,\footnote{\textsuperscript{31}We will see later that only in the saturated model can we solve for the maximum likelihood parameter in this way.}

$$\hat{β}_j = \log \left( \frac{π_j}{1 - π_j} \right) = \log \left( \frac{y_j}{m_j - y_j} \right)$$

We however usually work with constraints; consider the reparametrization where

$$η_j = α_0 + δ_j,$$

for $j = 1, \ldots, p$, with the identifiability constraint $δ_1 = 0$. Under the canonical link, we have

$$π_j = \frac{\exp(α_0 + δ_j)}{1 + \exp(α_0 + δ_j)}$$

for $j = 1, \ldots, p$. By invariance,

$$\hat{δ}_j = \log \left( \frac{π_j/(1 - π_j)}{π_1/(1 - π_1)} \right)$$

for $j = 2, \ldots, p$ and $\hat{α}_0 = \logit(π_1)$.

Note

$$δ_j = \log \left( \frac{π_j/(1 - π_j)}{π_1/(1 - π_1)} \right)$$

$$⇒ e^{δ_j} = \frac{π_j/(1 - π_j)}{π_1/(1 - π_1)}$$

where $π_j/(1 - π_j)$ is the odds or the success in a Bernoulli trial for factor level $j$, so that $e^{δ_j}$ is the odds ratio comparing factor level $j$ to factor level 1.
In the new parametrization, we have

\[ \mathcal{L}_n(\alpha_0, \delta_2, \ldots, \delta_p) = \prod_{i=1}^{p} \left( \frac{m_j}{y_j} \right) \left( \frac{e^{\alpha_0 + \delta_j}}{1 + e^{\alpha_0 + \delta_j}} \right)^{y_j} \left( \frac{1}{1 + e^{\alpha_0 + \delta_j}} \right)^{m_j - y_j} \]

Therefore

\[ \ell_n(\alpha_0, \delta_2, \ldots, \delta_p) = \text{const} + \sum_{j=1}^{p} \left( y_j (\alpha_0 + \delta_j) - m_j \log(1 + e^{\alpha_0 + \delta_j}) \right) \]

and so

\[ \frac{\partial \ell_n}{\partial \alpha_0} = \sum_{j=1}^{p} \left( y_j - m_j \frac{e^{\alpha_0 + \delta_j}}{1 + e^{\alpha_0 + \delta_j}} \right) \]

\[ = \sum_{j=1}^{p} (y_j - m_j \pi_j) \]

\[ = \sum_{j=1}^{p} (y_j - \mu_j) \]

we have thus

\[ \frac{\partial \ell_n}{\partial \delta_j} = y_j - m_j \expit(\alpha_0 + \delta_j) \]

\[ = y_j - m_j \pi_j \]

\[ = y_j - \mu_j \]

for \( j = 2, \ldots, p \). [Recall the linear model case] This is an example of saturated model where we had as many parameters as we had data points.

Rather than a factor predictor, we may have a continuous predictor

**Example 2.8 (Binomial response with covariate (continuous predictor))**

Suppose \( Y_1, \ldots, Y_n \) are independent with \( Y_i \sim \mathcal{B}(m_i, \pi_i) \) i.e. for data point \( i \), we have a single covariate \( x_i \), where we model

\[ \pi_i = \expit(\beta_0 + \beta_1 x_i) \]

for \( i = 1, \ldots, n \), i.e.

\[ \logit(\pi_i) = \log \left( \frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 x_i \]
and \((\beta_0, \beta_1)\) are the parameters of the model to be estimated. The log-likelihood is

\[
\ell_n(\beta_0, \beta_1) = \text{const} + \sum_{i=1}^{n} (y_i(\beta_0 + \beta_1 x_i) - m_i \log \left( 1 + e^{\beta_0 + \beta_1 x_i} \right))
\]

and we get the same estimating equations as before by partial differentiation

\[
\frac{\partial \ell_n}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - m_i \expit(\beta_0 + \beta_1 x_i))
\]

\[
= \sum_{i=1}^{n} (y_i - m_i \pi_i)
\]

\[
= \sum_{i=1}^{n} (y_i - \mu_i)
\]

and

\[
\frac{\partial \ell_n}{\partial \beta_1} = \sum_{i=1}^{n} (y_i - m_i \expit(\beta_0 + \beta_1 x_i)) x_i
\]

\[
= \sum_{i=1}^{n} x_i (y_i - m_i \pi_i)
\]

\[
= \sum_{i=1}^{n} x_i (y_i - \mu_i)
\]

These forms are again identical to what we would have for the linear model. \(^{32}\)

2.4 Iteratively reweighted least-squares

Note

For the Binomial case, \(Y_i \sim \mathcal{B}(m_i, \pi_i)\), we rescaled \(Y_i \to Y_i/m_i\) and referred to \(1/m_i\) as the “dispersion”. However, for a Binomial sample with \(m\) values \(m_1, \ldots, m_n\), this representation requires a different “dispersion” parameter for each \(i\).

Thus we instead choose to write the dispersion function

\[
a_i(\phi) = \frac{\phi}{w_i}
\]

for \(i = 1, \ldots, n\) where \(w_i\) is a known ‘weight’, \(i.e.\) previously we wrote

\[
f_{Y_i}(y_i; \pi_i) = \exp \left( \frac{y_i \log \left( \frac{\pi_i}{1 - \pi_i} \right) + \log(1 - \pi_i)}{1/m_i} + \log \left[ \left( \frac{m_i}{m_i y_i} \right) \right] \right)
\]

\(^{32}\)This is prioritizing the orthogonality of the residuals with the covariates, generalized in the GLM setting.
which we now rewrite as

\[ f_{Y_i}(y_i; \pi_i) = \exp \left( \frac{w_i}{\phi} \left[ y_i \log \left( \frac{\pi_i}{1-\pi_i} \right) + \log(1 - \pi_i) \right] + \log \left( \frac{m_i}{m_i y_i} \right) \right) \]

i.e. \( \phi = 1 \) and \( w_i = m_i \).

For the GLM model based on an exponential-dispersion family model parametrized by coefficients \( \beta = (\beta_1, \ldots, \beta_p)^\top \), we have the score equations

\[
\frac{\partial}{\partial \beta_j} \ell_n(\beta, \phi) = \sum_{i=1}^{n} \frac{w_i}{\phi} \left( y_i \frac{\partial \theta_i}{\partial \beta_j} - \frac{\partial b(\theta_i)}{\partial \beta_j} \right) = 0
\]

for \( j = 1, \ldots, p \) or equivalently

\[
\sum_{i=1}^{n} \frac{w_i}{\phi} \left( y_i - \frac{\partial b(\theta_i)}{\partial \theta_i} \right) \frac{\partial \theta_i}{\partial \beta_j} = 0.
\]

We have that

\[
\mathbb{E}(Y_i) = b(\theta_i) = \mu_i
\]

\[
\text{Var}(Y_i) = a(\phi) b(\theta_i) = \frac{\phi}{w_i} V(\mu_i)
\]

Thus, the score equations can be rewritten

\[
\sum_{i=1}^{n} \frac{w_i}{\phi} \left( y_i - \mu_i \right) x_{ij} = 0
\]

assuming a link function \( g(t) \) where

\[
\eta_i = x_i \beta = g(\mu_i).
\]

Indeed, the log-likelihood for a single observation, in canonical form, is given by

\[
\ell(y_i; \theta, \phi) = \frac{y_i \theta - b(\theta)}{a(\phi)} + c(y, \phi)
\]

and we require an expression for \( \partial \ell/\partial \beta_j \). By the chain rule,

\[
\frac{\partial \ell(\beta)}{\partial \beta_j} = \frac{\partial \ell(y_i; \theta, \phi)}{\partial \theta} \frac{\partial \theta}{\partial \beta_j} + \frac{\partial \ell(y_i; \theta, \phi)}{\partial \phi} \frac{\partial \phi}{\partial \beta_j}
\]

and from \( \partial \theta/\partial \beta_j = \mu \) and \( \partial \phi/\partial \beta_j = V \), we derive \( \partial \mu/\partial \theta = V(\mu) \) and from \( \eta = \sum x_j \beta_j \), \( \partial \eta/\partial \beta_j = x_j \) which leads to the above result

\[33\text{In order to compute } \frac{\partial \theta}{\partial \beta_j}, \text{ we need to know what link function is used.} \]
Consider the random variable \( g(Y_i) \): by a (first-order) Taylor approximation, we may write
\[
g(Y_i) \approx g(\mu_i) + (Y_i - \mu_i) \dot{g}(\mu_i)
\]
Let
\[
Z_i = \eta_i + (Y_i - \mu_i) \dot{g}(\mu_i) = x_i \beta + (Y_i - \mu_i) \dot{g}(\mu_i).
\]
We have that \(^{34}\)
\[
E(Z_i) = \eta_i = x_i \beta
\]
\[
\text{Var}(Z_i) = \text{Var}(Y_i) \left( \dot{g}(\mu_i) \right)^2 = \left( \dot{g}(\mu_i) \right)^2 \frac{\phi V(\mu_i)}{w_i} \equiv v_i
\]
Thus
\[
Z_i = x_i \beta + \varepsilon_i
\]
for \( i = 1, \ldots, n \) where \( \text{Var}(\varepsilon_i) = v_i \) and we may estimate \( \beta \) using least-squares using this linear model formulation via
\[
\hat{\beta} = \left( X^\top W X \right)^{-1} X^\top W z
\]
if we treat the \( Z_i \)'s as known quantities, where
\[
W = \text{diag} \left( v_1^{-1}, \ldots, v_n^{-1} \right)
\]
a weighted-least square problem. So an iterative procedure can be constructed as follows:

Algorithm 2.1 (Iteratively reweighted least-squares)
1. Initialize: choose \( \beta^{(0)} \) (or \( \mu_i^{(0)} \)) to give \( \eta_i^{(0)} = g \left( \mu_i^{(0)} \right) \).
2. Compute
\[
\hat{Z}_i^{(1)} = \hat{\eta}_i^{(0)} + \left( y_i - \hat{\mu}_i^{(0)} \right) \dot{g} \left( \hat{\mu}_i^{(0)} \right)
\]
3. Form
\[
\hat{W}_i^{(1)} = \frac{w_i}{\left( \dot{g} \left( \hat{\mu}_i^{(0)} \right) \right)^2 V \left( \hat{\mu}_i^{(0)} \right)}
\]
\(^{34}\)Since \((Y_i - \mu_i)\dot{g}(\mu_i)\) is a zero-mean random variable whose variance can be determined separately from the mean of \( Z_i \).
and
\[ W^{(1)} = \text{diag} \left( \hat{w}_1^{(1)}, \ldots, \hat{w}_n^{(1)} \right) \]
where \( w_i \) are specified by the choice of the model.

4. Compute
\[ \hat{\beta}^{(1)} = \left( X^\top W^{(1)} X \right)^{-1} X^\top W^{(1)} \hat{Z}^{(1)} \]

5. Return to 2. with updated
\[ \hat{\eta}_i^{(1)} = x_i \hat{\beta}^{(1)}, \quad \hat{\mu}_i^{(1)}, \quad \hat{Z}_i^{(1)} \]

and proceed through 3.-5.

This procedure produces a sequence of estimates \( \hat{\beta}^{(1)}, \hat{\beta}^{(2)}, \ldots, \hat{\beta}^{(t)} \) which will (in most cases) converge to a fixed point.

6. Stopping rule: stop iterating when \( \hat{\beta}^{(t)}, \hat{\mu}^{(t)} \) and \( \hat{\eta}^{(t)} \) do not change.

This algorithm is known as **iteratively reweighted least-squares** and under regularity conditions and as \( n \to \infty \),
\[ \hat{\beta} \xrightarrow{d} N \left( \beta, I(\beta)^{-1} \right) \]
where \( \hat{\beta} \) is the fixed point of the algorithm (i.e. the value of \( \hat{\beta}^{(t)} \) at termination).  

Here
\[ I(\beta) = \frac{1}{\phi} X^\top W X, \quad [W]_{ii} = \frac{w_i}{V(\mu_i)(\hat{g}(\mu_i))^2} \]
where \( \mu_i \) is estimated by \( \hat{\mu}_i = x_i \hat{\beta} \).

At termination, we check that the score equations are solved by \( \hat{\beta} \). Some R code is provided to fit a Poisson model on page 16 of the handout.

### 2.5 Quadratic algorithms

Consider the score function \( \dot{\ell}_n(\theta) \) arising in a parametric statistical model, and a first-order approximation to \( \dot{\ell}_n(\theta) \) about some arbitrary point \( \theta = \theta_0 \), a \((p \times 1)\) vector. Thus
\[ \dot{\ell}_n(\theta) \approx \dot{\ell}_n(\theta_0) + \dot{\ell}_n(\theta_0)(\theta - \theta_0) \]

\[ \text{In which case either no solution to the score equations or poor starting values choice.} \]

\[ \text{Properties of convergence for this algorithm are hard to study since function changes with every newly collected dataset, and depends on the link function and the data.} \]
where $\ell_n(\theta_0)$ is a $(p \times p)$ matrix of mixed partials. Rearranging gives

$$\theta = \theta_0 + (\ell_n(\theta_0))^{-1} (\ell_n(\theta) - \ell_n(\theta_0))$$

**Quadratic approximation**

We have

$$\theta \simeq \theta_0 + (\ell_n(\theta_0))^{-1} (\ell_n(\theta) - \ell_n(\theta_0))$$

Replacing $\theta$ by $\hat{\theta}_n$, $\ell_n(\hat{\theta}_n) = 0_p$, a $(p \times 1)$ vector so

$$\hat{\theta}_n \simeq \theta_0 - (\ell_n(\theta_0))^{-1} \ell_n(\theta_0)$$

This suggest a recursive approach to finding $\hat{\theta}_n$

- Initiate $\theta_0$ at some value $\hat{\theta}_0$
- for $m = 1, 2 \ldots$, define

$$\hat{\theta}_m = \theta_{m-1} - (\ell_n(\theta_{m-1}))^{-1} \ell_n(\theta_{m-1})$$

- track $\{\hat{\theta}_m\}$ until convergence i.e.

$$|\hat{\theta}_m - \hat{\theta}_{m-1}| < \varepsilon_1 \quad \text{or} \quad |\ell_n(\hat{\theta}_m)| < \varepsilon_2$$

for tolerances $\varepsilon_1, \varepsilon_2$.

This is termed **Newton’s method** or **Newton-Raphson method**. As a variant on this algorithm, if feasible, we may replace $-\ell_n(\theta)$ by its expectation

$$\mathcal{I}_n(\theta) = -\mathbb{E} (\ell_n(\theta)) = \mathbb{E} (\ell_n(\theta)\ell_n(\theta)^\top)$$

where the second equality holds in the Exponential family. Recall

$$\ell_n(\theta) = \frac{\partial \ell_n(\theta)}{\partial \theta} = \sum_{i=1}^n \frac{\partial}{\partial \theta} \log f_Y(y, \theta)$$

$$\ell_n(\theta) = \frac{\partial^2 \ell_n(\theta)}{\partial \theta \partial \theta^\top} = \sum_{i=1}^n \frac{\partial^2}{\partial \theta \partial \theta^\top} \log f_Y(y, \theta)$$

respectively $(p \times 1)$ and $(p \times p)$ and where $\mathcal{I}_n(\theta)$ is the $n$-data Fisher information, where $\mathcal{I}_n(\theta) = n\mathcal{I}(\theta)$ and where $\mathcal{I}(\theta)$ is the unit Fisher information.
The recursion based on

\[ \mathbf{\hat{\theta}_n} = \mathbf{\hat{\theta}_{n-1}} + (\mathbf{I}_n(\mathbf{\hat{\theta}_{n-1}}))^{-1}\mathbf{\hat{\ell}_n(\mathbf{\hat{\theta}_{n-1}})} \]

is termed **Fisher scoring**. Often, in this algorithm, \( \mathbf{I}_n(\mathbf{\hat{\theta}_n}) \) is replaced by estimates of the unit information, the so-called observed information:

\[ \hat{\mathbf{I}}(\mathbf{\hat{\theta}_n}) = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \ell_i(\mathbf{\theta})}{\partial \mathbf{\theta} \partial \mathbf{\theta}^\top} \right\}_{\theta=\mathbf{\hat{\theta}_n}} \]

### Example 2.9 (Poisson regression)

Suppose \( Y_i \sim \mathcal{P}(\mu_i) \) and \( \mu_i = \exp(x_i \beta) \) (i.e. the canonical log link), where

\[ x_i = (x_{i1}, \ldots, x_{ip})^\top \quad (1 \times p) \]

\[ \beta = (\beta_1, \ldots, \beta_p) \quad (p \times 1) \]

We have

\[ \ell_n(\beta) = \sum_{i=1}^{n} (-\mu_i + y_i \log(\mu_i)) + \text{const} \]

\[ \Rightarrow \hat{\ell}_n(\beta) = \sum_{i=1}^{n} \left( -\frac{\partial \mu_i}{\partial \beta} + y_i \frac{\partial \log(\mu_i)}{\partial \beta} \right) \]

a \((p \times 1)\) system, where the partial derivatives are given by

\[ \frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial}{\partial \beta_j} \exp(x_i \beta) \]

\[ = x_{ij} \exp(x_i \beta) \]

\[ = x_{ij} \mu_i \]

\[ ^{37} \text{In R, \texttt{glm} uses the Fisher scoring algorithm, which is very efficient.} \]

\[ ^{38} \text{Sample averages in place of the true Fisher information, in large sample the latter is more stable, but in moderate sample size, we have no guarantee. The estimate still depends on n in the sense that the bigger the sample, the better the approximation} \]
for \( j = 1, \ldots, p \) and where \( \mathbf{x}_i \beta = \sum_{j=1}^p x_{ij} \beta_j \). Now, we also need

\[
\frac{\partial \ell_n(\beta)}{\partial \beta_j} = \sum_{i=1}^n \left( -x_{ij} \mu_i + \frac{y_i}{\mu_i} x_{ij} \mu_i \right) = \sum_{i=1}^n x_{ij} (y_i - \mu_i)
\]

for \( j = 1, \ldots, p \) and where

\[
\frac{\partial \ell_n(\beta)}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^n x_{ij} x_{ik} \mu_i
\]

for \( j = 1, \ldots, p \) and \( k = 1, \ldots, p \).

Thus Newton’s method uses the recursion for \( m = 1, 2, \ldots \)

\[
\hat{\beta}_m = \hat{\beta}_{m-1} - \left( D(\hat{\beta}_{m-1}) \right)^{-1} \ell_n(\hat{\beta}_{m-1})
\]

where \( D(\beta) \) is the \( p \times p \) matrix with \((j, k)^{th}\) element

\[
-\sum_{i=1}^n x_{ij} x_{ik} \mu_i
\]

Note that in this case

\[
\mathbb{E} \left( \frac{\partial^2 \ell_n(\beta)}{\partial \beta_j \partial \beta_k} \right) = \sum_{i=1}^n x_{ij} x_{ik} \mu_i
\]

(as the expectation is over the distributions of \( Y \)'s for fixed \( x \)'s).

**Example 2.10 (Binomial data)**

Let \( Y_1, \ldots, Y_n \), with \( Y_i \sim \text{B}(m_i, \pi_i) \) and \( \pi_i = \text{expit}(\eta_i) = e^{\eta_i}/(1 + e^{\eta_i}) \), \( \eta_i = x_i \beta \), i.e. \( \eta_i = \log \left( \frac{\pi_i}{1 - \pi_i} \right) = \text{logit}(\pi_i) \), choosing the canonical link. Thus

\[
\ell_n(\beta) = \frac{\sum_{i=1}^n y_i x_i \beta - \log(1 + \exp(x_i \beta))}{1/m_i} + \text{const}
\]

\[
\Rightarrow \frac{\partial \ell_n(\beta)}{\partial \beta_j} = \sum_{i=1}^n \frac{x_{ij} (y_i - \mu_i)}{1/m_i}
\]

However, if \( x_i \beta = \Phi^{-1}(\pi_i) \), the probit link, the likelihood derivatives are more difficult to
compute. In that case,
\[ \ell_n(\beta) = \sum_{i=1}^{n} y_i \log \left( \frac{\Phi(x_i \beta)}{1 - \Phi(x_i \beta)} \right) + \log(1 - \Phi(x_i \beta)) \frac{1}{m_i} + \text{const} \]

Now \( \frac{\partial \ell_n(\beta)}{\partial \beta_j} \) depends on both \( \Phi \) and \( \phi \), respectively the standard normal CDF and PDF. Thus a more complicated estimating equation is obtained.

2.6 Asymptotic properties

This review of asymptotic properties is explained on pages 18 to 21 in the handout; this is only a brief review. The key results for likelihood-based estimation and testing are

- **consistency**: \( \hat{\theta}_n \overset{p}{\rightarrow} \theta_0 \), for \( \theta_0 \) is the “true” value.
- **asymptotic normality**

\[
\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N_p \left( 0_p, \left( I(\theta_0) \right)^{-1} \right)
\]

**Note**
We have

\[
\frac{1}{\sqrt{n}} \hat{\ell}_n(\theta_0) = \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} \hat{\ell}_i(\theta_0) \right)
\]

\[
= \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{f}(Y_i; \theta_0)}{f(Y_i; \theta_0)} \right) \xrightarrow{d} N_p(0_p, I(\theta_0))
\]

which allows for **score test**. This result follows by the Central Limit theorem applied to the iid quantities that appear in the sum. Thus, for large \( n \), we have

\[
\hat{\theta}_n \sim N_p \left( \theta_0, \frac{1}{n} \left( \hat{I}(\hat{\theta}_n) \right)^{-1} \right)
\]

where \( \hat{I}(\hat{\theta}_n) \) is the estimated or observed Fisher information; as before, we use the estimate

\[
\hat{I}(\hat{\theta}_n) = \left\{ - \frac{1}{n} \sum_{i=1}^{n} \hat{\ell}(y_i; \hat{\theta}_n) \hat{\ell}(y_i; \hat{\theta}_n)^\top \right. \\
- \left. \frac{1}{n} \sum_{i=1}^{n} \Psi(y_i, \hat{\theta}_n) \right\}
\]
both \((p \times p)\) matrices, where
\[
\Psi(y; \bar{\theta}_n) = \ell(y; \bar{\theta}_n) = \frac{\partial^2 \ell(y; \theta)}{\partial \theta \partial \theta^\top} |_{\theta = \bar{\theta}_n}.
\]
The diagonal elements of \((\hat{I}(\bar{\theta}_n))^{-1}\) yield the estimated (squared) standard errors for the elements of \(\hat{\theta}_n\), and we may perform hypothesis tests concerning \(\theta\) using the test statistics
\[
\frac{\hat{\theta}_n - \theta}{\text{se}(\hat{\theta}_n)} \sim \mathcal{N}(0, 1) \quad \text{if } \theta_0 = 0
\]
–such a statistic is termed a \textbf{Wald statistic}. To extend to multivariate tests, we rely on the distributional result that if \(Y \sim \mathcal{N}_p(0, \Sigma)\), then
\[
Y^\top \Sigma^{-1} Y \sim \chi^2_p.
\]
–we may apply this result to (sub)vector(s) (of) \(\hat{\theta}_n\) to obtain a test statistic for simultaneous testing.\(^{39}\)

### 2.6.1. Likelihood ratio tests

To compare two competing hypotheses, we typically use the Likelihood Ratio Test (LRT).

If
\[
H_0 : \theta \in \Theta_0 \quad H_1 : \theta \in \Theta_1
\]
we use the statistic
\[
\lambda(y) = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}_n(\theta)}{\sup_{\theta \in \Theta_0 \cup \Theta_1} \mathcal{L}_n(\theta)} = \frac{\mathcal{L}_n(\hat{\theta}_n)}{\mathcal{L}_n(\hat{\theta}_n)}
\]
Clearly \(\lambda(y) \leq 1\); if \(\lambda(y) \leq k\), for some \(k < 1\), we reject \(H_0\) in favor of \(H_1\). Two cases to consider

1. \(H_0\) \textbf{completely} specifies the model parameters.
2. \(H_0\) partially specifies the model parameters.

For (1), it can be show (using a Taylor series expansion and the CLT; see handout) that
\(-2 \log(\lambda(Y)) \xrightarrow{d} Q \sim \chi^2_p\) under \(H_0\). For (2), if \(H_0\) has \(k_1\) free parameters, \(H_1\) has \(k_2\) “free” parameters, with \(k_2 > k_1\), where “free” means unspecified by the hypothesis. As \(H_0\) is a

\(^{39}\)This result is in fact the exact analog to the linear model one, leading up to Fisher-\(F\) test
restricted version of $H_1$, then $-2 \log(\lambda(Y)) \overset{d}{\to} Q \sim \chi^2_{x_2 - k_1}$.\(^{40}\)

The likelihood ratio test structure is built to compare **nested models**. Nesting for GLMs corresponds to restrictions on the terms in the linear predictor e.g. \(\eta_i = \beta_0 + \beta_1 x_i\) whereas \(H_1 : \eta_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2\) for \(i = 1, \ldots, n\) (as in the linear model).

### 2.6.2. Model comparison

Different models for \(\mu_i = \mathbb{E}(Y_i | X_i = x_i)\) yield different log-likelihood values. The simplest model is where \(\hat{\mu}_i = \hat{\mu}_0\) (the **null model**) (independence on \(x\)). In contrast to that, the most complex model where \(\hat{\mu}_i = \hat{y}_i\), (\(\hat{\mu}_i\) is the “fitted value”). This one is a fit which matches the data exactly. For other models, we have

\[
\hat{\mu} = \mu(x_i \hat{\beta}) = g^{-1}(x_i \hat{\beta}),
\]

where \(g^{-1}\) is the inverse link function.

Let

\[
\bar{\theta} = \theta(\hat{\mu}_0 1_n) \quad \hat{\theta} = \theta(\hat{\mu}) \quad \tilde{\theta} = \theta(y)
\]

be \((n \times 1)\) vectors of canonical parameters derived under the three scenarios (null model, weighted and fitted values). For the (weighted) Exponential-Dispersion family

\[
\ell_n(\theta) = \sum_{i=1}^n \left[ \frac{w_i (y_i \theta_i - b(\theta_i))}{\phi} + c(y_i, \phi) \right]
\]

where \(\theta = (\theta_1, \ldots, \theta_n)\). Therefore, we can compute \(2(\ell_n(\tilde{\theta}) - \ell_n(\bar{\theta}))\). We have that

\[
2(\ell_n(\tilde{\theta}) - \ell_n(\bar{\theta})) = 2 \sum_{i=1}^n \frac{w_i}{\phi} \left[ y_i (\tilde{\theta}_i - \bar{\theta}_i) - (b(\tilde{\theta}_i) - b(\bar{\theta}_i)) \right]
\]

\[
= \frac{D(y; \hat{\mu})}{\phi}
\]

where \(D(y; \hat{\mu})\) is termed the **deviance** of the fitted model that defines \(\hat{\mu}\). The quantity \(D^*(y; \hat{\mu}) = D(y; \hat{\mu})/\phi\) is termed the **scaled deviance**.\(^{37}\)

\(^{40}\)The likelihood ratio test belong to Neyman-Pearson theory, it is explicit that the model appearing in the numerator is nested in the model appearing in the model in the denominator; the numerator is a restricted version and we have an explicit nesting of the models. The ANOVA \(F\)-test is a special case of the LRT; the theory which enables this result doesn’t hold for non-nested models. One needs to Taylor-expand one of the model to get to the second; there is no guarantee that you can do that if the models are not nested.
Example 2.11
In the Poisson model,

$$\ell_n(\theta) = \sum_{i=1}^{n} (-\mu_i + y_i \log(\mu_i)) + \text{const}$$

In the full model, $\hat{\mu}_i \equiv y_i$. We can define the scaled deviance quantity as

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

$$= D(y, \hat{\mu})$$

In the binomial model, $Y_i \sim \text{B}(m_i, \pi_i)$, where as usual we look at the proportion scale $Y_i/m_i$. In this case,

$$\ell_n(\theta) = \sum_{i=1}^{n} \left( y_i \log(\mu_i) + \frac{(1 - y_i) \log(1 - \mu_i)}{1/m_i} + \text{const} \right)$$

The full model has $\hat{\mu}_i \equiv y_i$, so we again have a simple deviance formula,

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

The deviance measures the discrepancy in fit between full and fitted models. According to likelihood ratio (LR) theory, under regularity condition (and asymptotically)

$$2 \left( \ell_n(\theta) - \ell_n(\hat{\theta}) \right) = -2 \log(\lambda(y)) \sim \chi^2_{n-p}$$

if the model represented by $\hat{\theta}$ is an adequate model with $p$ parameters.

In expectation, if the fitted model is adequate, $D(y; \hat{\mu})/\phi \approx n - p$ which defines a heuristic model adequacy assessment (i.e. if $D(y; \hat{\mu})/\phi \approx n - p$, then the model can be considered adequate).

**Note**
If $\phi$ is unknown, a (consistent) estimator of $\phi$ is

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

**Note**
R terms the $D(y; \hat{\mu})$ quantity the “residual deviance”, $n - p$ is termed the “residual degrees
of freedom".  

Suppose $\phi$ is known, but two nested models are fitted. Recall

**Definition 2.1 (Nested model)**
A model $M_A$ is nested inside model $M_B$ by the imposition of equality constraints at the interior of the parameter space.  

Say $M_A$ has $p_A$ parameters, and $M_B$ has $p_B$ parameters, with $p_B > p_A$ (i.e. we consider fixing $p_B - p_A$ quantities to obtain $M_A$ from $M_B$)  

Then, looking at the difference in deviance between $M_A$ and $M_B$

$$
\frac{D(y; \hat{\mu}_A) - D(y; \hat{\mu}_B)}{\phi} = 2(\ell_n(\hat{\theta}_B) - \ell_n(\hat{\theta}_A)) \sim \chi^2_{p_B - p_A}
$$

if $M_A$ is an adequate simplification of $M_B$.  

In summary: for the Poisson and Binomial cases, $\phi = 1$, thus we have an assessment of

1. **model adequacy**: $D(y; \hat{\mu})/(n - p) \approx 1$?  
2. **model comparison**: $D(y; \hat{\mu}_A) - D(y; \hat{\mu}_B) \sim \chi^2_{p_B - p_A}$ provided $M_A$ is an adequate simplification of $M_B$.  

**2.6.3. Pearson $X^2$ statistic**

The statistic

$$X^2 = \sum_{i=1}^{n} \left( \frac{y_i - \hat{\mu}_i}{\sqrt{\text{Var}(\hat{\mu}_i)}} \right)^2$$

is also a measure of goodness-of-fit or model adequacy.  

Under the usual asymptotic arguments (namely Central Limit Theorem),

$$\frac{y_i - \hat{\mu}_i}{\sqrt{\text{Var}(\hat{\mu}_i)}} \sim \mathcal{N}(0, \phi)$$

under regularity conditions (see McCullagh and Nelder). Therefore, when $n$ is large,

$$\frac{X^2}{\phi} \sim \chi^2_{n - p}$$

---

41 Analogous to the linear model, and since asymptotically the deviance is $\chi^2$ distributed. There is some logic to that terminology, although it is not used by textbook. The “null deviance” corresponds to the deviance under the null model.

42 Using a point on the boundary, say setting a parameter to $\infty$, then the asymptotic theory of LR breaks down.

43 This statistic would work even for moment-based estimation and used for model adequacy assessment.
Note
To estimate $\phi$, we may use $\hat{\phi} = X^2/(n-p)$

2.7 Residuals

For the GLM setting, several types of residual quantity may be considered.

- **Pearson residual**

  $$r_{pi} = \frac{y_i - \mu_i}{\sqrt{\text{Var}(\mu_i)}}$$

  $$(= \frac{\text{obs-fitted}}{\text{se(fitted)}})$$

  The residuals should be zero mean and constant variance if the model is adequate.\(^{44}\)

Note

For a Poisson model with canonical log link, we have that $\ell_n(\beta) = \sum_{i=1}^{n} x_i^T (y_i - \mu_i)$ a $(p \times 1)$ system as before. As $\ell_n = 0_p$, we must have $\sum_{i=1}^{n} x_i^T (y_i - \hat{\mu}_i) = 0$ i.e. we have $x_j (y - \hat{\mu}) = 0$ for $j = 1, \ldots, p$ i.e. a plot of $x_j$ versus $y - \hat{\mu}$ should reveal no systematic variation in mean level.

Note

Although the Pearson residual is zero mean and constant variance (under correct specification), the distribution may be highly skewed. \(^{45}\)

- **Anscombe residuals**

  Consider a transform $y \rightarrow A(y)$ made to ensure that the distribution of $A(y)$ is approximately normal.\(^{46}\) If $Y$ follows and Exponential-Dispersion family distribution, then it can be shown that

  $$A(t) = \int_{-\infty}^{t} \frac{d\mu}{V^{1/2}(\mu)}$$

  defines the appropriate transformation.\(^{47}\)

Example 2.12

If $Y \sim \mathcal{P}(\mu), V(\mu) = \mu$. In that case, we have that

$$A(t) = \int_{0}^{t} \frac{d\mu}{\mu^{1/2}} = \frac{3}{2} t^{3/2}$$

\(^{44}\)The residuals will be odd-looking compared to the linear model, where you have a continuous response. If $y_i$ is discrete values, the appearance of the residuals plot may exhibit lines, for common values of $y_i$ or $\hat{\mu}_i$. Ignore those strange features and look at the more general aspect of the plot.

\(^{45}\)Rescaling and shifting will take care of the two first moments, but does not impose restrictions on the higher moments. The normality assumption does not hold

\(^{46}\)Akin to Box-Cox transforms, which can be applied in any regression context.

\(^{47}\)The $V$ function is the one that appears in the Exponential-dispersion family equations.
i.e. we consider transformed $Y \rightarrow \frac{3}{2}y^{\frac{2}{3}}$ and hence define the residual by

$$\frac{A(y_i) - A(\bar{\mu}_i)}{A(\bar{\mu}_i)\sqrt{V(\mu_i)}}$$

with $A(t) = \frac{3}{2}t^\frac{2}{3}$ where $\dot{A}(t) = \frac{d}{dt}A(t)$. Recall that

$$\text{Var}(Y) = V(\mu) \Rightarrow \text{Var}(A(y)) \simeq (\dot{A}(\mu))^2 V(\mu)$$

so that in the Poisson case,$^{48}$ we examine

$$r_{A_i} = \frac{3}{2} \left( \frac{y_i^\frac{2}{3} - \hat{\mu}_i^\frac{2}{3}}{\hat{\mu}_i^\frac{1}{3}} \right).$$

○ Deviance residuals: Recall that the deviance is defined via

$$2(\ell_n(\theta) - \ell_n(\bar{\theta})) = 2 \phi \sum_{i=1}^{n} w_i \left( y_i(\bar{\theta}_i - \tilde{\theta}_i) - (b(\bar{\theta}_i) - b(\tilde{\theta}_i)) \right)$$

$$= \frac{1}{\phi} \sum_{i=1}^{n} d_i$$

where

$$d_i = 2w_i \left( y_i(\bar{\theta}_i - \tilde{\theta}_i) - (b(\bar{\theta}_i) - b(\tilde{\theta}_i)) \right)$$

The deviance residual$^{49}$ is defined by

$$r_{d_i} = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

where

$$\text{sign}(t) = \begin{cases} +1 & \text{if } t > 0 \\ -1 & \text{if } t < 0. \end{cases}$$

Discussion 2.1 (R example – residual plot for Poisson model)

Look at the handouts for a toy example of residuals plots for a Poisson model, using a constant-mean response model versus a model including the continuous covariate used to generate the data (p.22). Looking at the residual plots, we see that the model underfits $X$ for low values. There are specific features that arise because of the nature of the data. Looking at the Pearson residuals, the Anscombe residuals and the deviance residuals, we

$^{48}$For the Poisson case, a more obvious transformation would be take to take logarithm, or for the square root. The transformation is solely for the final stage of residuals check. The variance estimate should be zero mean and symmetrically distributed.

$^{49}$Not to be confused with residual deviance in R which stands for deviance

41
notice they are more or less the same at the tail for the low values of \( X \). The normal approximation holds well; the low response end; the residual for the GLM data exhibit features: strong lines, since the \( Y \) are discrete values. The message should be that such structures are inevitable because of the nature of the data. We have a lot of zero counts potentially for values of \( X \) between 80 and 120, and the transform still yields zero, which is clear in the plot of Figure 4.

This completes the basic theory for GLM. We now jump into particular sets of data and particular models.
Section 3
Models for count data

Discussion 3.1
In this section, we examine Poisson response data and count data; we could look at factors
levels count to begin with. Consider the example in the handout, with data on tumor type
and location site on page 24.

With a linear model, we could explain the difference in mean with factor variable. Here, we
can think of replications data for different individuals; we will think of them as aggregate
count; we want to model the variation in expected count to see if there was some systematic
variation.

We have an investigation of death rate due to cancer, for nine age groups, we have counts of
lung cancer death, and the corresponding (approximate) population corresponding to those
combinations. Ignoring the population, we could look at the main effect plus interaction
model ignoring the variation in population size. Accounting for the population size, its not
obvious that we should look at a Poisson model. We could look at a Binomial model. We
can turn however to the approximation of the Binomial distribution by a Poisson likelihood
and fit a Poisson GLMs. We would have a GLM where $\log(\mu_i) = \log(m_i) + \log(\pi_i)$ and we
will have to include fixed $m_i$ in the model, such a term is called an offset.

In the Scottish Lip cancer, there are difference in districts due to economic activities, size
and geographic features (latitude and longitude). This comes more into a spatial modelling
of a Poisson process; the bigger the area you look at, the higher the count observed should
be. Aggregation to that level (Scotland versus region), the count should be high. We will
be looking at how to analyze such data.

We have further examples with contingency tables, either two-way or multi-way.

3.1 Poisson regression and log-linear models

Recall for $Y_i \sim P(\mu_i)$, with canonical link $\theta_i = \log(\mu_i) = x_i \beta$ and $g(t) = \log(t)$. Other
possible links are the identity link ($g(t) = t$) or the square-root link ($g(t) = \sqrt{t}$).

3.1.1. Approximations

(a) $Y \sim P(\mu)$ implies $\frac{Y - \mu}{\sqrt{\mu}} \sim N(0, 1)$ for large values of $\mu$ (i.e. large count rate) using the
Central Limit theorem.

(b) Transformation to constant variance: if $S = h(Y)$, by a Taylor approximation

$$S \approx h(\mu) + (Y - \mu) h'(\mu) + \frac{(Y - \mu)^2}{2} h''(\mu)$$
If \( h(t) = \sqrt{t} \), then
\[
S = Y^{\frac{1}{2}} \approx \mu^{\frac{1}{2}} + \frac{(Y - \mu)}{2\mu^{\frac{3}{2}}}
\]
yielding
\[
\mathbb{E}(S) \approx \mu^{\frac{1}{2}}
\]
\[
\text{Var}(S) \approx \frac{1}{4\mu}\text{Var}(Y - \mu) = \frac{1}{4}
\]
and \( \text{Var}(S) \) does not depend on \( \mu \) (for \( \mu \) large). This is called a variance stabilizing transformation. In fact, McCullagh and Nelder go in detail on p. 196 and show
\[
\mathbb{E}(Y^{\frac{1}{2}}) \approx \mu^{\frac{1}{2}} \left(1 - \frac{1}{8\mu}\right)
\]
\[
\text{Var}(Y^{\frac{1}{2}}) \approx \mu^{\frac{3}{4}} \left(1 - \frac{3}{8\mu}\right)
\]
which work up to order \( 1/\mu \).

(c) Transform to symmetry: \( h(t) = t^\frac{3}{2} \), an Anscombe transform, and
\[
\mathbb{E}\left(Y^\frac{3}{2}\right) \approx \mu^\frac{3}{2} \left(1 - \frac{1}{9\mu}\right)
\]
\[
\text{Var}\left(Y^\frac{3}{2}\right) \approx \frac{4\mu^\frac{5}{2}}{9} \left(1 + \frac{1}{6\mu}\right)
\]

3.1.2. Log-likelihood

\[
\ell_n(\mu) = \sum_{i=1}^{n} (y_i \log(\mu_i) - \mu_i) + \text{const}
\]

3.1.3. Deviance

The deviance in the Poisson case is given by
\[
D(y; \hat{\mu}) = 2 \sum_{i=1}^{n} \left( y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right)
\]

\(^{50}\)Transforming the original \( y \) and not the mean
If \( g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j \), then

\[
\frac{\partial \ell_n(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} \dot{\mu}_{i0} \left( \frac{y_i}{\mu_i} - 1 \right)
\]

where

\[\dot{\mu}_{i0} = \frac{\partial \mu_i}{\partial \beta_0} .\]

Under the canonical log link,

\[
\frac{\partial \mu_i}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \left( \exp \left( \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j \right) \right)
\]

\[= \exp \left( \beta_0 + \sum_{j=1}^{p} x_{ij} \beta_j \right) \]

\[= \mu_i \]

which implies that

\[
\frac{\partial \ell_n}{\partial \beta_0} = \sum_{i=1}^{n} (y_i - \mu_i) = 0
\]

to yield estimates. Therefore, at \( \beta = \hat{\beta} \), we have

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

and therefore the deviance reduces to

\[
2 \sum_{i=1}^{n} y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right).
\]
3.2 Implementation in R

3.2.1. Types of Poisson data

1. Factor predictors only
   - counts cross-classified by one or more factors
   - “contingency table”
   - no obvious “response” variable other than the count itself
   (a) Two-way tables (two factors) (into an $I \times J$ table).\(^{51}\)
   (b) Multiway tables (three or more factors)

2. Factor predictors and covariates
   - individual-level analysis possible

3. Offset models
   Suppose that we choose to model the expected value for the $i^{\text{th}}$ datum as $\mu_i = E_i g^{-1}(\eta_i)$ say where $E_i$ is a known constant. $E_i$ allows us to consider systematic variations in $E(Y_i)$ which is not dependent on the predictors.
   In the Poisson case, with log link, we have
   
   $\log(\mu_i) = \log(E_i) + \eta_i = \log(E_i) + x_i \beta$

   and $\log(E_i)$ is termed an offset. We can consider $\log(E_i)$ as a predictor with known coefficient equal to 1.

Example 3.1 (Cancer dataset, page 24)
We have a two-way contingency table with two factors, type and site, with respectively 4 and 3 levels; one might want to test whether the type of cancer is independent of the location.

We have $\log(\mu_{jk}) = \log(\mu) + \alpha_j + \beta_k$ for $j = 1, 2, 3, 4$ and $k = 1, 2, 3$. This would be classical of a two-way ANOVA. This model is overparametrized, we thus set identifiability constraints. In R, the first level of each factor is used as the baseline group, thus $\alpha_1, \beta_1$ are zero.\(^{52}\) There are four basic models to consider:

- $M_0$: null model
- $M_1$: tumour type only main effect model
- $M_2$: site only main effect model
- $M_3$: tumour type plus site additive main effects model

---

\(^{51}\)We will be interested in symmetry and other simple structures for tables

\(^{52}\)Same as in the linear regression case.
A fifth model, with interactions, could be considered and everything for can be computed without fitting the model. The number of parameters of the model, to assess the adequacy of the model, can be determined by the same rules as for the linear model with factors.

\[ M_0 : 1 \]
\[ M_1 : 1 + (4 - 1) = 4 \]
\[ M_2 : 1 + (3 - 1) = 4 \]
\[ M_3 : 1 + (4 - 1) + (3 - 1) = 6 \]

We can perform model comparison via analysis of deviance comparisons: since all models are nested within \( M_3 \), we can compare the deviance change against a \( \chi^2 \) distribution with degrees of freedom \( p_2 - p_1 \). The easiest way to select the best model is to start with the more complicated model and try to simplify it, that is compare in this case \( M_3 \) with \( M_2 \).

The change in deviance

\[ \Delta \text{Deviance} : 196.9 - 51.795 = 145.105 \]
\[ \Delta \text{rdf} : 6 - 3 = 3 \]

To test \( H_0: M_2 \) is an adequate simplification of \( M_3 \), compare \( \Delta \) Deviance with \( \chi^2_{\Delta \text{rdf}} \) — here \( \chi^2_3(0.95) = 7.815 \), so we reject \( H_0 \), since the value of the test statistic is more extreme than the critical value. Similarly, \( M_1 \) is shown not to be an adequate simplification of \( M_3 \). Is \( M_3 \) an adequate model? For \( M_3 \) to be considered adequate, we would need to

\[
\frac{\text{Deviance}}{\text{dof}} = \frac{51.795}{6} = 8.6325
\]

to be near 1. We can conclude that \( M_3 \) is not adequate to explain the data.

The interaction model, given by

\text{tumor + site + tumor:site} \quad \text{or} \quad \text{tumor*site}

fits the data exactly as \( \hat{\mu}_{jk} = y_{jk} \) using MLE. Here, we refer to this model as the saturated model. The deviance for this model will be by construction zero, since the fitted model is the most complicated model in this case. There is no utility fitting that structure; we can’t compute the standard errors; we could exploit the Poisson structure to estimate the variance using the relationship between mean and variance using the estimated mean. Note that \texttt{R} also reports the AIC, which could be used for model comparison. Recall that AIC is based on the likelihood approximation, and that the lowest value of AIC indicates better fit, subject to the penalty for model complexity. Also recall that it can only be used for nested
models, we can use it for comparable likelihood (for example two models with different data transformation, for example log and identity transformation). You cannot use it to compare link functions, as you are fundamentally changing the structure of the data.

**Example 3.2 (Lung cancer incidence, page 27)**
In the lung cancer example on page 30, we can approximate the Binomial variables with low rate by a Poisson model. The counts cannot be directly compared since the population size from each case is very different. Then, if \( X \sim B(m, \pi) \) we can consider for \( X \) the Poisson approximation \( X \sim \mathcal{P}(\lambda) \) with \( \lambda \approx m\pi \). Taking logs, we get for the data

\[
\log(\mu_i) = \log(m_i) + \log(\pi_i)
\]

– we are interested in the rate per thousand of death in the population, \( \pi_i \); thus the offset \( E_i \) would be the \( m_i \) term. We have four smoking groups (smoke) and age group (age).

Taking into account the fact that we have to account the population size are different using an offset. Again consider a canonical link function. We have similar models as before. The syntax in R to use an offset is via \(~\text{offset}()\) . Since in this case \( \text{pop} \) is the population size, we need the offset to be on the log scale, so we get in this case \( \text{glm(dead~offset(log(pop)))} \).

We could perform and analysis of deviance, with \( M_3 \) beating off \( M_2 \); the simplication is not adequate; the variation seems to be explained by \( M_3 \), and the approximation is close to 1, so the model seems to capture well features of the data.

Looking at the summary, we have (intercept), whose parameter correspond to the baseline with levels 1 in both case.

**Example 3.3 (Scottish Lip Cancer data)**
This is an example of Poisson regression with covariates and factor predictors. We have an offset setting, with again depending on the population size by region. The expected number of cases takes into consideration how populous or geographically extended the region is and what would be expected taking these factors into account. We can get a sequence of model fits using \( \text{anova} \) command in the GLM. We are fitting one parameter per covariate in this case. Deviance in the table indicates the difference between the residual deviances.

Model 3 is the best among the chosen model, but is not adequate to explain the data.

### 3.2.2. GLM Analysis Checklist

1. Fit and compute deviances for a range of models exhibiting a nesting structure
2. Use “Analysis of Deviance” to compare nested models. We compare the change in deviance \( \Delta \text{Deviance}/\phi \sim \chi^2_{\text{d.r.d.f}} \), i.e. if \( \Delta \text{Deviance}/\phi \) is ‘not too large’, we do not reject the simpler model as an adequate simplification of the more complex model.
Note
If $\phi$ is unknown, it can be replaced by an estimated value $\hat{\phi}$ in the analysis of deviance.\(^{53}\)
\(R\) uses the **Pearson-based** estimate

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} w_i \left( \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \right)^2$$

– other estimates may be used, and model-based (ML) estimates of $\phi$ may also be used.

In the Analysis of Deviance, we compute $\hat{\phi}$ under the **more complex** model.

3. Test the ‘final’ model for adequacy, looking at Deviance/$\phi \sim \chi^2_{n-p}$ if the model is adequate. We use the heuristic check that

$$\frac{\text{Deviance}}{\phi} \approx n - p$$

for an adequate model.

4. Inspect the residuals

### 3.2.3. Information Criteria

We define the AIC to be $-2\ell_n(\hat{\beta}) + 2p$ and the BIC to be $-2\ell_n(\hat{\beta}) + p \log(n)$. We can compare models using AIC or BIC, choosing the lowest value to yield the “best” model.\(^{54}\)

Note
Only compare **nested models**.\(^{55}\)

**Example 3.4 (Scottish Lip Cancer model selection)**

See handout on page 31. The analysis of variance quantity can be carried out using the `anova` command. Model `glm2.off` from the Analysis of Deviance is given as the most significant fitted model, and the value of Deviance change from model `glm2.off` to model `glm3.off` gives 0.276 against $\chi^2_1$, so the later model does not yield significative improvement in fit. Increase in latitude implies increase in Lip cancer. Inspecting the residual deviance over degrees of freedoms, we have $159.71/53 \approx 3$, so the model is not adequate, even though it was best among the considered models. We can add an additional interaction effects, as we have two continuous covariates. Comparison in the model in terms of deviance yield

\(^{53}\)R does not scale the deviance, therefore in models like the Gamma or the inverse Gaussian, one needs to scale appropriately the difference in deviance.

\(^{54}\)Doesn’t take into account potential aspects of interest such as residuals and predictive capacities. Note that we would like asymptotically to select the right model provided it is within the compared models. Recall that AIC is an inconsistent model criterion. The performance in finite sample is rather unpredictable.

\(^{55}\)In general, how to compare non-nested models is unanswered. One could use predictions $p(y^*|y) \approx p(y^*|\hat{\beta})$ using hold-back sample data or future forecast for data available. Another possible option is cross-validation, taking random subsets of size 10% or 25% and check the discrepancy between fitted values and residual values.
that the interaction yield a significant improvement in deviance, thus we know include the
significantly (yet not strongly significant) negative interaction between latitude and work
in AFF. It is rather hard to interpret. Again, the heuristic comparison for model adequacy
shows that the interaction model is not adequate in explaining the variation in the data.
One could also have compared the AIC values: 373.69 vs 368.84, so the later model would
have been selected on this basis.

Adding higher order terms and quadratic terms could be used here, but no significative im-
provement would result. The assumption of the Poisson model of mean-variance equality re-
relationship may not be valid. We could use the extension to the quasi-Poisson quasi
poisson. Under this model, the parameter estimates stay the same (we are using the same link func-
tion, and as such estimates for \( \mu_i \) yields identical estimates to poisson), but the standard
ersors change and as a result the P-values (standard errors are scaled by \( \hat{\phi}^2 \)). Note the dis-
ersion parameter is 2.95, as opposed to 1 with the Poisson family. We still have \( \text{E}(Y_i) = \mu_i \)
and \( \text{Var}(Y_i) = \phi \mu_i \), but where \( \phi \) is unknown and estimated using the Pearson residuals.
The parameters estimates may no longer be significant, as the Poisson model is not valid.
Due to the small sample size, we have little power to detect the effect of AFF, and it is
questionable whether AFF significant. The AIC is not reported since the quasipoisson is
based on an estimating equation rather than a likelihood, and as such no justification for
using. Could use the pseudo-likelihood in place, and compute the objective function.

Other links could be used and could be potentially of interest, for example using either
family=poisson(link=identity) or family=poisson(link=sqrt). Neither should im-
prove on the log link.

Example 3.5 (Political Affiliations of College Students)
The data is provided in a 4 × 3 two-way table. To explain the affiliation, we have the same
form as previously, with the offset excluded, featuring constant mean, either covariate, the
additive and the interaction model (or saturated model). We could use backward model
selection, starting from the saturated model.

Comparing the additive and the full factorial model, we have 6 degrees of freedom, and the
deviance for the saturated model is of course zero by construction. Comparing the model
with the interaction yields the difference 16.39 against a \( \chi^2_6 \) distribution, which has 0.95
quantile is 12.59, so the change in deviance is significant and we reject model 1 (additive:
\( A + C \)) as an adequate simplification of model 2 (the saturated model: \( A \ast C \)). We should
not entertain any simplification from the full factorial model \( A \ast C = A + C + A.C \).

The analysis of deviance comparing one factor against the additive model yields strong
significance of both factors. Only the “saturated” model \( A \ast C \) is adequate. Using the log
link, this model sets the usual contrast parametrization

\[
\log(\mu_{jk}) \begin{cases} 
\beta_1 & j = k = 1 \\
\beta_1 + \beta_j^{(A)} & j = 2, 3, 4; k = 1 \\
\beta_1 + \beta_k^{(C)} & j = 1; k = 2, 3 \\
\beta_1 + \beta_j^{(A)} + \beta_k^{(C)} + \gamma_{jk}^{(AC)} & j = 2, 3, 4; k = 2, 3 
\end{cases}
\]

We can compute even in the case of a single data point the standard errors and translate them into statements about the $\beta$'s using invariance of the $\delta$-method. We cannot in this model compare the common influence of a factor. We cannot evaluate the adequacy of the model, and cannot address the Poisson assumption in this model. Under this model, the fitted values is precisely the data point.

We could have specialized model, with for example constant level for Independent for the four college types, while for Democrats, we could have significant changes across college. Specific values in the above parametrization could then be set to zero. The model cannot be easily fitted in R. The levels of the factors would be by assumption different. We are usually making the assumption that the levels are exchangeable, whereas the assumptions that say $\beta_3^{(C)} = 0$ says different and cannot be fitted routinely.\(^\text{56}\)

**Example 3.6 (Classroom behavior data)**

Data can be found page 30, the analysis is present on page 35 of the handout. Again, we perform backward selection approach.

We look at an Analysis of Deviance and conclude the Deviant model and Adversity factors are significant; note the sequential fitting of the `anova` function.

1. Starting with the saturated model `deviant*atrisk*adversity`, on line 19-20, we drop the three way interaction as $\Delta$ Deviance is 0.943 and the change in degrees of freedom is $(2 - 1)(2 - 1)(3 - 1) = 2$, so as the 0.95 quantiles of the chi-square distribution is $\chi^2_{2}(0.95) = 5.99$, we can drop the 3-way interaction (that is the model without 3-way interaction is an adequate simplification of the saturated model).

2. Drop 2-way interaction `atrisk:adversity`. In this case $\Delta$ Deviance=10.378 and $\Delta$rdf=$(2 - 1)(3 - 1) = 2$, but this time, we reject the hypothesis that the simpler model omitting `atrisk:adversity` is an adequate simplification.

In R, when `anova` does the comparison, it will drop the terms sequentially regardless of any considerations for the pertinence of the covariates; its a one way sequence of models, there is no including-excluding variables. On lines 95-105, we have the Analysis of Deviance for the two way interacts. We should proceed to re-introduce the `atrisk:adversity`, but

\(^{56}\text{You would need to maximize the likelihood yourself.}\)
drop other two-way interactions. For this, update could be used. See Exercise 2, Q1.\textsuperscript{57}

For the “all main effects plus all 2-way interactions” model, the deviance is 0.943 and the residual degrees of freedom are 2, so \( \hat{\phi} \approx 0.47 \). We might conclude from that that the Poisson model is adequate. There is less variation for Poisson than one would expect, this underdispersion is not particularly worrying.

3.3 Goodness-of-fit

The deviance test says that for an adequate model,

\[
\frac{D(y; \hat{\mu})}{\phi} \sim \chi^2_{n-p}
\]

What if \( \phi \) is unknown? In this case, we may replace \( \phi \) by \( \hat{\phi} \) from the Pearson estimator, \textit{i.e.}

\[
\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \left( \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \right)^2
\]

Similarly, for the model comparison analysis of deviance test,

\[
\frac{D_0 - D_1}{\phi} \sim \chi^2_{p_1-p_0}
\]

where \( D_0, p_0 \) are the deviance and degrees of freedom for the simpler model and the corresponding for \( D_1, p_1 \) for the more complicated model. We may again replace \( \phi \) by \( \hat{\phi} \) to complete the test.

This follows as the estimation of \( \hat{\phi} \) is based on a consistent procedure, and the fact that for both Pearson and Deviance-based estimators, we have (asymptotically) variation that can be ignored. We have that

\[
\frac{(D_0 - D_1)/\phi(p_1 - p_0)}{D/\hat{\phi}(n - p)} \xrightarrow{d} \mathcal{F}(p_1 - p_0, n - p)
\]

where \( \hat{\phi} \) computed from the “adequate model”, which corresponds to the \( D \) quantity in the denominator above.\textsuperscript{58} The two \( \chi^2 \) distributions divided by their degrees of freedom are asymptotically independent. The assumption for convergence is under the null that the simpler model is adequate. Thus

\[
\frac{(D_0 - D_1)/(p_1 - p_0)}{D/(n - p)} \sim \mathcal{F}(p_1 - p_0, n - p) \quad \text{as } n \to \infty
\]

\textsuperscript{57}Note that in \texttt{R}, you can use the \texttt{step} function to do backward and forward selection based on deviance, AIC or BIC, with option \texttt{direction ="both"}.  

\textsuperscript{58}This may (not necessarily) be \( M_1 \).
but this dependence on $n$ is problematic, as the asymptotic result is for large $n$.

Thus, as $\nu \to \infty$, $\chi^2_{\nu}/\nu \to 1$, the above is therefore approximately distributed asymptotically to

$$\frac{(D_0 - D_1)/(p_1 - p_0)}{D/(n - p)} \to \frac{\chi^2_{p_1 - p_0}}{p_1 - p_0}$$

as required.

Pearson $X^2$

For the Poisson model,

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

—the familiar form of the chi-squared statistic. For a two-way contingency table,

$$X^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \left( \frac{y_{jk} - \hat{\mu}_{jk}}{\sqrt{\hat{\mu}_{jk}}} \right)^2$$

Asymptotically (in the $\mu_{jk}$),

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

for a model with $p$ parameters that defines the $\hat{\mu}_{jk}$ (and the model is adequate).

Here, $n = J \times K$. For the main effects only model, $p = (J - 1) + (K - 1) + 1$, i.e. $p = J + K - 1$. We thus have $n - p = JK - (J + K - 1) = (J - 1)(K - 1)$ i.e. if the main model is adequate,

$$X^2 \sim \chi^2_{(J-1)(K-1)}$$

This is related to the chi-squared test of ‘independence’. In the main effects model under the log link, we have that

$$\log(\mu_{jk}) = \beta_{j}^{(A)} + \beta_{k}^{(B)}$$

for $j = 1, \ldots, J; k = 1, \ldots, K$ (plus the usual identifiability constraints). This means, $\mu_{jk} = \exp(\beta_{j}^{(A)}) \exp(\beta_{k}^{(B)}) = \lambda_j^{(A)} \lambda_k^{(B)}$ (i.e. the two factors modify $\mu_{jk}$ in an independent fashion).

Corresponding to the main factors plus the baseline
Note
Recall that the deviance in models with an intercept takes the form
\[ D(y, \hat{\mu}) = 2 \sum_{i=1}^{n} y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right) = G^2 \]
- this form always pertains when we deal exclusively with factor predictors.

For the change in deviance,
\[ D_0(y, \hat{\mu}^{(0)}) - D_1(y, \hat{\mu}^{(1)}) = 2 \sum_{i=1}^{n} y_i \log \left( \frac{\hat{\mu}_i^{(1)}}{\hat{\mu}_i^{(0)}} \right) \]

3.4 Contingency tables

A table of counts formed by cross-classifying subjects by two or more factors is termed a **contingency table**. We have considered
- main effects
- interactions
- saturated models

Other models of interest can be considered. For simplicity, consider three factors \(A, B, C\) with \(J, K, L\) levels respectively.

The **independence model** \(A + B + C\) is
\[ \log(\mu_{jkl}) = \beta_j^{(A)} + \beta_k^{(B)} + \beta_l^{(C)} \]

This model has \(1 + (J - 1) + (K - 1) + (L - 1)\) parameters. The next model of interest is the **partial independence** model, which we can express as \(A + B * C = A + B + C + B.C\), with three mains effects and the interactions between \(B\) and \(C\). This means
\[ \log(\mu_{jkl}) = \beta_j^{(A)} + \beta_k^{(B)} + \beta_l^{(C)} + \gamma_{kl}^{(BC)} \]
- classification by factor \(A\) is independent of cross-classification by \(B\) and \(C\), \(i.e.\ A \perp (B, C)\) and \(A\) is independent of \(B\) and \(C\). We will distinguish between the cases of independence, with \(A + B + C\) as opposed to **partial independence** \(A + B * C\) so \(A \perp (B, C)\) \(i.e.\ the effect of \(A\) on the (expected) response is independent of the joint effect of \((B, C)\) - at each level of \(A\), the joint distribution of responses at levels of \((B, C)\) is identical. We see
\[ \log(\mu_{jkl}) = \beta_j^{(A)} + \beta_k^{(B)} + \beta + t^{(C)} + \gamma_{kl}^{(BC)} \]
– for fixed \( j \) the contribution to the expected response (on the log transformed scale) from factors \((B, C)\) is always

\[
\hat{\beta}_k^{(B)} + \hat{\beta}_l^{(C)} + \hat{\gamma}_{kl}^{(BC)}
\]

whatever value of \( j \) is chosen. We also have **conditional independence** of the form \( A + B + C + A.B + B.C \equiv (A \ast B) + (B \ast C) \). In the linear predictor, we got

\[
\log(\mu_{jkl}) = \beta_j^{(A)} + \beta_k^{(B)} + \beta_l^{(C)} + \gamma_{jk}^{(AB)} + \gamma_{kl}^{(BC)}
\]

– conditional on \( B \), the factors \( A \) and \( C \) influence the expected response independently and \( A \perp C \mid B \) \(^{60}\) – in the two-way tables defined by the levels \( k = 1, \ldots, K \), the factors \( A \) and \( C \) affect the response in an independent fashion.

Graphically,

\[
\begin{align*}
A &\quad B \quad C \\
C &\quad B \quad A
\end{align*}
\]

A test for conditional independence can be based on the analysis of deviance for the two models

\[
H_0 : A + B + C + A.B + B.C \quad \text{and} \quad H_1 : A + B + C + A.B + A.C + B.C
\]

3.5 Structured log-linear models for square tables

For **two** factors \( A, B \) where \( A \) and \( B \) have the same number of levels (*i.e.* a square contingency table), we may also inspect tailored models. *E.g.* economic status, for longitudinal study, one would suspect that for multiple observations of a variable, say location, would stay in the same country or state, and otherwise the migration to another province say would happen at random.

○ **Quasi-independence:**

We have

\[
\log(\mu_{jk}) = \beta_j^{(A)} + \beta_k^{(B)} + 1_{j=k}\gamma_j
\]

– more complex than the independence model due to the extra \( J \) parameters \( \gamma_1, \ldots, \gamma_J \); this model relaxes the strict independence model, and we can compare it with the independence

\(^{60}\)Influence of the factors on the response, not the factors themselves

55
model using analysis of deviance. This model contains

\[ 1 + (J - 1) + (J - 1) + J \]

corresponding respectively the intercept, the row effects \( \beta_j^{(A)} \) for \( j = 2, \ldots, J \), the column effects \( \beta_k^{(B)} \) for \( k = 2, \ldots, J \) and the diagonal \( \gamma_1, \ldots, \gamma_J \).

- **Symmetry:**

\[
\log(\mu_{jk}) = \log(\mu_{kj}) \quad (i.e. \gamma_{jk} = \gamma_{kj})
\]

In this model, the number of parameters is given by the cells on the diagonal and the lower triangular,\n
\[ 1 + (J - 1) + \frac{J(J - 1)}{2} \]

This comes about for example in genetics for tests of association based on a symmetry for transmission of alleles.

**Note**

This model implies **marginal homogeneity** – the distribution of counts across levels of \( A \) is identical to the distribution across levels of \( B \). For e.g., if \( J = 3 \),

\[
\begin{array}{ccc}
\mu_{11} & \mu_{12} & \mu_{13} \\
\mu_{21} & \mu_{22} & \mu_{23} \\
\mu_{31} & \mu_{32} & \mu_{33}
\end{array}
\]

By the Poisson assumption, for \( j = 1, \ldots, J \)

\[
Y_{j\bullet} = Y_{j1} + Y_{j2} + Y_{j3} \sim \mathcal{P}(\mu_{j1} + \mu_{j2} + \mu_{j3})
\]

\[
\sim \mathcal{P}(\mu_{1j} + \mu_{2j} + \mu_{3j})
\]

so the distribution of \( Y_{j\bullet} \) is the same as the distribution of \( Y_{\bullet j} \).

We have

\[
\log(\mu_{jk}) = \beta_j^{(A)} + \beta_k^{(B)} + \gamma_{jk}^{(AB)}
\]

out with the constraints

\[
\beta_j^{(A)} = \beta_k^{(B)} \quad \text{if } j = k
\]

\[
\gamma_{jk}^{(AB)} = \gamma_{kj}^{(BA)} \quad \text{if } j \neq k
\]

- **Quasi-symmetry**
This model relaxes the marginal homogeneity assumption by setting

$$\log(\mu_{jk}) = \beta_j^A + \beta_k^B + \gamma_{jk}^{(AB)}$$

with only the constraints

$$\gamma_{jk}^{(AB)} = \gamma_{kj}^{(BA)} \quad \text{if} \quad j \neq k$$

are imposed.

In R, the quasi-symmetry model cannot be fitted routinely using the glm function; however, the gnm library allows for such model fitting (using the Diag and Symm functions). The Swedish election data on page 40 of the handout provides an example of such code with simple parametric models. Quasi-symmetry and quasi-independence fit are much better compared to the independence or symmetry. We cannot compare the quasi-independence against the quasi-symmetry; this is not possible since the models are not nested. A model in which people switch for parties in a similar political stripe rather than at random voting is more plausible. The models (rather than the individual parameters) are really of interest here. See also Exercise 2 for more discussion of this.

### 3.6 Iterative proportional fitting

For a two-factors models, parameter estimation under a log-linear specification can be carried out analytically. For example, the model \( A + B \), parametrized as

$$\log(\mu_{jk}) = \beta_0 + \beta_j^A + \beta_k^B$$

for \( j = 1, \ldots, J \) and \( k = 1, \ldots, K \) with constraints \( \sum_{j=1}^J \beta_j^A = \sum_{k=1}^K \beta_k^B = 0 \). Under this setting, we have that, given \( \beta_1^B, \ldots, \beta_K^B \)

\[
\hat{\beta}_0 = \frac{1}{J} \sum_{j=1}^J \log \left( \frac{y_{j\bullet}}{\omega_\bullet^{(B)}} \right) \\
\hat{\beta}_j^A = \log \left( \frac{y_{j\bullet}}{\omega_\bullet^{(B)}} \right) - \hat{\beta}_0
\]

where

$$\omega_\bullet^{(B)} = \sum_{k=1}^K e^{\beta_k^B} = \sum_{k=1}^K \omega_k^{(B)}$$

say. So given \( \beta_1^B, \ldots, \beta_K^B \) we can estimate \( \beta_0, \beta_1^A, \ldots, \beta_J^A \). Therefore, a simple iterative scheme to estimate all parameters is

\[61\] The constant, saturated or one factor can be done routinely using MLE
- fix \( \beta_1^{(B)}, \ldots, \beta_K^{(B)} \), estimate \( \beta_1^{(A)}, \ldots, \beta_J^{(A)} \) and \( \beta_0 \) as \( \hat{\beta}_1^{(B)}, \ldots, \hat{\beta}_J^{(A)}, \hat{\beta}_0 \).
- fix \( \beta_1^{(A)}, \ldots, \beta_J^{(A)} \) to the estimated values, then estimate \( \beta_1^{(B)}, \ldots, \beta_K^{(B)}, \beta_0 \).
- Iterate until convergence.

The MLE calculation for a log-linear additive model is explained in more details in the handout on iterative estimation for Poisson log-linear model. A more complicated model with three factors is given in section 3.6 in the handout on page 43. Indeed, we have in the two case that

\[
\log(\mu_{jk}) = \beta_0 + \beta_j^{(A)} + \beta_k^{(B)}
\]

and we perform ML estimation on the ‘pseudo’-data

\[
y_{jk}^{(B)} = y_{jk} \exp \left( -\beta_k^{(B)} \right)
\]

Before moving to Binomial data (which we will see can be viewed as Poisson model conditional on sums, which turn out to be binomial distributed), we have a last discussion on Poisson models.

### 3.7 Overdispersion and underdispersion

In the Poisson model, the dispersion parameter \( \phi \) is set to be 1. For modelling purposes, this is restrictive and in practice, we often observe count data where there is evidence of **overdispersion** (i.e. \( \text{Var}(Y|X) > E(Y|X) \)) or underdispersion (i.e. \( \text{Var}(Y|X) < E(Y|X) \)). We can look at the construction of a parametric model which looks at overdispersion, using the negative binomial distribution.

**Model for overdispersion**

Suppose \( Y|Z = z \sim \mathcal{P}(\mu z) \) where \( \mu > 0, Z \sim f_Z \) such that \( P(Z > 0) = 1 \) for \( E(Y) > 0 \). Using the iterated expectation and iterated variance formulas, we can easily verify that

\[
E(Y) = E_Z \left( E_{Y|Z}(Y|Z = z) \right) = \mu E_Z(Z)
\]

and

\[
\text{Var}(Y) = E_Z \left( \text{Var}_{Y|Z}(Y|Z = z) \right) + \text{Var}_Z \left( E_{Y|Z}(Y|Z = z) \right)
= \mu E_Z(Z) + \mu^2 \text{Var}_Z(Z).
\]

The most common construction has \( Z \sim \text{G}(\theta_z, \theta_z) \) i.e. \( E_Z(Z) = 1 \) and \( \text{Var}_Z(Z) = \frac{1}{\theta_z} \) for \( \theta_z \) indicating the dependence on the distribution of \( Z \). Therefore, for \( Y \), we have \( E(Y) = \mu, \text{Var}(Y) = \mu + \frac{\mu}{\theta_z} \).

The choice of the Gamma mixing distribution is partly encouraged by the fact that this makes \( Y \sim \text{NB}(\mu, \theta_z) \), the negative binomial distribution. It can be viewed as a sum of
geometric distributed random variables, or a location shift of the binomial distribution such that the support is $\text{supp}(Y) = \{0, 1, 2, \ldots\}$ where

$$f_Y(y; \mu, \theta) = \frac{\Gamma(\theta_z + y)}{\Gamma(\theta_z) y!} \left(\frac{\mu}{\mu + \theta_z}\right)^y \left(\frac{\theta_z}{\mu + \theta_z}\right)^{\theta_z} 1_{y \in \mathbb{Z}^+}$$

This parametrization is most useful as it is parametrized in terms of the mean; it is a more useful parametrization for GLM than having a probability parameter say $p = \frac{\mu}{\mu + \theta_z}$.

Note
It is clear from the construction that as $\theta_z \to \infty$, this model reverts to the Poisson model (we have asymptotically a Poisson distribution from standard distribution theory, as $Z$ become degenerate at 1).

In a GLM setting, this model can be adapted to allow dependence on covariates and factors: specifically, we would set

$$g(\mu_i) = x_i \beta$$

where $g$ is some link function, in the Poisson case for e.g. $\log(\mu_i) = x_i \beta$. This model can be estimated in a straightforward fashion and be used to model count data: it is a standard parametric model that can be estimated using ML.

In R, the function `glm.nb` in the MASS library can be used to estimate $\beta$ and $\theta_z$. For example

```r
glm.nb(formula, link = log)
```

and if $\theta_z$ was very large, we could conclude about the appropriateness, all things being equal, of the Poisson model assumption. See page 45 on the handout.\(^\text{62}\)

Note
If $\theta_z$ is estimated, comparison with the Poisson model in terms of fit (for the same linear predictor specification) is not straightforward. If $\varphi = \frac{1}{\theta_z}$, the negative binomial model reduces to the Poisson if $\varphi = 0$, therefore the Poisson model is nested inside the negative binomial, so we may compare models using the likelihood ratio test of the hypotheses

$$H_0 : \varphi = 0 \text{ against } H_1 : \varphi > 0$$

but as $\varphi = 0$ is a point on the boundary of the parameter space, the problem is non-

---

\(^\text{62}\)Direct comparison of these models is not however straightforward; one might wonder about the existence of a nesting structure here. If we think of $1/\theta_z = 0$ is setting the value of the parameter at the boundary of the parameter space will destroy the regularity of the testing, so the problem becomes non-regular.
The asymptotic distribution of the test statistic is no longer a simple $\chi^2_1$.\textsuperscript{63} It can be shown that the likelihood ratio statistic, when appropriately transformed

$$-2 \left( \ell^{(p)}_n(\hat{\beta}) - \ell^{(N_2)}_n(\hat{\beta}, \hat{\varphi}) \right) \xrightarrow{d} V$$

where $V$ has a mixture distribution.

$$P(V = 0) = \frac{1}{2} \quad \text{and} \quad P(V \leq v) = \frac{1}{2} + \frac{1}{2} \int_0^v f_Q(t) \, dt$$

where $Q \sim \chi^2_1$, namely a mixture of a point mass at zero and $\chi^2_1$ with equal probability. We have thus $P(V > v) = \frac{1}{2} \int_v^\infty f_Q(t) \, dt$, so we only need to divide the $P$ value by half of what it would be under the regular case, i.e. LRT is completed by using this non-standard asymptotic distribution to compute critical regions or $P$ values.

Underdispersion is regarded as a less important problem and is less studied. It is however possible to construct underdispersed models.

**Models for underdispersion**

The zero-inflated Poisson model has probability mass function

$$f_Y(y; \pi, \mu) = \pi \delta_{\{0\}}(y) + (1 - \pi) \frac{e^{-\mu} \mu^y}{y!} 1_{y \in \mathbb{Z}^+}$$

i.e.

$$P(Y = 0) = \pi + (1 - \pi)e^{-\mu}$$

$$P(Y = y) = (1 - \pi) \frac{e^{-\mu} \mu^y}{y!} \quad \text{for } y = 1, 2, \ldots$$

This is a simple two parameters parametric model which can be estimated in R using the library `vglm` or `VGAM`). For some settings of $(\mu, \pi)$, this model exhibits underdispersion – $\pi, \mu$ can be made dependent on predictors in the GLM setting.

---

\textsuperscript{63}As the validity of the Taylor series expansion around the parameter is not valid; for this particular kind of test, see Lawless (1987) and Chernoff (1954).
Binomial data arise as counts of “positive” responses when two possible outcomes may be observed in a sequence of independent trials. We assume

$$Y \sim B(m, \pi)$$

with $m$ fixed and known. As before, we will consider the transform $Y \mapsto Y/m$, so that

$$E(Y) = \pi, \quad \text{Var}(Y) = \frac{\pi(1 - \pi)}{m} = \phi V(\pi)$$

In a GLM setting, we consider $Y_1, \ldots, Y_n$ independent, but allow $\pi$ to be dependent on $i$ and then $\pi_i$ to vary with covariate/factor predictor values. We model

$$\eta_i = x_i \beta = g(\pi)$$

for link function $g$. Multinomial data arise when $K \geq 3$ outcomes may be observed in the original independent trials. For datum $i$, we have $Y_i = (Y_{i1}, \ldots, Y_{iK})$ – (a $K$ components vector) and write

$$Y_i \sim M(m_i, \pi_1, \ldots, \pi_K)$$

with $m_i$ a fixed, known positive integer. In this model, the joint mass function for the vector random vector $Y_i$ is

$$f_{Y_i}(y_i) \equiv f_{Y_{i1}, \ldots, Y_{iK}}(y_{i1}, \ldots, y_{iK}) = \frac{m_i!}{y_{i1}! \cdots y_{iK}!} \pi_1^{y_{i1}} \pi_2^{y_{i2}} \cdots \pi_K^{y_{iK}}$$

for $0 \leq y_{ik} \leq m_i$, $\sum_{k=1}^{K} y_{ik} = m_i$, $y_{ik}$ integer-valued where $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$. This is a $K-1$ dimensional discrete multivariate distribution determined by $K-1$ parameters as

$$y_{iK} = m_i - \sum_{k=1}^{K-1} y_{ik}$$

$$\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k.$$
We have

\[ E(Y_{ik}) = m_i \pi_k, \quad \text{for } k = 1, \ldots, K \]
\[ \text{Var}(Y_{ik}) = m_i \pi_k (1 - \pi_k), \quad \text{for } k = 1, \ldots, K \]

Indeed, marginally \( Y_{ik} \sim B(m_i, \pi_k) \). It transpires that multivariate margins e.g. \( f_{Y_1, Y_2, Y_3} \) and conditional distributions (e.g. \( f_{Y_1, Y_2|Y_3, Y_4} \), etc.) are also multinomial. In the original full multinomial model, it can be shown that

\[ \text{Cov}(Y_{ij}, Y_{ik}) = -m_i \pi_j \pi_k. \]

There is thus negative correlation between the variables. \(^{64}\)

### 4.1 Binomial Model and Logistic Regression

Suppose \( Y_i \sim B(m_i, \pi_i) \), for \( i = 1, \ldots, n \). Transform \( Y_i \) to the proportion scale: \( Y_i \mapsto Y_i/m_i \) and consider the likelihood for those new data. By previous constructions, the likelihood is

\[ f_{Y_i}(y_i, \pi_i) = \exp \left( y_i \log \left( \frac{\pi_i}{1 - \pi_i} \right) + \log(1 - \pi_i) + \log \left( \frac{m_i}{m_i y_i} \right) \right) \]

and recall that the canonical parameter \( \theta_i = \log \left( \frac{\pi_i}{1 - \pi_i} \right) \) with dispersion parameter \( \phi = 1 \) and weight \( w_i = m_i \). Here, \( \mu_i = E(Y_i) = \pi_i \) therefore the canonical link is given by \( \theta_i = \logit(\pi_i) \) so the canonical link is the **logistic** or logit link

\[ g(t) = \log \left( \frac{t}{1 - t} \right). \]

The likelihood for the corresponding GLM with linear predictor \( \eta_i = \mathbf{x}_i \beta \) yields the log likelihood

\[ \ell_n(\beta) = \sum_{i=1}^{n} \left( y_i \log \left( \frac{\pi_i}{1 - \pi_i} \right) + \log(1 - \pi_i) + \text{constant} \right) \]
\[ = \sum_{i=1}^{n} w_i \left( y_i \mathbf{x}_i \beta - \log(1 + e^{\mathbf{x}_i \beta}) \right) + \text{constant} \]

\(^{64}\)Indeed, if \( Y_1 \) is large, since the variables sum to one, this entails \( Y_2 \) has to be small.
thus

\[
\frac{\partial \ell_n(\beta)}{\partial \beta} = \sum_{i=1}^{n} m_i \left( y_i x_i^\top - \frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}} x_i^\top \right)
\]

= \sum_{i=1}^{n} m_i (y_i - \pi_i) x_i^\top

a \ (p \times 1) \ system \ usually. \ This \ requires \ solving \ numerically \ \frac{\partial \ell_n(\beta)}{\partial \beta} = 0, \ yielding \ \hat{\beta}.

Saturated model

In this setting, we set \( \hat{\pi}_i = y_i \) for \( i = 1, \ldots, n \). The deviance statistic is easy to derive as

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

Note

Suppose \( m_i = 1 \) for \( i = 1, \ldots, n \) (a Bernoulli setup); in this case, the deviance is then

\[
D(y; \hat{\pi}) = 2 \sum_{i=1}^{n} (y_i \log y_i + (1 - y_i) \log(1 - y_i) - y_i \logit(\hat{\pi}) - \log(1 - \hat{\pi}))
\]

Now \( y_i = 0 \) or \( y_i = 1 \), so

\[
y_i \log(y_i) = (1 - y_i) \log(1 - y_i) = 0
\]

Therefore

\[
D(y; \hat{\pi}) = -2 \sum_{i=1}^{n} y_i x_i^\top \hat{\beta} - 2 \sum_{i=1}^{n} \log(1 - \hat{\pi}_i)
\]

= \(-2\hat{\beta}^\top X^\top y - 2 \sum_{i=1}^{n} \log(1 - \hat{\pi}_i)\).

But

\[
\frac{\partial \ell_n(\hat{\beta})}{\partial \beta} = \sum_{i=1}^{n} (y_i - \pi_i) x_i^\top = 0_p
\]

which entails that \( X^\top y = X^\top \hat{\pi} \) where \( \hat{\pi} = (\hat{\pi}_1, \ldots, \hat{\pi}_n)^\top \). \( \footnote{This \ is \ a \ simple \ moment \ equation.} \)
the deviance term, we have
\[
D(y; \hat{\pi}) - 2\hat{\eta}^\top \hat{\pi} - 2 \sum_{i=1}^n \log(1 - \hat{\pi}_i)
\]
where \(\hat{\eta} = X\hat{\beta}\), therefore \(D\) is a function of \(\hat{\beta}\) (that is, given \(\hat{\beta}\), the quantity \(D\) is fixed). Now as \(n \to \infty\), \(\hat{\beta} \xrightarrow{p} \beta_0\), the true value, by usual consistency properties of MLE. Asymptotically, \(D\) is a function of \(\beta_0\). That is, we don’t have \(D(y; \hat{\pi}) \sim \chi^2_{n-p}\) for large \(n\) uniformly across \(\beta_0\) values, i.e. deviance assessment for the binomial model with each \(m_i = 1\), the properties of the deviance statistic are non-standard; it should not be used to assess goodness of fit of a given model.

Consider the simpler case \(Y_i \sim B(1, \pi)\), the Pearson \(X^2\) statistic
\[
X^2 = \sum_{i=1}^n \frac{(y_i - \bar{y})^2}{\bar{y}(1 - \bar{y})} = n
\]
see McCullagh and Nelder for a proof. This is fixed to the sample size, – no use as a goodness of fit statistic. Furthermore, the deviance in this model turns out to be
\[
-2n (\bar{y} \log(\bar{y}) + (1 - \bar{y}) \log(1 - \bar{y}))
\]
– asymptotics depend on the true value of \(\pi\).

Example 4.1 (Plots of uncertainty and deviance for Bernoulli with \(m = 1\))
See the plot for \(m = 1\); we have regular asymptotic, that is consistency of \(\hat{\beta} \xrightarrow{p} \beta_0\), and the uncertainty corresponding to the inverse of the Fisher information, which depends on values of \(\beta\) usually. The deviance plot however is not \(\chi^2_{n-p}\) and the distribution depends on the parameters very heavily, the deviance is too large when the deviance is small, and too small when the value of \(\beta\) is large; the asymptotic distribution must be derived for any particular value of \(\beta\). The problem persists when \(m = 3, 5\) or even \(m = 100\); it starts looking good when \(m = 1000\), the problem has dissipated.

We will see that the problem is alleviated for deviance comparison, as we can get a valid expansion.

Example 4.2
Deviance comparison for models, with \(Y_i \sim B(1, \pi_i)\), with \(M_1 : \logit(\pi_i) = \beta_0 + \beta_1 x_i\) against \(M_0 : \logit(\pi_i) = \beta_0\) i.e. to test \(H_0 : \beta_1 = 0\), we can use \(\chi^2_1\) and this is what we observe uniformly in the plots. We see a QQ-plot for 1000 replicates with a sample size of 1000.

\[66\]Whenever we did deviance analysis, we managed to get a pivotal quantity that did not depend on \(\beta_0\), thus the asymptotics derived earlier do not hold
Figure 2: Deviation from true parameter $\beta_{true}$ as a function of the parameter estimate

\[ \beta_1 - \beta_{true} \]

- $m = 1$
- $m = 3$
- $m = 5$
- $m = 10$
- $m = 100$
- $m = 1000$
Figure 3: Deviance plot as function of number of observations and value of $\beta_1$
Figure 4: Densities estimates and QQ plot for $\beta_1 - \beta_{\text{true}}$

Plotting the chi-square empirical quantiles against the $\chi^2_1$ alleged hypothetical distribution. The plot is good, with some deviations in the tails.

Thus, although deviance itself as a goodness-of-fit is not reliable, the comparison in deviance is.

Example 4.3 (Horseshoe crab data)
A single binary response with a continuous covariate. In R, for binomial data, we have to supply both $m_i, y_i$ data. Thus, in this case, you can supply only one column, since we have binary data, so $I(\text{satell}>0)$ as a factor is ok. Otherwise, we can supply `cbind(m_i-y_i, y_i)`, or $y_i/m_i$ with `weights= m_i`.

The next bit of the code reformulate the data, with a categorization according to deciles of the width. The last part of the code gives the fitted value. In model m2, since part of the explanatory power is passed from `width` to `I(width^2)`, we need to compare the changes
in deviance. The plot illustrates the categorized data proportion. The analysis of deviance shows that the latter inclusion of the quadratic term is not necessary.

A Wald test is perhaps more reliable as a one-parameter test, if

$$W_n = \frac{\hat{\beta}_2}{\text{se}(\beta_2)}$$

The comparison of the analysis of deviance can be though of as a multivariate version of the Wald test, and asymptotically equivalent. Here $W^2_n \sim \chi^2_1$, and we can assess using the test. Since we cannot use the model assessment criterion that $\text{Dev}/(n-p) \sim \chi^2_{n-p}$, as the asymptotics depend on the data generating mechanism. We could have left skew, right skew, offset to the left or the right, etc. For $m < 10$, for assessment of the adequacy of the model, the deviance criterion is completely unreliable. The asymptotics are more subject to break down than Poisson data.

**Example 4.4 (Aids and AZT)**

We are looking at the presence of absence of symptoms for a drug for HIV. We could legitimately use the Poisson likelihood for this data, with four points with two levels and two factors. This is somewhat similar to the models which were seen earlier on, with null model, either factor (azt or race), additive factor (azt+race) and a saturated model (azt*race). This is a $2 \times 2 \times 2$; although the data could be regarded as grouped data, we could convert those groups into individual data to later include covariates. We would need to supply four lines in the data frame. Here, we do the individual data analysis. For all quantities of interest with factors, the two approaches will give same results and estimates. In this case, we can use saturated models and still have residual degrees of freedom available, so we do not fit exactly the data. Looking at the effect of race from the summary, we see that the factor appears not significant. Indeed, the race term is not necessary based on the analysis of deviance. The azt estimate gives a contrast between the treated and the control groups. We are fitting the model

$$\logit(\pi_{jk}) = \log \left( \frac{\pi_{jk}}{1 - \pi_{jk}} \right) = \beta_0 + \beta_{(\text{AZT})}$$

so for control, we get $\beta_0$ and the treated groups effect is $\beta_0 + \beta_{(\text{AZT})}$ on the logit scale. Here, the Intercept is -1.0361 and $\beta_{(\text{AZT})} = -0.7218$. The odds are decreased when AZT is administered. The log-odds ratio gives

$$\log \left( \frac{\pi_{21}/(1 - \pi_{21})}{\pi_{11}/(1 - \pi_{11})} \right) = \beta_{(\text{AZT})}$$
\( \pi_{jk-j} \) indexes AZT=yes \((j = 2)\), for \( k = 1 \) and 2. Recall that an odd is 

\[
p(E) \quad \text{and the log-odds} \quad \log \left( \frac{p(E)}{1 - p(E)} \right)
\]

so the log-odds ratio for \( E, F \) is given by

\[
\log \left( \frac{p(E)/(1 - p(E))}{p(F)/(1 - p(F))} \right)
\]

This is perhaps more interesting to compare when we have azt and race; make sure you understand the indexing and interpretation.

**Example 4.5 (Birthweight data)**

We want to predict the low birthweight; the low indicator is a discretization of a covariate. The 2.5kg cutoff has a meaning full interpretation and is a quantity of interest. Including all covariates but actual birthweight, with smoking, race all having three levels, ptl indicator of premature birth. We use the drop1 command to remove a single term; in fact many could be dropped. This seems to be justified on the base of the level, dropping ftv, age and ui. In the end, we keep weight of the mother (lwt) and race, smoking, ptd and ht factors. We could also add polynomial terms to this model.

The theory we developed for GLM holds in the case of Binomial data, we could use residuals when \( n \) is large, but we are still in look for an adequacy assessment of the model. We now look at the links between the two models we have seen for count data, using the canonical links for both models.

### 4.2 Logistic regression and log-linear models

**Note**

If \( Y_1 \sim \mathcal{P}(\mu_1) \) and \( Y_2 \sim \mathcal{P}(\mu_2) \) and \( Y_1 \perp \perp Y_2 \), namely the two are independent. Then

\[
Y_1 | Y_1 + Y_2 = m \sim \mathcal{B}(m, \pi) \quad \text{where} \quad \pi = \frac{\mu_1}{\mu_1 + \mu_2}.
\]

Consider a simple log linear model for Poisson data \( y_1, y_2 \) in this setting:

\[
\log(\mu_1) = \beta_0 \\
\log(\mu_2) = \beta_0 + \beta_1
\]

Then

\[
\ell(\beta_0, \beta_1) = y_1 \beta_0 + y_2 (\beta_0 + \beta_1) - (e^{\beta_0} + e^{\beta_0+\beta_1}) + \text{const}.
\]
We now do an additional reparametrization; set \( \psi = e^{\beta_0} + e^{\beta_0 + \beta_1} = e^{\beta_0}(1 + e^{\beta_1}) \). Ignoring constant terms,

\[
\ell(\beta_1, \psi) = (y_1 + y_2) \log(\psi) - \psi + y_2 \beta_1 - (y_1 + y_2) \log(1 + e^{\beta_1}) \\
= \ell_M(\psi; m) + \ell_{Y_2|M}(\beta_1; y_2, m)
\]

where \( m = y_1 + y_2 \). \( \ell_M(\psi; m) \) is the likelihood that would arise from datum \( M = Y_1 + Y_2 \sim \mathcal{P}(\psi) \) being observed to equal \( m \) and \( \ell_{Y_2|M}(\beta_1; y_2, m) \) is the conditional likelihood for \( Y_2 \) given \( M = m \), with \( Y_2|M = m \sim \mathcal{B}(m, \pi) \) for some \( \pi = \pi(\beta_1) \).

To estimate \( \beta_1 \), we may restrict attention to \( \ell_{Y_2|M} \). In the conditional Binomial model,

\[
\beta_1 = \log \left( \frac{\mu_2}{\mu_1} \right)
\]

but from the conditional likelihood formed, we can deduce that \( \beta_1 = \text{logit}(\pi) \) which implies in turn

\[
\pi = \frac{\mu_2/\mu_1}{1 + \mu_2/\mu_1} = \frac{\mu_2}{\mu_1 + \mu_2}.
\]

This result extends to more than two random variables.

If \( Y_1, \ldots, Y_K \) with \( Y_k \sim \mathcal{P}(\mu_k) \) independent Poisson for \( k = 1, \ldots, K \). Then

\[
(Y_1, \ldots, Y_k) \bigg| \sum_{k=1}^K Y_k = m \sim \mathcal{M}(m, \pi_1, \ldots, \pi_K)
\]

where \( \pi_k = \mu_k / \sum_{l=1}^K \mu_l \) for \( k = 1, \ldots, K \).

Suppose we model

\[
\log(\mu_k) = \beta_0 + \beta_1 x_k
\]

for some continuous covariate \( x \) taking values \( x_1, \ldots, x_K \) for the \( K \) random variables. Let

\[
\psi = \sum_{k=1}^K \exp(\beta_0 + \beta_1 x_k) = \exp(\beta_0) \sum_{k=1}^K \exp(\beta_1 x_k).
\]

Reparametrize to \( \psi, \beta_1 \); we have for the log-likelihood

\[
\ell_n(\psi, \beta_1) = \ell_M(\psi; m) + \ell_{Y|M}(\beta_1; y_1, \ldots, y_K, m)
\]
where
\[ \ell_m(\psi; m) = m \log(\psi) - \psi \]
is the Poisson log-likelihood
\[ \ell_{YM}(\beta_1; y, m) = \beta_1 \sum_{k=1}^{K} x_k y_j - m \log \left( \sum_{k=1}^{K} e^{\beta_1 x_k} \right) \]
is the multinomial log-likelihood
and
\[ \pi_k = \frac{\exp(\beta_1 x_k)}{\sum_{l=1}^{K} \exp(\beta_1 x_l)} \]

Therefore all the information concerning \( \beta_1 \) lies in the conditional likelihood denoted \( \ell_{YM} \)
for \( Y_1, \ldots, Y_{K-1} \mid \sum_{k=1}^{K} Y_k = m \)
For the extension: again, \( Y_k \sim \mathcal{P}(\mu_k) \) page 57), but now \( \log(\mu_k) = \beta_0 + \beta_k \), for \( k = 1, \ldots, K \)
and \( \beta_1 \equiv 0 \) (i.e. the model with one factor predictor). Here let
\[ \psi = \sum_{k=1}^{K} \exp(\beta_0 + \beta_k) = \exp(\beta_0) \times \sum_{k=1}^{K} \exp(\beta_k) = \exp(\beta_0) \times \left( 1 + \sum_{k=2}^{K} \exp(\beta_k) \right). \]

Again, we have
\[ \ell_n(\psi, \beta_2, \ldots, \beta_K) = \ell_M(\psi; m) + \ell_{YM}(y_1, \ldots, y_K; m, \beta_2, \ldots, \beta_K) \]
with
\[ (1) : \ m \log(\psi) - \psi \]
\[ (2) : \ \sum_{k=1}^{K} y_k \beta_k - m \log \left( \sum_{k=1}^{K} \exp(\beta_k) \right) \]

Here, we have a multinomial (conditional likelihood) with
\[ \pi_k = \frac{e^{\beta_k}}{\sum_{l=1}^{K} e^{\beta_l}} = \frac{e^{\beta_k}}{1 + \sum_{l=2}^{K} e^{\beta_l}} \]
and therefore, equivalent inferences available for Poisson and Binomial/Multinomial formulations (under the chosen canonical links).
Interpreting the parameters in a logistic regression

For binary regression with a logistic link, we have

\[ \pi(x) = P(Y = 1|X = x) = \frac{\exp(x\beta)}{1 + \exp(x\beta)} = \text{expit}(x\beta) \]

that is

\[ x\beta = \log \left( \frac{\pi(x)}{1 + \pi(x)} \right) = \logit(\pi(x)) \]

the log-odds on \(Y = 1\) at \(X = x\). For continuous predictors,

\[ \begin{align*}
\logit(\pi(x)) &= \beta_0 + \beta_1 x \\
\therefore \log \left( \frac{\pi(x_1)/(1 - \pi(x_1))}{\pi(x_0)/(1 - \pi(x_0))} \right) &= \beta_1 (x_1 - x_0).
\end{align*} \]

Here \(\beta_1\) measures the change in log-odds for unit change in \(x\). For a factor predictor taking two levels,

\[
\begin{array}{c|cc}
X & 1 & 2 \\
\hline
0 & \pi_{01} & \pi_{02} \\
1 & \pi_{11} & \pi_{12} \\
\end{array}
\]

and

\[ \logit(\pi_{ik}) = \begin{cases} 
\beta_0 & \text{if } k = 1 \\
\beta_0 + \beta_2 & \text{if } k = 2
\end{cases} \]

and therefore

\[ \log \left( \frac{\pi_{12}/(1 - \pi_{12})}{\pi_{11}/(1 - \pi_{11})} \right) = \beta_2 \]

and \(\beta_2\) is the log-odds ratio. If \(X\) takes \(K\) levels,

\[ \log \left( \frac{\pi_{1k}/(1 - \pi_{1k})}{\pi_{11}/(1 - \pi_{11})} \right) = \beta_k \text{ for } k = 2, \ldots, K \]

and \(\beta_K\) is the log-odds for level \(k\) relative to baseline. Models with factor predictors only lead to contingency tables

- one factor: \(2 \times K\) table
- two factors: \(2 \times K \times L\) table
4.3 Case-control and $2 \times 2$ designs

Consider a binary response $Y$ with 0 indicates “unaffected” and 1 “affected” and a single binary factor $X$ taking value 0 if “unexposed” and 1 if “exposed”.

Two possible experimental designs

1. **Prospective approach:**
   At the start of study, discover $X$ and follow up to discover $Y$

2. **Retrospective approach:**
   Observe $Y$ to determine inclusion in study, then measure $X$.

In the resulting $2 \times 2$ table where we have

\[
\begin{array}{cc|c}
 & 0 & 1 \\
 \hline
0 & Y_{00} & Y_{01} \\
1 & Y_{10} & Y_{11} \\
\end{array}
\]

For (1), we can condition on the column totals $m_{0*}, m_{1*}$, while for (2), we **must** condition on the row column $m_{0*}, m_{1*}$.

These assumptions only allow us to study certain conditional probabilities.

1. $Y_{k*} \sim B(m_k, \pi_k)$ for $k = 0, 1$ i.e. Binomial in the columns where $\pi_k = P(Y = 1 | X = k)$ for $k = 0, 1$ whereas in
2. $Y_{*j} \sim B(m_j, \varpi_j)$ for $j = 0, 1$ i.e. Binomial in the rows and $\varpi_j = P(X = 1 | Y = j)$ for $j = 0, 1$.

Scientifically, we are interested in $P(Y = 1 | X = 0)$ versus $P(Y = 1 | X = 1)$; that is, what is the effectiveness rate within the people that were exposed or not (e.g. treatment or therapy or exposed/not exposed to a noxious substance in the environment), but not $P(X = 1 | Y = 0)$ versus $P(X = 1 | Y = 1)$.

Can design (2) informs us about any quantities of interest? Yes, as by Bayes theorem

\[
\pi_1 = P(Y = 1 | X = 1) = \frac{P(X = 1 | Y = 1) P(Y = 1)}{P(X = 1)} = \frac{\varpi_1}{P(X = 1)}
\]

therefore
\[
\frac{\pi_1}{1 - \pi_1} = \frac{P(X = 1 | Y = 1) P(Y = 1)}{P(X = 1 | Y = 0) P(Y = 0)} = \frac{\varpi_1}{\varpi_0} \frac{P(Y = 1)}{P(Y = 0)}
\]

and
\[
\frac{\pi_0}{1 - \pi_0} = \frac{P(X = 0 | Y = 1) P(Y = 1)}{P(X = 0 | Y = 0) P(Y = 0)} = \frac{1 - \varpi_1 P(Y = 1)}{1 - \varpi_0 P(Y = 0)}
\]
Therefore, dividing $\frac{\pi_1}{1-\pi_1}$ by $\frac{\pi_0}{1-\pi_0}$ yields

$$\frac{\pi_1}{1-\pi_1} \frac{1-\pi_1}{\pi_0} = \frac{\varpi_1/(1-\varpi_1)}{\varpi_0/(1-\varpi_0)}$$

and the odds ratio in the designs are precisely equal. The odds ratio can be estimated equivalently from the two designs (1) and (2). These retrospective design is extremely common in epidemiology. It is only an odds ratio, and we cannot learn about $\pi$’s, roughly approximating the denominator $\frac{1-\pi_0}{1-\pi_1} \approx 1$ for very rare disease and get $\frac{\pi_1}{\pi_0}$. The conditions under which this hold are very specific.

There are many measures of interest for contingency tables, namely rate of incidence, odds on incidence, risk difference risk ratio, odds ratio, excess relative risk, etc. These measures are on the prospective design; all the quantities are estimable. For definitions, we refer the reader to page 58. With any given link function, these quantities can be estimated. Usually, in epidemiology, $\pi$ is often times the identity link; the measure can be transformed back to the scale of the data. Depending of the measure used, the standard errors may change.

For the retrospective case, the use of the Delta method is necessary and transformations of $\hat{\pi_j}$ quantities such as $\log(\hat{\pi_i})$ are needed to derive the properties of the estimators.

**Example 4.6 (Handout - Smoking and lung cancer case-control study)**

We have

$$\hat{\psi} = \frac{\hat{\pi}_{11}/\hat{\pi}_{01}}{\hat{\pi}_{10}/\hat{\pi}_{00}} = \frac{y_{00}y_{11}}{y_{01}y_{10}}$$

in this particular example, by invariance properties of the maximum likelihood estimates. The standard error is the square root of the sum of the reciprocal of counts. We derive a Wald-type test statistic for the log-odds ratio. From the handout, we have $\log(\hat{\psi}) \sim N(\log(\psi), y_0^{-1} + y_0^{-1} + y_1^{-1} + y_1^{-1})$. We have a strong significant effect of smoking of lung cancer, the statistic is big and positive so the effect is positive (smoking increases incidence of lung cancer). Since zero is excluded from the 95% confidence interval, we conclude that the effect is significant. On the log-odds ratio scale, we want the quantity to exclude 1. We cannot study the prevalence of cancer from this table, since the number of cases and controls matched is selected a priori.

**Example 4.7 (Aspirin and Myocardial Infarction Randomized trial)**

This prospective study is a randomized trial with a placebo group. We have fairly low case of fatal myocardial infarction. What is interesting here is that the Wald statistic is negative, so relatively “yes” (i.e. taking aspirin) is associated with “no” response, meaning
that aspirin is protective against fatal myocardial infraction. One might argue whether other measures should be used in place of the odds-ratio. Here, the confidence interval is monotonically transformed using \( \exp(\cdot) \) for the log-scale. We have \( \text{PE}/(1-\text{PE}) \) with the odds ratio. Don’t fall in the epidemiological tendency of confounding the risk ratio with the log-odds ratio; they are not the same (only approximate) and you should clearly state the necessary assumptions.

**Extension to \( 2 \times 2 \times K \)**

Suppose we have \( Y, X, Z \) with \( X, Y \) are binary response and factor with \( Z \) – factor predictor taking \( K \) levels (for e.g. multicenter case-control study). We need to model

\[
P(Y = 1|X = 1, Z = k) = \pi_{11k} \\
P(Y = 1|X = 0, Z = k) = \pi_{10k}
\]

for each \( k = 1, \ldots, K \). We might be interested in comparing

\[
\psi_k = \frac{\pi_{11k}/(1-\pi_{11k})}{\pi_{10k}/(1-\pi_{10k})}
\]

– odds ratio in the \( k \)th \( 2 \times 2 \) table; think of it as a \( K \) parallel contingency tables. We can use the binomial GLM approaches in the 2 factor problem. We can fit main effects and interaction between \( X \) and \( Z \) to explain the variation in \( Y \). e.g. fit \( Y \sim X + Z \) or \( Y \sim X \times Z \).

See Exercise 3. To summarize the relation between \( Y \) and \( X \) of interest, one needs to get average over the levels of \( Z \) in some sense, if no simple relationship holds and \( Z \) modifies the interaction between the variables.

**Note**

Other link functions than logit may be used (e.g. probit, complementary log-log, etc.)

**4.4 Overdispersion for Binomial data**

Suppose \( Y_i \sim \text{B}(m_i, \pi_i) \) for \( i = 1, \ldots, n \). Now we can write \( Y_i \overset{d}{=} \sum_{j=1}^{m_i} Y_{ij} \) where \( Y_{ij} \sim \text{B}(\pi_i) \) are independent. If we allow \( \pi_i \) to be a **random** quantity,

\[
\begin{align*}
\mathbb{E}(\pi_i) &= \lambda_i \quad (0 < \lambda_i < 1) \\
\text{Var}(\pi_i) &= \kappa \lambda_i(1 - \lambda_i) \quad (\kappa > 0)
\end{align*}
\]

then

\[
\begin{align*}
\mathbb{E}(Y_i) &= m_i \lambda_i \\
\text{Var}(Y_i) &= m_i \lambda_i(1 - \lambda_i)[1 + (m_i - 1)\kappa]
\end{align*}
\]
by repeated application of the iterated variance formula and remark that $[1+(m_i-1)\kappa] \geq 1$.

Indeed, since $E_\pi (\pi^2) = \text{Var}_\pi (\pi) + (E_\pi (\pi))^2 = \lambda(\kappa(1-\lambda) + \lambda)$

\[
\text{Var}(Y) = E_\pi(\text{Var}_{|\pi}(Y)) + \text{Var}_\pi (E_{|\pi}(Y)) = mE_\pi (\pi(1-\pi)) + m^2 \text{Var}_\pi (\pi) = m\lambda(1 - [\kappa(1-\lambda) + \lambda]) + m^2 \kappa \lambda(1-\lambda) = m\lambda(1-\lambda)[1 + (m - \kappa)]
\]

as claimed.

The variance of the new $Y_i$ is greater than that implied by the binomial model, i.e. the model with

\[
E(Y_i) = m_i, \quad \text{Var}(Y_i) = \phi m_i \lambda_i (1 - \lambda_i)
\]

with $\phi > 1$ is a suitable model for overdispersion. This model is not based on a particular parametric model: however, we may still model.

\[
\log \left( \frac{\lambda_i}{1 - \lambda_i} \right) = x_i \beta
\]

The attention went from modelling $\pi$ to modelling its expected value $\lambda_i$ and use the estimating function/equation

\[
\sum_{i=1}^n (y_i - m_i \lambda_i(\beta)) x_i^T = 0
\]

to estimate $\beta$. Finally, we may use

\[
\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^n \frac{(y_i - m_i \lambda_i(\hat{\beta}))^2}{m_i \lambda_i(\beta)(1 - \lambda_i(\beta))}
\]

to estimate $\phi$. This formulation represents ‘clustering’ amongst the $Y_{ij}, j = 1, \ldots, m_i$ around $\pi_i$.

This is what is used in R under the quasibinomial family. For purely parametric approach, take $\pi_i \sim B(\alpha, \beta)$. This leads to a closed form for the distribution of $Y_i$, which is the so-called Beta-Binomial model (exercise).

Zero-inflation model with a point mass can be used to model underdispersion. It is very hard to introduce model for underdispersion, some proposals include power constructions (taking power of a discrete distribution and renormalizing – one can increase as well in that way by taking low powers). Other methods include the use of a Markov chain that has
any stationary distribution that you want by specifying the probability transition matrix.
Inference is incredibly difficult to do for those models, they are artificial in a way.

4.5 Multinomial responses

Here, each response $Y_i$ can take on of the $K$ values, $Y_i \equiv (Y_{i1}, \ldots, Y_{iK})$ is polytomous.
The multinomial distribution is used for such data; it has PMF (proportional to)
$$
\prod_{k=1}^{K} \pi_{ik}^{y_{ik}}
$$
for datum $i$, where $\pi_{ik}$ is defined for an individual by $\pi_{ik} = P\left(Y_i^{(R)} = k|X_i = x_i\right)$ in general.
Therefore, the log likelihood is
$$
\sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log(\pi_{ik})
$$
where, in a GLM, we specify
$$
g(\pi_{ik}) = x_i \beta
$$
(where $Y_i^{(R)} = k \Leftrightarrow Y_{ik} = 1, Y_{ij} = 0 \forall j \neq k$)

Multinomial data

Let $Y_i \sim \mathcal{M}(m_i, \pi_1, \ldots, \pi_K)$ and $Y_i \overset{d}{=} \sum_{j=1}^{m_i} Y_{ij}$ where $Y_{ij} \sim \mathcal{M}(1, \pi_1, \ldots, \pi_K)$ independent for $j = 1, \ldots m_i$. See the handout for grouped representation; we could also have rows with one entry for each individual, for the individual-level representation. To fit the model, we use the multinom function from the nnet library. We can also use a log-linear model and interpret the results as arising from Poisson model.

Example 4.8 (Birth malformation data)

See the data on page 60; the log-linear model has 3 or 6 parameters, with different (and category-specific) responses. We can have constant intercept for the model for the different outcomes, or use a model where the counts are varying with dose level. We lose a parameter when moving from the Poisson to the Multinomial representation.

The fit using the multinom function exploit the link between Poisson log-linear model and Multinomial logistic models, i.e we can fit the model where $\pi_k(x_i) = P(Y_{ij} = k|X_i = x_i)$
for $j = 1, \ldots, m_i$ and $k = 1, \ldots, K$ where
\[
\log \left( \frac{\pi_k(x_i)}{\pi_1(x_i)} \right) = \beta_{0k} + x_i \beta_{1k}
\]
That is,
\[
\pi_k(x_i) = \frac{\exp(\beta_{0k} + x_i \beta_{1k})}{1 + \sum_{l=2}^{K} \exp(\beta_{0l} + x_i \beta_{1l})}
\]
The choice of category 1 as a baseline is arbitrary, but it is the default in R.

Moving from the three parameters Poisson model to the two parameter Multinomial model, we lose one parameter, namely the baseline. It reports the coefficients, standard errors and residual deviance and AIC. In the example, we add the non-baseline slope. Although there are differences in the three level of outcome; the fact there are different $\pi$ for different outcome categories already indicates the change. The more important is that the proportions in each row change as the dose level change. It is that dependence of the covariates that we are interested in (i.e. the slopes).

Recall that for identifiability, $\beta_{01} = \beta_{11} = 0$.

This imposes the specific dependence structure and how the changes in covariates impacts the proportions. When dropping the logistic link, however, we will see that the formulation used above does not hold anymore.

Other models for multinomial response

- **Cumulative logit:** for ordinal response (i.e. ordered categorical), we instead construct a model for the cumulative logits.

  That is, $P(Y_i^{(R)} \leq k \mid X_i = x_i) = \pi_1(x_i) + \cdots + \pi_k(x_i)$ is the probability of observing a response in categories $\{1, \ldots, k\}$ and we model
\[
\log \left( \frac{P(Y_i^{(R)} \leq k \mid X_i = x_i)}{1 - P(Y_i^{(R)} \leq k \mid X_i = x_i)} \right) = \log \left( \frac{\pi_1(x_i) + \cdots + \pi_k(x_i)}{\pi_{k+1}(x_i) + \cdots + \pi_K(x_i)} \right)
\]
as a function of the covariates. To construct the likelihood for this model, we need explicit forms for $\pi_k(x_i)$ by inverting this more complicated link function. Note that
\[
\pi_k(x_i) = P(Y_i^{(R)} = k \mid X_i = x_i) = P(Y_i^{(R)} \leq k \mid X_i = x_i) - P(Y_i^{(R)} \leq k - 1 \mid X_i = x_i)
\]
Proportional odds model: in this case, we look at

$$\logit \left( P \left( Y_i^{(R)} \leq k \mid X_i = x_i \right) \right) = \beta_{0k} + x_i \beta_1$$

where $\beta_1$ is common for all $k$ and $\beta_{01} < \beta_{02} < \cdots < \beta_{0(K-1)}$. For scalar $x_i$,

$$\logit \left( P \left( Y_i^{(R)} \leq k \mid X_1 \right) \right) - \logit \left( P \left( Y_i^{(R)} \leq k \mid X_2 \right) \right) = (x_1 - x_2) \beta_1.$$

Therefore the (cumulative) log-odds quantities change linearly with changes in $x$ for each $k$. Inference follows from the likelihood involving $\pi_k(x_i)$.

Example 4.9 (Mental impairment)

We have 40 observations with ordered categories for outcome. We have two covariates to model, socio-economic status and number of life events (LE). There is an underlying continuous variable and we discretize the continuum into four categories.

The fitting is done using the `polr` function from the `MASS` library, used in exactly the same way as the `glm` function. The response is recorded on an individual basis, 12 in the first category, 12 in the second, and 7 and 9 for the last categories. $1/2$ corresponds in our notation to $\beta_{01}$ (i.e. the probability left of $k$). Note that the non-intercept terms are reported on the negative scale, i.e. $-\beta_1$. See the note on page 68. We can carry analysis of deviance using the `drop1` function and drop `ses` on the basis of the 5% level. We can use the `VGAM` library via the `vglm`. The default uses the $\geq$ probabilities, and to get the model corresponding to the class notes, use `reverse=F` option. The number of degrees of freedom is here $3 \times 40 - 5$ instead of 35, since `VGAM` treats the data as vector, and the degree of freedom for the multivariate data would be counts of individuals; it differs with the interpretation we have had up to this point. Otherwise, the output is more standard and the sign for the parameters is correct in this model.

There are many families available in the `VGAM` library, for adjacent categories, the usual multinomial fitting the logit, cumulative odds, etc. The probabilities odds and probabilities reported y the model, whether logit($P( Y = k \mid X = x)$, $P( Y = k + 1) / P( Y = k)$ or $P( Y = k + 1) / P( Y = k$ or $Y = k + 1)$ yielding $\pi_i(x)$ can be compared for the different models.

---

$^{67}$In this case, using the last category as the baseline and look at $k = 1, \ldots, k-1$ since the cumulative probability is 1. We can also look at $P \left( Y_i^{(R)} \geq k \mid X_i = x_i \right)$ would lead naturally to the choice of $k = 1$ as the baseline.

$^{68}$Since the data was ordered in the table to begin with.
Model checking for binomial/multinomial data

For binomial/multinomial data, Pearson’s chi-squared measure $X^2$ may be used for “grouped” data. $(M_i > 1)$ with $Y_i \sim \mathcal{B}(m_i, \pi_i)$ where

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - m_i \pi_i)^2}{m_i \pi_i (1 - \pi_i)}$$

is approximately $\chi^2_{n-p}$ distributed when $n$ is large – this formula for $X^2$ extends to the multinomial case. For individual level data $(m_i = 1)$, the asymptotic approximation is poor. For binomial responses, we instead look at different tailored measures to look at goodness of fit.

Misclassification statistics

For binomial data, $Y_i \in \{0, 1\}$, we form the table where

$$\hat{Y}_i = \begin{cases} 0 & \text{if } \hat{\pi}_i < \tau \\ 1 & \text{if } \hat{\pi}_i \geq \tau \end{cases}$$

for some threshold $\tau$: with $a + b + c + d = n$. The resulting misclassification table gives an

Table 1: Misclassification table (or confusion matrix)

<table>
<thead>
<tr>
<th>$Y_i$</th>
<th>Predicted $\hat{Y}_i$</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>$Y_i$</td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td></td>
<td>$\hat{Y}_i$</td>
<td>$c$</td>
<td>$d$</td>
</tr>
</tbody>
</table>

indication of how well the model predicts the observed data – we may report misclassification rates or the sensitivity and specificity of the model.

Definition 4.1 (Sensitivity and specificity)

We define the sensitivity to be $\alpha(\tau) = \frac{d}{c+d}$, for the proportion of correctly predicted ones, and the specificity as $\beta(\tau) = \frac{a}{a+b}$ for the proportion of true zeros. If we think about plotting the functions as $\tau$ changes.

This leads the receiver operating characteristic curve (ROC curve), which plots $\alpha(\tau)$ versus $1 - \beta(\tau)$ for $0 < \tau < 1$. A “good” model is one that classifies zeros and ones accurately when the threshold $\tau$ is optimally chosen: such as model has an ROC curve that, for some $\tau$, is near to the (0,1) corner of the unit square. This can be formalized by inspecting the area under the ROC curve (AUC), which should be near 1 for a “good” model.
“optimal” choice of the threshold is the value of \( \tau \) defining the point on the ROC curve closest to \((0, 1)\).

However, we may think that badly predicting zero for one is worst thing than mispredicting a one as a zero. This leads to differential concern for misclassification; we can change the
calculation to adjust for this fact. A more potent criticism is a purely within sample procedure. You might want to build an increasingly complex model to get no misclassification; however, as in the linear regression setting when driving the $R^2$ statistic by adding irrelevant covariates. There is no penalty for the complexity for the model, just as for the Pearson $X^2$ statistic. It is thus not enough; using a hold-back sample of say 10% and look at the predictive ability for this residual data. Training set (test set) construction is widely used for this kind of model. Note that there is also a package in R to plot the ROC curve.

4.6 Conditional Inference

The idea is to substitute in sufficient statistic in the place of “nuisance parameters”, parameters that are associated with the covariates are not nuisance parameters, however, parameters such as the intercept tells us nothing about the influence of the parameters on the model.

Conditional Logistic Regression

Let $Y_i \sim \mathcal{B}(\pi_i(x_i))$ where $\logit(\pi_i(x_i)) = \beta_0 + x_i \beta$. In this setup, $\beta$ is the parameter of interest, while $\beta_0$ is a nuisance parameter. By standard arguments (the full details are given in the printed notes), we can access for $n$ data under the logistic link model write

$$L_n(\beta_0, \beta) = \exp \left( \frac{T_1(y)\beta_0 + T_2(x, y)\beta}{\prod_{i=1}^{n} (1 + \exp (\beta_0 + x_i \beta))} \right)$$

where

$$T_1(y) = \sum_{i=1}^{n} y_i \quad T_2(x, y) = \sum_{i=1}^{n} y_i x_i$$

are sufficient statistics for $\beta_0, \beta$ respectively.

Consider conditioning on $T_1(y) = t$ and computing the conditional likelihood

$$L_{n_T}(\beta, t) = P(Y_1 = y_1, \ldots, Y_n = y_n \mid T_1(y) = t) = \frac{\exp(T_2(x, y)\beta)}{\sum_{z \in \mathcal{Y}(t_1)} \exp(T_2(x, z)\beta)}$$

where

$$\mathcal{Y}(t_1) = \left\{ z = (z_1, \ldots, z_n) : \sum_{i=1}^{n} z_i = t \right\}$$

When $n$ is small, we may view this method as being attractive as we have one less parameter to estimate and the enumeration of the $z$ is easier to do. The generality of this approach is not obvious, but we may use this procedure for other nuisance parameters.
Stratification or matching, or small number of observations are occasions where we would one to use a conditional likelihood approach and not estimate the nuisance parameters. This approach is no free lunch, since there is same additional computational burden to the estimation of the sum in the denominators, possibly large. Marginalization over strata may be quite a large computation to do. The distribution is intractable from a Bayesian perspective, since the numerator and denominator both involve the parameters, so we cannot avoid the calculation. In matched pairs analysis, we could use this to marginalize some of the observations.

4.7 Matched pair

This generalizes to more than the simple $2 \times 2$ table case, but the former is most illustrative for our purposes.

The design for the data collected at two time points: $j = 1, 2$. For a collection of individuals, we record a binary response at $j = 1$ and $j = 2$. The cell counts in the resulting $2 \times 2$ table can be modeled using $M(n; \pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$, assuming $n$ individuals are sampled independently from the population. Inference about $\pi_{00}, \ldots, \pi_{11}$ is straightforward.

However, the focus of interest is the difference between the ‘marginal’ probabilities, i.e. $\pi_{1\ast} = \pi_{10} + \pi_{11} = P(Y = 1$ at $j = 1)$. Similarly, $\pi_{\ast 1} = \pi_{01} + \pi_{11} = P(Y = 1$ at $j = 2)$. The estimation of $\pi_{\ast 1}$ and $\pi_{1\ast}$ by invariance of the MLE estimates. The problem now is that those estimates $\hat{\pi}_{1\ast}$ and $\hat{\pi}_{\ast 1}$ are dependent. In fact,

$$\text{Cov} (\hat{\pi}_{1\ast}, \hat{\pi}_{\ast 1}) = \frac{\pi_{00}\pi_{11} - \pi_{01}\pi_{10}}{n}$$

In this study, we wish to test, $\hat{\pi}_{1\ast} = \hat{\pi}_{\ast 1}$ (which implies $\pi_{0\ast} = \pi_{\ast 0}$) – this is an assumption of marginal homogeneity. Let $\delta = \pi_{1\ast} - \pi_{\ast 1}$, then $\delta = 0 \Rightarrow \pi_{1\ast} = \pi_{\ast 1} \Rightarrow \pi_{10} = \pi_{01}$, corresponding to a symmetry structure. So if $\delta = \hat{\pi}_{1\ast} - \hat{\pi}_{\ast 1}$ we have by the Delta method

$$\sqrt{n}(\hat{\delta} - \delta) \overset{d}{\to} N(0, V(\pi))$$

where $V(\hat{\pi}) = (\hat{\pi}_{10} + \hat{\pi}_{01}) - (\hat{\pi}_{10} - \hat{\pi}_{01})^2$.\textsuperscript{71} For more details, see the handout. The test

\textsuperscript{69}The multinomial is still a valid model, conditioning on a sample size, with observations arising from a target population with observations picked at random observed twice.

\textsuperscript{70}Giving $2n$ observations with independent pairs

\textsuperscript{71}This is not a score test, and we should be using under the null with the second term being equal to zero. Both tests, along with the Wald test, are equivalent on the null and use asymptotic and large sample approximation results. The tests are not equivalent under the alternative. Not squaring the test allows to get the direction of the result, as opposed to the squared Wald statistic, where $W_n^2 \sim \chi_1^2$.  

83
statistic is one of a symmetry test.

Example 4.10 (Prime minister approval)
See the handout. The analysis and testing is easy to get, however the big difference is in the dependence structure for the multinomial data. For the conditional analysis, we can use the `clogit` function in R. The function looks at the binary respond for approve, disapprove. The `strata` function (from the `survival` library) identifies the pairs. You supply 3200 observations with an id variable. The Z statistic is rather similar to the one above, but it uses a different formulation. We have not written the parametric model for this, so the coefficient is not interpretable yet, but we will derive shortly that the parameter is actually an odds ratio. If there is pairing in a experiment, one has to acknowledge it.

Conditional logistic regression for matched pairs or stratified data

Let \((Y_{1i}, Y_{2i})\) be the responses for study \(j = 1, 2, \ldots\) for individual \(i\); \(\pi_{ij} = P(Y_{ij} = 1)\) is the probability of “\(Y = 1\)” response for individual \(i\) on study \(j\). Consider the model

\[
\text{logit}(\pi_{ij}) = \begin{cases} 
\beta_{0i} & \text{if } j = 1 \\
\beta_{0i} + \zeta & \text{if } j = 2
\end{cases}
\]

This is indicating that the data is not IID; individuals are acknowledged to be different, with somewhat similar effect to a random effect model. That is, \(\beta_{0i}\) measures an individual-level (stratum-specific) component of the response probabilities. Thus \(\zeta\) is the common log-odds ratio comparing \(Y = 1\) responses for \(j = 2\) with \(j = 1\).

See the indications. This model has \(n + 1\) parameters (one per data pair), since we have 2 observations per individual, we could have identifiability. However, theory with models in which the number of parameters grows linearly with the number of data points is not standard. The likelihood \(L_n(\beta_{01}, \ldots, \beta_{0n}, \zeta)\) for the data is

\[
\prod_{i=1}^{n} \left( \text{expit}(\beta_{0i} y_{i1} (1 - \text{expit}(\beta_{0i}))^{1-y_{i1}} \text{expit}(\beta_{0i} + \zeta)^{y_{i2}} (1 - \text{expit}(\beta_{0i} + \zeta))^{1-y_{i2}} \right)
\]

We may thus consider the \(\beta_{0i}\) as nuisance parameters and replace them with sufficient statistic and condition on those. We see from the interpretation that individuals that do not change their mind do not affect the likelihood value, while those who change from one strata to the other contribute to the evaluation of \(\zeta\). We have

\[
P(Y_{1i} = y_{1i}, Y_{2i} = y_{i2} \mid S_i = s) = \begin{cases} 
1 & \text{if } (s, y_{1i}, y_{i2}) = (0, 0, 0) \text{ or } (2, 1, 1) \\
\frac{e^{\gamma_{12s}}}{{1 + e^{\gamma_{12}}}^s} & \text{if } (s, y_{1i}, y_{i2}) = (1, 0, 1) \text{ or } (1, 1, 0)
\end{cases}
\]
The parameter $\tilde{\zeta}$ has a different form, given by $\tilde{\zeta} = \log \left( \frac{n_{01}}{n_{10}} \right)$ and $\tilde{\sigma} = \sqrt{n_{01}^{-1} + n_{10}^{-1}}$.

The conditional likelihood approach is very appealing when the number of parameter is very large or grows with the data.
5.1 Gamma GLM

The Gamma model for continuous response $Y$ can form the basis of a GLM. The canonical parameter is $\theta = \frac{-1}{\mu}$, the cumulant function is $b(\theta) = -\log(-\theta)$ and the dispersion parameter is $\phi = 1/\nu$ assuming $G(\mu, \nu)$ where $\mathbb{E}(Y) = \mu, \mathbb{Var}(Y) = \mu/\nu$ (here, $\alpha = \nu, \beta = \nu/\mu$ in the usual variance formulation).

The link function is the “inverse”, yet in $R$ we have $g(t) = \frac{1}{t}$ so the minus sign is removed and as such the interpretation is changed. The log link $g(t) = \log(t)$ is also a plausible choice, which would capture more perhaps of the extreme variability.

The linear predictors are

\[
\eta = \beta_0 + \beta_1 x \\
\eta = \beta_0 + \beta_1 / x \\
\eta = \beta_0 + \beta_1 / x + \beta_2 x
\]

In the clotting example, we assume that the two subjects are IID. It is possibly the case that with $\log(t)$ with reconstruction of the mean was sensible, since the values could be negative. The analysis is fairly straightforward; the one slightly different of the output is the dispersion parameter which is not fixed to 1. This does not make a difference for the analysis of deviance.

We could add a binary indicator for difference in the patient. The fit using the Pearson residuals looks okay; it is easier to evaluate with continuous family of response like standard residual plots as compared to the discrete data case. The way the clothing time affects the patient is similar, by looking at the parameter estimate and significance. Note that the dispersion parameter changes when fitting a different model, for example the additive model. In this case, all parameters seem very significant, and the additive model is more appropriate at explaining the data.

**Note**

To compute the Analysis of Deviance test statistics, to compare nested models, we use

\[
\frac{\Delta \text{Deviance}}{\hat{\phi}} \sim \chi^2_{\Delta \text{df}}
\]

where $\hat{\phi}$ is computed for the more complex model.

Using the `drop1` function with `test="Chisq"` allows to look at single term deletion. In the
end both interactions will be dropped. There is significant difference in patient type, but
the dependence on \( \log(u) \) and \( 1/\log(u) \) is the same for both patients.

The gamma model can be extended like the binomial by allowing for different dispersion for
different individuals.

**Note**

If \( Y_i \sim G \) and \( \mathbb{E}(Y_i) = \mu_i \), \( \text{Var}(Y_i) = \frac{\mu_i^2}{\nu} \) where \( \nu_i = w_i \nu \) with \( w_i \) **fixed** (known quantity) but possibly changing with \( i \). Then the individual-level dispersions can be incorporated using “case weights” yielding for example the Deviance.

\[
D(y; \tilde{\mu}) = -2 \sum_{i=1}^{n} w_i \left[ \log \left( \frac{y_i}{\tilde{\mu}_i} \right) - \left( \frac{y_i - \tilde{\mu}_i}{\tilde{\mu}_i} \right) \right]
\]

–introduced using the weights argument in glm. Finally, the dispersion estimate given in
\( R \) is the Pearson based estimate of \( \phi = \frac{1}{\nu} \) is

\[
\hat{\phi} = \frac{1}{n - p} \sum_{i=1}^{n} \left( \frac{y_i - \tilde{\mu}_i}{\tilde{\mu}_i} \right)^2
\]

if \( w_i \equiv 1 \ \forall \ i. \)

The Gamma model is used in survival analysis or reliability analysis. The tractability is
however not as easy as with a Weibull or an exponential distribution, can’t use glm function
unless there is no censoring, i.e. we can’t fit censored data using this technology.

**5.2 Quasi-likelihood**

The GLM estimating equations (score equations) take the general form

\[
\sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{V(\mu_i)\tilde{g}(\mu_i)} x_{ij} = 0, \quad j = 1, \ldots, p
\]

where \( \mu_i \) is a function of the unknown parameters \( \beta \) and in this case \( g \) is the link function:

\[
\eta_i = x_i \beta
\]

\( g(t) \) defines the relationship between \( \eta_i \) and \( \mu_i \) i.e. \( g(\mu_i) = \eta_i \).

Suppose now we relax the need for a ‘distribution-based’ construction of the estimating
equation, i.e. break the link encapsulated by \( V(\mu) \). Instead, let \( \mathbb{E}(Y_i) = \mu_i \) and \( \text{Var}(Y_i) = \sigma^2 V(\mu_i) \) for an arbitrary \( V(\cdot) \) returning a positive quantity.

Let \( U_i = \frac{Y_i - \mu_i}{\sigma^2 V(\mu_i)} \). Then \( \mathbb{E}(U_i) = 0 \) and \( \text{Var}(U_i) = (\sigma^2 V(\mu_i))^{-1} \) and also –\( \mathbb{E} \left( \frac{\partial U_i}{\partial \mu_i} \right) = \)
\[
\frac{1}{\sigma^2 V(\mu_i)} \text{ thus } U_i \text{ has the same properties as a score random variable.}
\]

\[
S(Y_i, \theta) = \frac{\partial}{\partial \theta} \log f_T(Y_i; \theta).
\]

Therefore, we could use

\[
\sum_{i=1}^{n} U_i(\beta) = 0 \quad (p \times 1)
\]

to estimate \( \beta \) estimators derived as the solution to this “quasi-score” equation are consistent and asymptotically normally distributed.

The corresponding ‘quasi-likelihood’ is obtained from the form of \( U_i \), i.e. the quasi-likelihood contributions the form

\[
Q(\mu_i, y_i) = \int_{y_i}^{\mu_i} \left( \frac{y_i - t}{\sigma^2 V(t)} \right) \, dt
\]

and the quasi-deviance \( D_i(\mu_i; y_i) = -2\sigma^2 Q(\mu_i, y_i) \).

Asymptotic variance

If \( g(\mu) = X\beta \ (n \times 1) \), \( V(\mu) = \text{diag} \ (V(\mu_1), \ldots, V(\mu_n)) \) an \((n \times n)\) covariance matrix, \( B(\mu) \) is the \( n \times p \) matrix with \((i, j)^{th}\) term \([B]_{ij} = \frac{\partial \mu}{\partial \beta_j}\) for \( i = 1, \ldots, n \) and \( k = 1, \ldots, p \) and

\[
U(\beta) = \frac{1}{\sigma^2} B^T (V(\mu))^{-1} (Y - \mu)
\]

an \((n \times 1)\) system. Then \( E(U(\beta)) = 0_n \) and \( \text{Var}(U(\beta)) = \frac{1}{\sigma^2} B^T V^{-1} B \) and \( \hat{\beta} \) satisfies \( U(\hat{\beta}) = 0_n \) (an \((n \times 1)\) system).

Thus

\[
E(\hat{\beta}) \approx \beta
\]

\[
\text{Var}(\hat{\beta}) \approx \sigma^2 \left( B^T V^{-1} B \right)^{-1}.
\]

When \( n \) is large, by analogy with weighted least squares, \( \hat{\beta} \) is computed using recursive methods e.g. Fisher scoring

\[
\beta^{(t+1)} = \beta^{(t)} + \left( B^{(t)} V^{(t)} B^{(t)} \right)^{-1} B^{(t)} V^{(t)} (y - \mu^{(t)})
\]
yields $\hat{\beta}, \hat{B}, \hat{V}$ and $\hat{\sigma}^2$ where

$$\hat{\sigma}^2 = \frac{1}{n - p} \sum_{i=1}^{n} \left( \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} \right)^2$$
Index

aliased, 12

case-control, 73
  prospective approach, 73
clustering, 76
conditional inference, 82
consistency, 35
contingency table, 43, 54
  independence, 54
  independence model, 54
  partial independence, 54
contingency tables
  measures, 74
  odds ratio analysis, 74
covariates, 4
dependent variable, 4
deviance, 37
  scaled, 37
discrete predictor, 4

exponential family, 18
  canonical parameter, 18
  dispersion parameter, 18
exponential-dispersion family, 18

factors, 4
Fisher information, 32
Fisher scoring, 33

generalized linear models, 18

identifiability, 13
incomplete designs, 13
independent variable, 4

information criterion
  Akaike’s (AIC), 49
  Schwartz’s (BIC), 49

interaction, 12
interactions, 14
  higher order, 15
  without main effects, 15
inverse link function
  expit, 22

least squares
  estimation, 5
  generalized, 6
  geometric interpretation, 8
  iteratively reweighted, 31
  ordinary, 6
  weighted, 6
least-squares criterion, 9

likelihood ratio test, 36

linear model, 4
linear predictor, 21

link function, 21, 22
  canonical, 23
  log, 23
  logistic (logit), 22
  reciprocal, 23
  complementary log-log, 22
  probit, 22
logistic regression
  parameters interpretation, 72
  connection with Poisson regression, 69
main effect and interaction, 12
marginal homogeneity, 56
matched pairs, 83
misclassification statistics, 80
multicollinearity, 13
multinomial data, 61
  model checking, 80
multinomial model
cumulative logit, 78
grouped representation, 77
individual-level representation, 77
logit, 77
proportional odds, 78

multinomial regression

links with Poisson, 71

nested models, 37, 39
Newton-Raphson method, 32
null model, 37

odd, 69
odds ratio, 26
offset, 43, 46
orthogonal
sequential fitting, 14
outcome, 4

overdispersion
Binomial, 75
Poisson, 58

Pearson $X^2$ statistic, 39
polytomous, 77
projection, 9
projection matrix, 11

quadratic form, 5
quasi-deviance, 88
quasi-likelihood, 87
quasi-score equation, 88

residuals
Anscombe, 40
deivance, 41
Pearson, 40
response variable, 4
ROC curve, 80

saturated model, 27
score equations, 29

score test, 35
sensitivity, 80
specificity, 80

square tables
log-linear
quasi-independence, 55
symmetry, 56

underdispersion
Poisson, 58

Wald statistic, 36
Creative Commons Attribution-Non Commercial-ShareAlike 3.0 Unported

You are free:
to Share - to copy, distribute and transmit the work
to Remix - to adapt the work

Under the following conditions:
Attribution - You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
Noncommercial - You may not use this work for commercial purposes.
Share Alike - If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar license to this one.

With the understanding that:
Waiver - Any of the above conditions can be waived if you get permission from the copyright holder.
Public Domain - Where the work or any of its elements is in the public domain under applicable law, that status is in no way affected by the license.
Other Rights - In no way are any of the following rights affected by the license:
Your fair dealing or fair use rights, or other applicable copyright exceptions and limitations;
The author’s moral rights;
Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights.

Course notes for MATH 523: Generalized Linear Model
Léo Raymond-Belzile
Full text of the Legal code of the license is available at the following URL.