MATH 523 - Generalized Linear Models

Pr. David A. Stephens

Course notes by
Léo Raymond-Belzile
Leo.Raymond-Belzile@mail.mcgill.ca

The current version is that of February 23, 2022
Winter 2013, McGill University

Please signal the author by email if you find any typo.

These notes have not been revised and should be read carefully.

Contents

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**).[1] $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})$ 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 $X_{i1}, X_{i1}^2, \ldots$). We could also have a **discrete** predictor *e.g.* $X \in \{1, 2, \ldots, L\}$, then

$$X_{ij} = \begin{cases} 1 & \text{if } X_i = j \\ 0 & \text{if } X_i \neq j \end{cases} = \mathbf{1}_j(X)$$

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.* $X_{ij} X_i^{(1)} \sqrt{X_i^{(2)}}$, etc.

We will denote **observed data** by lowercase variables; $y_i, \boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ a $(1 \times p)$ for the same individual, and by $\underset{\sim}{\boldsymbol{x}}_j = (x_{ij}, \ldots, x_{nj})^\top$ the $(n \times 1)$ vector for j accross individuals.

The **(general) linear model** is defined by considering the conditional expectation of Y_i, given $X_i = x_i$. We specify that

$$\mathsf{E}\left(Y_i | X_i = x_i\right) = \boldsymbol{x}_i \boldsymbol{\beta}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a $(p \times 1)$ column vector of parameters. Considering all data simultaneously we have

$$\mathsf{E}\left(\boldsymbol{Y} | \boldsymbol{X} = \boldsymbol{x}\right) = \boldsymbol{X}\boldsymbol{\beta}$$

[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 $\boldsymbol{X}$ is the $(n \times p)$ **design matrix** with $(i,j)^{\text{th}}$ element x_{ij} and where $\boldsymbol{\beta}$ is $(p \times 1)$.

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

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ is a $(n \times 1)$ vector of random variables with

$$\mathsf{E}(\varepsilon_i) = 0, \quad i = 1, \ldots, n \qquad\qquad \mathsf{Var}(\boldsymbol{\varepsilon}) = \mathsf{E}\left(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right) = \boldsymbol{\Sigma}$$

where $\boldsymbol{\Sigma}$ is an $(n \times n)$ symmetric positive definite matrix.[2] In the simplest case, $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the $(n \times n)$ identity matrix, that is $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_n$ are identical in terms of moments and are uncorrelated. More commonly,[3]

$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_n)$$

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

1.1 Estimation

In the linear model, we need to estimate $\boldsymbol{\beta}$ and σ^2. $\boldsymbol{\beta}$ is estimated typically using the **least-squares** criterion; we choose $\boldsymbol{\beta}$ to minimize

$$S(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

We minimize this quadratic form analytically by solving

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \boldsymbol{0}$$

a $(p \times 1)$ vector of equations (since $S(\boldsymbol{\beta})$ is a smoothly varying function). [4]

[2] We could also have $\boldsymbol{\Sigma}$ to be also function of $\boldsymbol{X}$, but this is not the case in this course. For example, one could have increase in variance with increase in values of $\boldsymbol{X}_i$, using in the linear case least squares.

[3] Finite-sample hypothesis testing and maximum likelihood estimation will require distributional assumption about the $\boldsymbol{\varepsilon}$, usually a normality assumption. In general, we need not be that restrictive to get optimality results in the context of the Gauss-Markov theorem.

[4] Making the normality assumption, $S(\boldsymbol{\beta})$ turns out to be the negative log-likelihood of the mode. We can get the solution of this system of equations analytically.

We have

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{Y} + 2\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{X}\boldsymbol{\beta}$$

so equating to zero, we obtain the **normal equations** that $\widehat{\boldsymbol{\beta}}$ solves

$$(\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{X})\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

and if $(\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{X})$ is non-singular we have

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{Y}. \tag{1.1}$$

If we make the simplifying assumption that $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$, then this simplifies to

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y},$$

the ordinary least square (OLS) formula.[5] The estimator

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{Y}$$

- for general $\boldsymbol{\Sigma}$ – is termed **Generalized Least Squares**
- if $\boldsymbol{\Sigma}$ is diagonal, say, $\boldsymbol{\Sigma} = \text{diag}\,(\sigma_1^2, \ldots, \sigma_n^2)$, then we often rewrite $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1} = \text{diag}\,(w_1, \ldots, w_n)$ so that

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

– termed **Weighted Least Squares**
- if $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$ – termed **Ordinary Least Squares.**

If $\boldsymbol{\Sigma}$ 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 $\boldsymbol{\Sigma}$ as a known quantity. In the OLS case, no estimation of σ^2 is necessary to estimate $\boldsymbol{\beta}$.

Note

$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. If we do a Cholesky decomposition, namely if $\boldsymbol{\Sigma} = \boldsymbol{M}\boldsymbol{M}^\top$, then

$$\boldsymbol{M}^{-1}\boldsymbol{Y} = \boldsymbol{M}^{-1}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{M}^{-1}\boldsymbol{\varepsilon}$$

or

$$\boldsymbol{Y}^* = \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$$

[5] In the GLS case, weighted least-squares or generalized least-squares, we use usually a two-step procedure, estimating the matrix $\boldsymbol{\Sigma}$ assumed to be function of few parameters, usually by some MLE equivalent method.

where $\mathsf{Var}\,(\boldsymbol{\varepsilon}^*) = \boldsymbol{I}_n$. 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 $\widehat{\boldsymbol{\beta}}$ are easy to derive using standard results for expectation and variance. We have unbiasedness of our estimator and

$$\mathsf{E}\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta} \qquad \mathsf{Var}\left(\widehat{\boldsymbol{\beta}}\right) = (\boldsymbol{X}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{X})^{-1}$$

which of course simplifies in the OLS case; if $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_n$, then $\mathsf{Var}\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$. These results hold in generality without any normality assumption. If $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_n)$, then $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1})$

Typically, σ^2 is unknown: it is usually estimated by

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

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 β'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

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

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

Note

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

[6] The inversion of $(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})$ 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.

asymptotically).[8]

1.3.1. Tests for linear restrictions

Suppose we have the null hypothesis

$$H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{C}$$

which imposes q linear constraints on $\boldsymbol{\beta}$, *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 accross the two models). In this case, let $\widehat{\boldsymbol{\beta}}_0$ be the parameter estimates under H_0 and $\widehat{\boldsymbol{\beta}}_1$ be the parameter estimates under $H_0 \cup H_1$. That is $\widehat{\boldsymbol{\beta}}_0$ is computed under the restriction imposed by H_0 and $\widehat{\boldsymbol{\beta}}_1$ is computed under no restrictions.

Let

$$S(\widehat{\boldsymbol{\beta}}_0) = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_0)^\top (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_0)$$
$$S(\widehat{\boldsymbol{\beta}}_1) = (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_1)^\top (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_1)$$

and note that $S(\widehat{\boldsymbol{\beta}}_0) \geq S(\widehat{\boldsymbol{\beta}}_1)$. Then

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

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

1.4 Geometric interpretation

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ 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 P-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 $\widehat{\boldsymbol{\Sigma}}$ quantity – the best possible estimate of $\boldsymbol{\Sigma}$ assuming the general model is appropriate

[10] `anova` or `aov` can be used in `R` to test for nested models. Most statistical procedures, like the $\mathcal{F}$-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 ($\mathbf{X}$). We can also think of $\mathbf{X}\boldsymbol{\beta} \in \mathbb{R}^n$, but restricted to the p-dimensional subspace

$$\mathcal{X}_p = \left\{ \mathbf{X} : \mathbf{X} = \sum_{j=1}^{p} \lambda_j \underset{\sim}{\boldsymbol{x}}_j \right\},$$

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

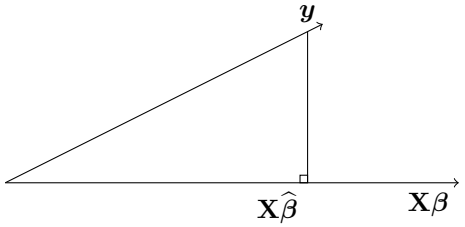
$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$$

where all observations are non-stochastic, $\boldsymbol{y}$ is the observed response, $\mathbf{X}\boldsymbol{\beta}$ is the fitted value and $\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}$ is the residual component. The **least-squares criterion** is to choose $\widehat{\boldsymbol{\beta}}$ as

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \end{aligned}$$

i.e. we choose $\widehat{\boldsymbol{\beta}}$ to minimize the "length" of the residual vector $(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})$. Clearly, $\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ is minimized when $\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ is **perpendicular** to $\mathcal{X}_p$ and $\mathbf{X}\widehat{\boldsymbol{\beta}}$ is the projection of $\boldsymbol{y}$ onto $\mathcal{X}_p$, the p-dimensional subspace.

Figure 1: Projection matrix



therefore the vector $\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ should be orthogonal to $\mathbf{X}\widehat{\boldsymbol{\beta}}$ *i.e.*

$$(\mathbf{X}\widehat{\boldsymbol{\beta}})(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = 0 \quad \Rightarrow \quad \mathbf{X}^\top \boldsymbol{y} = (\mathbf{X}^\top \mathbf{X})\widehat{\boldsymbol{\beta}}$$

which means that

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}$$

assuming that $\mathbf{X}^\top\mathbf{X}$ is non-singular. By Pythagoras' theorem, we have

$$\|\boldsymbol{y}\|^2 = \|\mathbf{X}\widehat{\boldsymbol{\beta}}\|^2 + \|\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2$$

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

The optimal estimate of $\boldsymbol{\beta}$, denoted by $\widehat{\boldsymbol{\beta}}$, defines the point in the 'model space', $\mathbf{X}\widehat{\boldsymbol{\beta}}$, that is nearest to $\boldsymbol{y}$. Note that also

$$\begin{aligned}\mathbf{X}^\top(\boldsymbol{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) &= \mathbf{X}^\top\boldsymbol{y} - (\mathbf{X}^\top\mathbf{X})\widehat{\boldsymbol{\beta}} \\ &= \mathbf{X}^\top\boldsymbol{y} - (\mathbf{X}^\top\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{y} \\ &= \mathbf{0}_p,\end{aligned}$$

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

For random variable $\boldsymbol{Y}$ and **estimator** $\widehat{\boldsymbol{\beta}}$

$$\mathsf{E}\left((\mathbf{X}\widehat{\boldsymbol{\beta}})^\top(\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\right) = 0$$

and also

$$\mathsf{E}\left(\mathbf{X}^\top(\boldsymbol{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}})\right) = \mathbf{0}_p$$

Finally, if $\boldsymbol{\beta}_0$ is the **true** value of $\boldsymbol{\beta}$, then

$$\begin{aligned}\mathsf{E}\left((\mathbf{X}\boldsymbol{\beta}_0)^\top(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_0)\right) &= 0 \\ \mathsf{E}\left(\mathbf{X}^\top(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}_0)\right) &= \mathbf{0}_p\end{aligned}$$

Note

The expectation $\mathsf{E}(\cdot)$ is done with respect to the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X} = \mathbf{X}$ at the true value $\boldsymbol{\beta}_0$. [12]

Summary

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

$$\begin{aligned}(\mathbf{X}\boldsymbol{\beta}_0)^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}_0) &= 0 \\ \mathbf{X}^\top(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}_0) &= \mathbf{0}_p\end{aligned}$$

[12] In general and unless stated otherwise, the expectations are with respect with respect to $\boldsymbol{Y}|\mathbf{X}$ under the true data generating mechanism.

respectively a (1×1) and $(p \times 1)$ systems, that is

$$\sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} \beta_j (y_i - \boldsymbol{x}_i \boldsymbol{\beta}) = 0$$
$$\sum_{i=1}^{n} x_{ij} (y_i - \boldsymbol{x}_i \boldsymbol{\beta}) = 0,$$

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

As $\mathbf{X}\boldsymbol{\beta} = \mathsf{E}(\boldsymbol{Y}|\mathbf{X}, \boldsymbol{\beta}) = \mu(\mathbf{X}; \boldsymbol{\beta})$, we may rewrite the previous expressions

$$\big(\mu(\mathbf{X}; \boldsymbol{\beta})\big)^\top \big(\boldsymbol{y} - \mu(\mathbf{X}; \boldsymbol{\beta})\big) = 0$$
$$\big(\dot{\boldsymbol{\mu}}(\mathbf{X}; \boldsymbol{\beta})\big)^\top \big(\boldsymbol{y} - \mu(\mathbf{X}; \boldsymbol{\beta})\big) = \mathbf{0}_p$$

wjere $\dot{\boldsymbol{\mu}}$ is the $(n \times p)$ matrix with $(i, j)^{\text{th}}$ element

$$\frac{\partial \mu_i(\mathbf{X}; \boldsymbol{\beta})}{\partial \beta_j} \mu_i(\mathbf{X}; \boldsymbol{\beta}) = \mathsf{E}(\boldsymbol{Y}_i|\mathbf{X}, \boldsymbol{\beta}) = \boldsymbol{x}_i \boldsymbol{\beta}.$$

Finally, note that

$$\widehat{\boldsymbol{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

where $\boldsymbol{H}$ is the $(n \times n)$ 'hat' matrix that defined the projection of $\boldsymbol{Y}$ onto the model space. Note that $\boldsymbol{H}$ is an idempotent matrix, *i.e.* $\boldsymbol{H}^\top \boldsymbol{H} = \boldsymbol{H}$.

1.5 Aliasing and multicollinearity

In the formulation of the linear model, the design matrix $\mathbf{X}$ is presumed to be full rank, so that $\mathcal{X}$ is a p-dimensional, the p columns of $\mathbf{X}$ are linearly independent, etc. If $\mathbf{X}$ 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

$$\mathsf{E}(Y_i|A_i = j) = \mu + \alpha_j, \qquad j = 1, 2, \ldots, a$$

–this model is often written

$$1 + \boldsymbol{A}$$

with parameters $\mu, \alpha_1, \ldots, \alpha_a$; the design matrix $\mathbf{X} = \left(\underset{\sim}{\boldsymbol{x}}_0, \underset{\sim}{\boldsymbol{x}}_1, \ldots, \underset{\sim}{\boldsymbol{x}}_a\right)$ where

$$\begin{aligned}\underset{\sim}{\boldsymbol{x}}_0 &= \mathbf{1}_n^\top \\ \underset{\sim}{\boldsymbol{x}}_j &= (x_{ij}, \ldots, x_{nj})^\top, \qquad j = 1, \ldots, a\end{aligned}$$

where

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

However, $\underset{\sim}{\boldsymbol{x}}_0 = \sum_{j=1}^a \underset{\sim}{\boldsymbol{x}}_j$, 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) $\widehat{\mu} = \bar{y}_1$, $\widehat{\alpha}_j = \bar{y}_j - \bar{y}_1$, for $j = 2, \ldots, a$.
(3) The third looks at the center tendency in the data and look at the differences from this as

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

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

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

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 μ 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.[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:

		B		
		1	2	3
A	1	x	x	o
	2	x	x	o
	3	o	o	x

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

$$\sum_{j=1}^{p} \lambda_j \underset{\sim}{x}_j \approx 0.$$

If there is a perfect multicollinearity such that

$$\sum_{j=1}^{p} \lambda_j \underset{\sim}{x}_j = 0,$$

[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 α_j, γ_{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 $\mathbf{X}$ is not full rank, $\mathbf{X}^\top\mathbf{X}$ is singular, even if there is "approximate" linear dependence, $\mathbf{X}^\top\mathbf{X}$ may be numerically singular, or numerically unstable on inversion.[16]

Note

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

$$\boldsymbol{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$$
$$\boldsymbol{Y} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{bmatrix} + \boldsymbol{\varepsilon}$$

may return different estimates for $\boldsymbol{\beta}_1$. This will occur if $\mathbf{X}_1^\top\mathbf{X}_2 \neq \mathbf{0}$. 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 $\mathbf{X}_1$, then $\mathbf{X}_2$, or $\mathbf{X}_2$, then fit $\mathbf{X}_1$ on residuals on the second model, or fit the two together resulting estimate –see handout for the linear algebra results. Only when $\mathbf{X}_1, \mathbf{X}_2$ are orthogonal will the least square estimates remain unchanged when doing sequential fitting versus fitting joint model – thus estimates of $\boldsymbol{\beta}$ 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 $\mathbf{X}_1$.

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.*

$$\begin{aligned}&\mathsf{E}\left(Y|X_1 = x_1, X_2 = x_2\right) - \mathsf{E}\left(Y|X_1 = x_1, X_2 = x_2'\right) \\ &\quad \neq \mathsf{E}\left(Y|X_1 = x_1', X_2 = x_2\right) - \mathsf{E}\left(Y|X_1 = x_1', X_2 = x_2'\right)\end{aligned}$$

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

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

[16] This can be diagnosted; 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.

implies[17]

$$\mathsf{E}\left(Y|X_1 = L, X_2 = x_2\right) = (\beta_0 + \beta_{0L}) + (\beta_1 + \beta_{1L}x_2)$$

for $l = 2, \ldots, L$ and for $L = 1$,

$$\mathsf{E}\left(Y|X_1 = 1, X_2 = x_2\right) = \beta_0 + \beta_1 x_2$$

i.e. β_{0L} is the modification to the intercept β_0 when $X_1 = L$, β_{1L} is the modification to slope β_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 $X_3 = x_3'$.[18]

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

$$1 + X_1 + X_2 + X_1.X_2 \tag{1.2}$$

is "legitimate", whereas

$$1 + X_1 + X_1.X_2 \tag{1.3}$$

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 ommitted variable on the two-way interaction.

Example 1.3

For the two factor predictors, model (1.3) might imply

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

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

$$\mathsf{E}\,(Y|X = x) = \beta_0 + \beta_2 x^2. \tag{1.4}$$

If we location-scale transform x to z, *i.e.* set $x = \mu + \sigma z$. [19] The model becomes

$$\begin{aligned} \mathsf{E}\,(Y|Z = z) &= \beta_0 + \beta_2(\mu + \sigma z)^2 \\ &= \beta_0 + \beta_2(\mu^2 + 2\mu\sigma z + \sigma^2 z_2) \\ &= \beta_0' + \beta_1' z + \beta_2' z^2 \end{aligned}$$

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 $\widehat{\boldsymbol{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ allows us to form the vector of residuals

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ &= (\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top)\boldsymbol{Y} \\ &= (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y} \end{aligned}$$

with observed values

$$\boldsymbol{r} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y}.$$

[19] Related to the γ_{jk} coefficient.

[20] In the case of covariates, putting ineractions, but ommiting 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 Farenheit to Celcius. See Nelder and McCullagh on this for more extensive discussion

As we have seen, by construction

$$\begin{aligned}\widehat{\boldsymbol{y}}^\top \boldsymbol{r} &= \widehat{\boldsymbol{y}}^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) \\ &= \widehat{\boldsymbol{y}}^\top (\boldsymbol{Y} - \widehat{\boldsymbol{y}}) = 0\end{aligned}$$

and

$$\underset{\sim}{\boldsymbol{x}}_j^\top \boldsymbol{r} = \underset{\sim}{\boldsymbol{x}}_j^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}, \qquad j = 1, \ldots, p$$

However, also, the model assumption that the $\boldsymbol{\varepsilon}$ components have constant variance and are uncorrelated can be check by inspection of the observed residuals. Patterns in the residual plots of $\underset{\sim}{\boldsymbol{x}}_j$ versus $\boldsymbol{r}$ indicate violation of the zero mean assumption made for $\boldsymbol{\varepsilon}$. [21]

The residuals are constructed as

$$\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$$

which imply

$$\begin{aligned}\mathsf{Var}\,(\boldsymbol{R}) &= (\boldsymbol{I}_n - \boldsymbol{H})\mathsf{Var}\,(\boldsymbol{Y})\,(\boldsymbol{I}_n - \boldsymbol{H})^\top \\ &= \sigma^2(\boldsymbol{I}_n - \boldsymbol{H});\end{aligned}$$

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

[21] $\boldsymbol{r}$ 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 $\mathsf{E}(Y|X=x)$ 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 (θ, ϕ) 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 ϕ is **known**, this model is referred to as the **exponential family**, whereas if ϕ is **unknown**, it is referred to as the **exponential-dispersion family**.[22]

Note

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

θ is the **canonical parameter**, ϕ is the **dispersion parameter**.

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

$$\mathsf{E}_{f_Y}\left(\dot{\ell}_\theta(\theta,\phi;Y)\right) = 0$$

where

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

and

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

where

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

[22] We are restricting our attention with two unknown parameters models, namely the case with θ unknown and ϕ known (one unknown parameter distribution) and the case where θ, ϕ are unknown. This covers many one-dimensional distributions of interest.

Here, θ is 1-dimensional, ℓ is suitably differentiable with respect to θ and all expectations are taken with respect to the "true" data generating mechanism (DGM), all expectations finite.

In the exponential family, we have

$$\begin{aligned}\mathsf{E}(Y) &= \dot{b}(\theta) = \frac{\partial}{\partial\theta}(b(\theta)) = \mu, \text{ say.}\\ \mathsf{Var}(Y) &= a(\phi)\ddot{b}(\theta) = a(\phi)\frac{\partial^2}{\partial\theta^2}b(\theta).\end{aligned}$$

This implies that

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

i.e. we have an explicit mean-variance relationship.[23]

Example 2.1 (Poisson model)
Let $Y \sim \mathcal{P}(\lambda)$, a Poisson random variable. We write

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

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 $\mathsf{E}(Y) = e^\theta = \mu, \mathsf{Var}(Y) = e^\theta = \mu$ and $V(\mu) = \mu$.

Example 2.2 (Gamma model)
Let $X \sim \mathcal{G}(\alpha,\beta)$, a Gamma random variable with density

$$\begin{aligned}f_Y(y;\alpha,\beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\beta y}\mathbf{1}_{y>0}\\ &= \exp\{(\alpha-1)\log(y) - \beta y + \alpha\log(\beta) - \log(\Gamma(\alpha))\}\end{aligned}$$

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

[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

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

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

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

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

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

Set

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

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

$$\begin{aligned} f_Y(y,\pi) &= \binom{m}{y}\pi^y(1-\pi)^{m-y}\mathbf{1}_{0\le y\le m} \\ &= \exp\left\{y\log\left(\frac{\pi}{1-\pi}\right) + m\log(1-\pi) + \log\left[\binom{m}{y}\right]\right\} \end{aligned}$$

In this form, $\mathsf{E}\,(Y) = m\pi$ and $\mathsf{Var}\,(Y) = m\pi(1-\pi)$. This form is less appealing as $\mathsf{E}\,(Y)$ depends on m. So consider the data transformation from $y \mapsto y/m$. For the "new" data, we

have

$$f_Y(y,\pi) = \exp\left\{my\log\left(\frac{\pi}{1-\pi}\right) + m\log(1-\pi) + \log\left[\binom{m}{my}\right]\right\}\mathbf{1}_{y\in\left\{0,\frac{1}{m},\frac{2}{m},\ldots,\frac{m}{m}\right\}}$$
$$= \exp\left\{\frac{y\log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)}{1/m} + \log\left[\binom{m}{my}\right]\right\}\mathbf{1}_{y\in\left\{0,\frac{1}{m},\frac{2}{m},\ldots,\frac{m}{m}\right\}}$$

Set

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

and

$$\mathsf{E}(Y) = \pi = \frac{e^\theta}{1+e^\theta} = \mu$$
$$\mathsf{Var}(Y) = \frac{\pi(1-\pi)}{m} = \frac{\mu(1-\mu)}{m} = \phi V(\mu)$$

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

$$\mu(\boldsymbol{x}) = \mathsf{E}(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x})$$

and suppose that $\mu(\boldsymbol{x})$ depends on a finite dimensional parameter $\boldsymbol{\beta}$, in particular $\mu(\boldsymbol{x};\boldsymbol{\beta}) \equiv \mu(\boldsymbol{x}\boldsymbol{\beta}) = \mu(\eta)$.[24] where

$$\eta = \boldsymbol{x}\boldsymbol{\beta} = \sum_{j=1}^{D} x_j\beta_j$$

$\boldsymbol{x}$ is $(1\times p)$ vector and $\boldsymbol{\beta}$ a $(p\times 1)$ vector. η is termed the **linear predictor** (linear in the parameters $\beta_1,\ldots,\beta_p$).

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

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

[24] Thus depends on the inner product.

In this case, if $\beta = \boldsymbol{x}\boldsymbol{\beta}$ takes values on $\mathbb{R}$ (*i.e.* with $\boldsymbol{x}, \boldsymbol{\beta}$ unconstrained in general), the transformation

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

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 μ and η 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 Φ denotes the inverse standard normal CDF.[25] In this case

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

2. the **complementary log-log** , where

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

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 one Nelder 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 sensibility 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

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

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,

$$\begin{aligned} f_Y(y_i, \theta_i, \phi) &= \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(\phi, y_i)\right) \\ &= \exp\left(\frac{y_i \boldsymbol{x}_i\boldsymbol{\beta} - b(\boldsymbol{x}_i\boldsymbol{\beta})}{a(\phi)} + c(\phi, y_i)\right) \end{aligned}$$

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

$$\mathcal{L}_n(\boldsymbol{\beta}, \phi, \boldsymbol{y}, \boldsymbol{x}) = \exp\left(\frac{\sum_{i=1}^n y_i \boldsymbol{x}_i\boldsymbol{\beta} - \sum_{i=1}^n b(\boldsymbol{x}_i\boldsymbol{\beta})}{a(\phi)} + \sum_{i=1}^n c(y_i, \phi)\right)$$

[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 β_j, for $j = 1, \ldots, p$.[29]

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

$$\begin{aligned} \mathsf{E}(Y) &= \dot{b}(\theta) = \mu \\ \mathsf{Var}(Y) &= a(\phi)\ddot{b}(\theta) = a(\phi)V(\mu) \end{aligned}$$

with linear predictor

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

We model $\mu = \mathsf{E}(Y_i | \boldsymbol{X}_i = \boldsymbol{x}_i)$ 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

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i; \boldsymbol{\theta}) \equiv \mathcal{L}_n(\boldsymbol{\theta})$$

The log-likelihood is

$$\ell_n(\boldsymbol{\theta}) = \log \mathcal{L}_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f_{Y_i}(y_i; \boldsymbol{\theta})$$

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

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

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

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

[29] Nowadays, the dimension reduction is less important as fourty years ago, unless sample size are very large

generally a $(p \times 1)$ system. Write $\dot{\ell}_n(\boldsymbol{\theta})$ for $\partial \ell_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$; the system of equations to be solved has components

$$\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \theta_j} = 0, \qquad \text{for } j = 1, \ldots, p$$

In general, $\widehat{\boldsymbol{\theta}}_n$ is such that $\dot{\ell}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}_p$ must be found numerically.[30]

Note

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

$$\ell_n(\boldsymbol{\theta}) = \ell_n(\widehat{\boldsymbol{\theta}}_n) + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n)^\top \dot{\ell}_n(\widehat{\boldsymbol{\theta}}_n) + \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n)^\top \ddot{\ell}_n(\widehat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n) + \text{remainder}$$

i.e.

$$\ell_n(\boldsymbol{\theta}) \simeq \ell_n(\widehat{\boldsymbol{\theta}}_n) + (\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n)^\top \dot{\ell}_n(\widehat{\boldsymbol{\theta}}_n) + \frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n)^\top \ddot{\ell}_n(\widehat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n)$$

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

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}, \qquad \text{for } j, k = 1, \ldots p.$$

But $\dot{\ell}_n(\widehat{\boldsymbol{\theta}}_n) = 0$, so on rearrangement we have that

$$\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\boldsymbol{\theta}) \simeq -\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n)^\top \ddot{\ell}_n(\widehat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n)$$

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

$$\exp\Big(\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\boldsymbol{\theta})\Big) = \frac{\mathcal{L}_n(\widehat{\boldsymbol{\theta}}_n)}{\mathcal{L}_n(\boldsymbol{\theta})} = \exp\left(\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n)^\top \ddot{\ell}_n(\widehat{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n)\right)$$

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

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

$$\widehat{\pi}_j = \frac{y_j}{m_j}, \qquad \text{for } j = 1, \ldots, p$$

We may write this as a GLM, by considering a factor predictor $\boldsymbol{X}$ taking p levels. In the

[30] Recall that maximum likelihood estimator has lowest variance asymptotically among the class of unbiased parametric estimators. In 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

$$\beta_j = \eta_j = \log\left(\frac{\pi_j}{1-\pi_j}\right)$$

so that

$$\pi_j = \frac{e^{\beta_j}}{1+e^{\beta_j}} \equiv \text{expit}(\beta_j),$$

for $j = 1, \ldots, p$. By invariance of ML estimation,[31]

$$\widehat{\beta}_j = \log\left(\frac{\widehat{\pi}_j}{1-\widehat{\pi}_j}\right) = \log\left(\frac{y_j}{m_j - y_j}\right)$$

We however usually work with constrasts; consider the reparametrization where

$$\eta_j = \alpha_0 + \delta_j,$$

for $j = 1, \ldots, p$. with the identifiability constraint $\delta_1 = 0$. Under the canonical link, we have

$$\pi_j = \frac{\exp(\alpha_0 + \delta_j)}{1 + \exp(\alpha_0 + \delta_j)}$$

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

$$\widehat{\delta}_j = \log\left(\frac{\widehat{\pi}_j/(1-\widehat{\pi}_j)}{\widehat{\pi}_1/(1-\widehat{\pi}_1)}\right)$$

for $j = 2, \ldots, p$ and $\widehat{\alpha}_0 = \text{logit}(\widehat{\pi}_1)$.

Note

$$\begin{aligned} \delta_j &= \log\left(\frac{\pi_j/(1-\pi_j)}{\pi_1/(1-\pi_1)}\right) \\ \Rightarrow \quad e^{\delta_j} &= \frac{\pi_j/(1-\pi_j)}{\pi_1/(1-\pi_1)} \end{aligned}$$

where $\pi_j/(1-\pi_j)$ 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.

[31] We will see later that only in the saturated model can we solve for the maximum likelihood parameter in this way.

In the new parametrization, we have

$$\begin{aligned}\mathcal{L}_n(\alpha_0, \delta_2, \ldots, \delta_p) &= \prod_{i=1}^{p} \binom{m_j}{y_j} \left(\frac{e^{\alpha_0+\delta_j}}{1+e^{\alpha_0+\delta_j}}\right)^{y_j} \left(\frac{1}{1+e^{\alpha_0+\delta_j}}\right)^{m_j-y_j} \\ &= \prod_{i=1}^{n} \binom{m_j}{y_j} \exp\left(y_j(\alpha_0+\delta_j)\right) \left(\frac{1}{1+e^{\alpha_0+\delta_j}}\right)^{m_j}\end{aligned}$$

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

$$\begin{aligned}\frac{\partial \ell_n}{\partial \alpha_0} &= \sum_{j=1}^{p} \left(y_j - m_j \frac{e^{\alpha_0+\delta_j}}{1+e^{\alpha_0+\delta_j}}\right) \\ &= \sum_{j=1}^{p} (y_j - m_j \pi_j) \\ &= \sum_{j=1}^{p} (y_j - \mu_j)\end{aligned}$$

we have thus

$$\begin{aligned}\frac{\partial \ell_n}{\partial \delta_j} &= y_j - m_j \text{expit}(\alpha_0+\delta_j) \\ &= y_j - m_j \pi_j \\ &= y_j - \mu_j\end{aligned}$$

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

$$\pi_i = \text{expit}(\beta_0 + \beta_1 x_i)$$

for $i = 1, \ldots, n$, *i.e.*

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

and (β_0, β_1) are the parameters of the model to be estimated. The log-likelihood is

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

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

$$\begin{aligned}
\frac{\partial \ell_n}{\partial \beta_0} &= \sum_{i=1}^{n} \left(y_i - m_i \text{expit}(\beta_0 + \beta_1 x_i)\right) \\
&= \sum_{i=1}^{n} (y_i - m_i \pi_i) \\
&= \sum_{i=1}^{n} (y_i - \mu_i)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial \ell_n}{\partial \beta_1} &= \sum_{i=1}^{n} \left(y_i - m_i \text{expit}(\beta_0 + \beta_1 x_i)\right) x_i \\
&= \sum_{i=1}^{n} x_i (y_i - m_i \pi_i) \\
&= \sum_{i=1}^{n} x_i (y_i - \mu_i)
\end{aligned}$$

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

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

[32] This is prioritizing the orthogonality of the residuals with the covariates, generalized in the GLM setting.

which we now rewrite as

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

i.e. $\phi = 1$ and $w_i = m_i$.

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

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

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

$$\begin{aligned}\mathsf{E}\left(Y_i\right) &= \dot{b}(\theta_i) = \mu_i\\ \mathsf{Var}\left(Y_i\right) &= a_i(\phi)\ddot{b}(\theta_i) = \frac{\phi}{w_i}V(\mu_i)\end{aligned}$$

Thus, the score equations can be rewritten

$$\sum_{i=1}^{n}\frac{w_i}{\phi}\frac{(y_i-\mu_i)x_{ij}}{\dot{g}(\mu_i)V(\mu_i)}$$

assuming a link function $g(t)$ where [33]

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

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

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

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

$$\frac{\partial\ell(y;\theta,\phi)}{\partial\beta_j} = \frac{\partial\ell(y;\theta,\phi)}{\partial\theta}\frac{\mathrm{d}\theta}{\mathrm{d}\mu}\frac{\partial\mu}{\partial\eta}\frac{\partial\eta}{\partial\beta_j}$$

and from $\dot{b}(\theta) = \mu$ and $\ddot{b}(\theta) = V$, we derive $\mathrm{d}\mu/\,\mathrm{d}\theta = V(\mu)$ and from $\eta = \sum x_j\beta_j, \partial\eta/\partial\beta_j = x_j$ which leads to the above result

[33] In order to compute, $\frac{\partial\theta_i}{\partial\beta_j}$, 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

$$g(Y_i) \simeq g(\mu_i) + (Y_i - \mu_i)\dot{g}(\mu_i)$$

Let

$$\begin{aligned} Z_i &= \eta_i + (Y_i - \mu_i)\dot{g}(\mu_i) \\ &= \boldsymbol{x}_i\boldsymbol{\beta} + (Y_i - \mu_i)\dot{g}(\mu_i). \end{aligned}$$

We have that [34]

$$\begin{aligned} \mathsf{E}(Z_i) &= \eta_i = \boldsymbol{x}_i\boldsymbol{\beta} \\ \mathsf{Var}(Z_i) &= \mathsf{Var}(Y_i)\left(\dot{g}(\mu_i)\right)^2 \\ &= \left(\dot{g}(\mu_i)\right)^2 \frac{\phi V(\mu_i)}{w_i} \equiv v_i \end{aligned}$$

Thus

$$Z_i = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i$$

for $i = 1, \ldots, n$ where $\mathsf{Var}(\varepsilon_i) = v_i$ and we may estimate $\boldsymbol{\beta}$ using least-squares using this linear model formulation via

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \boldsymbol{W}\mathbf{X}\right)^{-1}\mathbf{X}^\top \boldsymbol{W}\boldsymbol{z}$$

if we treat the Z_i's as known quantities, where

$$\boldsymbol{W} = \text{diag}\left(v_1^{-1}, \ldots, v_n^{-1}\right)$$

a weighted-least square problem. So an iterative procedure can be constructed as follows:

Algorithm 2.1 (Iteratively reweighted least-squares)

1. Initialize: choose $\boldsymbol{\beta}^{(0)}$ (or $\widehat{\mu}_i^{(0)}$) to give $\widehat{\eta}_i^{(0)} = g\left(\widehat{\mu}_i^{(0)}\right)$.
2. Compute

$$\widehat{Z}_i^{(1)} = \widehat{\eta}_i^{(0)} + \left(y_i - \widehat{\mu}_i^{(0)}\right)\dot{g}\left(\widehat{\mu}_i^{(0)}\right)$$

3. Form

$$\widehat{W}_i^{(1)} = \frac{w_i}{\left(\dot{g}\left(\widehat{\mu}_i^{(0)}\right)\right)^2 V\left(\widehat{\mu}_i^{(0)}\right)}$$

[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

$$\boldsymbol{W}^{(1)} = \text{diag}\left(\widehat{w}_1^{(1)}, \ldots, \widehat{w}_n^{(1)}\right)$$

where w_i are specified by the choice of the model.

4. Compute

$$\widehat{\boldsymbol{\beta}}^{(1)} = \left(\mathbf{X}^\top \boldsymbol{W}^{(1)} \mathbf{X}\right)^{-1} \mathbf{X}^\top \boldsymbol{W}^{(1)} \widehat{Z}^{(1)}$$

5. Return to **2.** with updated

$$\widehat{\eta}_i^{(1)} = \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}^{(1)}, \quad \widehat{\mu}_i^{(1)}, \quad \widehat{Z}_i^{(1)}$$

and proceed through **3.-5.**

This procedure produces a sequence of estimates $\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\boldsymbol{\beta}}^{(2)}, \ldots, \widehat{\boldsymbol{\beta}}^{(t)}$ which will (in most cases) [35] converge to a fixed point.

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

This algorithm is known as **iteratively reweighted least-squares** and under regularity conditions and as $n \to \infty$,

$$\widehat{\boldsymbol{\beta}} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{\beta}, \mathcal{I}(\boldsymbol{\beta})^{-1}\right)$$

where $\widehat{\boldsymbol{\beta}}$ is the fixed point of the algorithm (*i.e.* the value of $\widehat{\boldsymbol{\beta}}^{(t)}$ at termination). [36]

Here

$$\mathcal{I}(\boldsymbol{\beta}) = \frac{1}{\phi} \mathbf{X}^\top \boldsymbol{W} \mathbf{X}, \qquad [\boldsymbol{W}]_{ii} = \frac{w_i}{V(\mu_i)\left(\dot{g}(\mu_i)\right)^2}$$

where μ_i is estimated by $\widehat{\mu}_i = \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}$.

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

2.5 Quadratic algorithms

Consider the score function $\dot{\ell}_n(\boldsymbol{\theta})$ arising in a parametric statistical model, and a first-order approximation to $\dot{\ell}_n(\boldsymbol{\theta})$ about some arbitrary point $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, a $(p \times 1)$ vector. Thus

$$\dot{\ell}_n(\boldsymbol{\theta}) \simeq \dot{\ell}_n(\boldsymbol{\theta}_0) + \ddot{\ell}_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

[35] In which case either no solution to the score equations or poor starting values choice.

[36] 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 $\ddot{\ell}_n(\boldsymbol{\theta}_0)$ is a $(p \times p)$ matrix of mixed partials. Rearranging gives

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \left(\ddot{\ell}_n(\boldsymbol{\theta}_0)\right)^{-1} \left(\dot{\ell}_n(\boldsymbol{\theta}) - \dot{\ell}_n(\boldsymbol{\theta}_0)\right)$$

Quadratic approximation

We have

$$\boldsymbol{\theta} \simeq \boldsymbol{\theta}_0 + (\ddot{\ell}_n(\boldsymbol{\theta}_0))^{-1} \left(\dot{\ell}_n(\boldsymbol{\theta}) - \dot{\ell}_n(\boldsymbol{\theta}_0)\right)$$

Replacing $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}_n, \dot{\ell}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}_p$, a $(p \times 1)$ vector so

$$\widehat{\boldsymbol{\theta}}_n \simeq \boldsymbol{\theta}_0 - (\ddot{\ell}_n(\boldsymbol{\theta}_0))^{-1} \dot{\ell}_n(\boldsymbol{\theta}_0)$$

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

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

$$\widehat{\boldsymbol{\theta}}_m = \boldsymbol{\theta}_{m-1} - (\ddot{\ell}_n(\boldsymbol{\theta}_{m-1}))^{-1} \dot{\ell}_n(\boldsymbol{\theta}_{m-1})$$

- track $\{\widehat{\boldsymbol{\theta}}_m\}$ until convergence *i.e.*

$$\left|\widehat{\boldsymbol{\theta}}_m - \widehat{\boldsymbol{\theta}}_{m-1}\right| < \varepsilon_1 \quad \text{or} \quad \left|\dot{\ell}_n(\widehat{\boldsymbol{\theta}}_m)\right| < \varepsilon_2$$

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 $-\ddot{\ell}_n(\boldsymbol{\theta})$ by its expectation

$$\mathcal{I}_n(\boldsymbol{\theta}) = -\mathsf{E}\left(\ddot{\ell}_n(\boldsymbol{\theta})\right) = \mathsf{E}\left(\dot{\ell}_n(\boldsymbol{\theta})\dot{\ell}_n(\boldsymbol{\theta})^\top\right)$$

where the second equality holds in the Exponential family. Recall

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

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

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

The recursion based on

$$\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}_{n-1} + (\mathcal{I}_n(\widehat{\boldsymbol{\theta}}_{n-1}))^{-1}\dot{\ell}_n(\widehat{\boldsymbol{\theta}}_{n-1})$$

is termed **Fisher scoring**.[37] Often, in this algorithm, $\mathcal{I}_n(\widehat{\boldsymbol{\theta}}_n)$ is replaced by estimates of the unit information, the so-called observed information[38]

$$\widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}}_n) = \begin{cases} -\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \left.\dfrac{\partial^2 \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \left(\dfrac{\partial \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\dfrac{\partial \ell_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^\top \Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_n} \end{cases}$$

Example 2.9 (Poisson regression)

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

$$\begin{aligned} \boldsymbol{x}_i &= (x_{i1}, \ldots, x_{ip})^\top && (1 \times p) \\ \boldsymbol{\beta} &= (\beta_1, \ldots, \beta_p) && (p \times 1) \end{aligned}$$

We have

$$\begin{aligned} \ell_n(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(-\mu_i + y_i \log(\mu_i)) + \text{ const} \\ \Rightarrow \dot{\ell}_n(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\left(-\frac{\partial \mu_i}{\partial\boldsymbol{\beta}} + y_i\frac{\partial \log(\mu_i)}{\partial\boldsymbol{\beta}}\right) \end{aligned}$$

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

$$\begin{aligned} \frac{\partial \mu_i}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j}\exp(\boldsymbol{x}_i\boldsymbol{\beta}) \\ &= x_{ij}\exp(\boldsymbol{x}_i\boldsymbol{\beta}) \\ &= x_{ij}\mu_i \end{aligned}$$

[37] In `R`, `glm` uses the Fisher scoring algorithm, which is very efficient.

[38] 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 $\boldsymbol{x}_i\boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij}\beta_j$. Now, we also need

$$\begin{aligned}\frac{\partial \ell_n(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n}\left(-x_{ij}\mu_i + \frac{y_i}{\mu_i}x_{ij}\mu_i\right)\\ &= \sum_{i=1}^{n} x_{ij}(y_i - \mu_i)\end{aligned}$$

for $j = 1, \ldots, p$ and the matrix of mixed partials is given by

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

for $j = 1, \ldots, p$ and $k = 1, \ldots, p$.

Thus Newton's method uses the recursion for $m = 1, 2, \ldots$

$$\widehat{\boldsymbol{\beta}}_m = \widehat{\boldsymbol{\beta}}_{m-1} - \left(D(\widehat{\boldsymbol{\beta}}_{m-1})\right)^{-1} \dot{\ell}_n(\widehat{\boldsymbol{\beta}}_{m-1})$$

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

$$-\sum_{i=1}^{n} x_{ij}x_{ik}\exp(\boldsymbol{x}_i\boldsymbol{\beta})$$

Note that in this case

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

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

Example 2.10 (Binomial data)

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

$$\begin{aligned}\ell_n(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \frac{y_i\boldsymbol{x}_i\boldsymbol{\beta} - \log(1 + \exp(\boldsymbol{x}_i\boldsymbol{\beta}))}{1/m_i} + \text{ const}\\ \Rightarrow \quad \frac{\partial \ell_n(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n} \frac{x_{ij}(y_i - \mu_i)}{1/m_i}\end{aligned}$$

However, if $\boldsymbol{x}_i\boldsymbol{\beta} = \Phi^{-1}(\pi_i)$, the probit link, the likelihood derivatives are more difficult to

compute. In that case,

$$\ell_n(\boldsymbol{\beta}) = \frac{\sum_{i=1}^n y_i \log\left(\frac{\Phi(\boldsymbol{x}_i\boldsymbol{\beta})}{1-\Phi(\boldsymbol{x}_i\boldsymbol{\beta})}\right) + \log(1-\Phi(\boldsymbol{x}_i\boldsymbol{\beta}))}{1/m_i} + \text{ const}$$

Now $\frac{\partial \ell_n(\boldsymbol{\beta})}{\partial \beta_j}$ depends on both Φ and ϕ, 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** : $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathsf{p}} \boldsymbol{\theta}_0$, for $\boldsymbol{\theta}_0$ is the "true" value.
- **asymptotic normality**

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}_p\left(\mathbf{0}_p, (\mathcal{I}(\boldsymbol{\theta}_0))^{-1}\right)$$

Note

We have

$$\begin{aligned}\frac{1}{\sqrt{n}}\dot{\ell}_n(\boldsymbol{\theta}_0) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \dot{\ell}_i(\boldsymbol{\theta}_0)\right) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \frac{\dot{f}_Y(Y_i;\boldsymbol{\theta}_0)}{f_Y(Y_i;\boldsymbol{\theta}_0)}\right) \xrightarrow{d} \mathcal{N}_p(\mathbf{0}_p, \mathcal{I}(\boldsymbol{\theta}_0))\end{aligned}$$

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

$$\widehat{\boldsymbol{\theta}}_n \dot{\sim} \mathcal{N}_p\left(\boldsymbol{\theta}_0, \frac{1}{n}\left(\widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}}_n)\right)^{-1}\right)$$

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

$$\widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}}_n) = \begin{cases} \quad \frac{1}{n}\sum_{i=1}^n \dot{\ell}(y_i;\widehat{\boldsymbol{\theta}}_n)\dot{\ell}(y_i;\widehat{\boldsymbol{\theta}}_n)^\top \\ - \quad \frac{1}{n}\sum_{i=1}^n \Psi(y_i,\widehat{\boldsymbol{\theta}}_n) \end{cases}$$

both $(p \times p)$ matrices, where

$$\Psi(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}_n) = \ddot{\ell}(\boldsymbol{y}; \widehat{\boldsymbol{\theta}}_n) = \left. \frac{\partial^2 \ell(\boldsymbol{y}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right|_{\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_n}.$$

The diagonal elements of $\left(\widehat{\mathcal{I}}(\widehat{\boldsymbol{\theta}}_n)\right)^{-1}$ yield the estimated (squared) standard errors for the elements of $\widehat{\boldsymbol{\theta}}_n$, and we may perform hypothesis tests concerning $\boldsymbol{\theta}$ using the test statistics

$$\frac{\widehat{\boldsymbol{\theta}}_{nj}}{\widehat{\text{se}}(\widehat{\boldsymbol{\theta}}_{nj})} \dot{\sim} \mathcal{N}(0,1) \quad \text{if } \boldsymbol{\theta}_{0j} = 0$$

–such a statistic is termed a **Wald statistic** . To extend to multivariate tests, we rely on the distributional result that if $\boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, then

$$\boldsymbol{Y}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{Y} \sim \chi^2_p.$$

–we may apply this result to (sub)vector(s) (of) $\widehat{\boldsymbol{\theta}}_n$ 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

$$\begin{aligned} H_0 : &\quad \boldsymbol{\theta} \in \Theta_0 \\ H_1 : &\quad \boldsymbol{\theta} \in \Theta_1 \end{aligned}$$

we use the statistic

$$\lambda(\boldsymbol{y}) = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} \mathcal{L}_n(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta_0 \cup \Theta_1} \mathcal{L}_n(\boldsymbol{\theta})} = \frac{\mathcal{L}_n(\widehat{\boldsymbol{\theta}}_{n0})}{\mathcal{L}_n(\widehat{\boldsymbol{\theta}}_n)}$$

Clearly $\lambda(y) \leq 1$; if $\lambda(\boldsymbol{y}) \leq k$, for some $k < 1$, we reject H_0 in favor of H_1. Two cases to consider

1. H_0 **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 $-2\log(\lambda(\boldsymbol{Y})) \xrightarrow{d} Q \sim \chi^2_p$ 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-$\mathcal{F}$ test

restricted version of H_1, then $-2\log(\lambda(\boldsymbol{Y})) \xrightarrow{d} Q \sim \chi^2_{k_2-k_1}$.[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 $\mu_i = \mathsf{E}\left(Y_i | \boldsymbol{X}_i = \boldsymbol{x}_i\right)$ yield different log-likelihood values. The simplest model is where $\widehat{\mu}_i = \widehat{\mu}_0$ (the **null model**) (independence on $\boldsymbol{x}$). In constrast to that, the most complex model where $\widehat{\mu}_i = y_i$, ($\widehat{\mu}_i$ is the "fitted value"). This one is a fit which matches the data exactly. For other models, we have

$$\widehat{\boldsymbol{\mu}} = \mu(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}}) \equiv g^{-1}(\boldsymbol{x}_i\widehat{\boldsymbol{\beta}}),$$

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

Let

$$\bar{\boldsymbol{\theta}} = \theta(\widehat{\mu}_0 \mathbf{1}_n) \qquad \widehat{\boldsymbol{\theta}} = \theta(\widehat{\boldsymbol{\mu}}) \qquad \widetilde{\boldsymbol{\theta}} = \theta(\boldsymbol{y})$$

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

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

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$. Therefore, we can compute $2(\ell_n(\widetilde{\boldsymbol{\theta}}) - \ell_n(\widehat{\boldsymbol{\theta}}))$. We have that

$$\begin{aligned} 2(\ell_n(\widetilde{\boldsymbol{\theta}}) - \ell_n(\widehat{\boldsymbol{\theta}})) &= 2\sum_{i=1}^{n} \frac{w_i}{\phi} \left[y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - (b(\widetilde{\theta}_i) - b(\widehat{\theta}_i)) \right] \\ &= \frac{D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})}{\phi} \end{aligned}$$

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

[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 $\mathcal{F}$-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,

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

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

$$\begin{aligned} 2\left(\ell_n(\widetilde{\boldsymbol{\theta}}) - \ell_n(\widehat{\boldsymbol{\theta}})\right) &= 2\sum_{i=1}^{n}\left(y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)\right) \\ &= D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) \end{aligned}$$

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

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

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

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

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

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

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

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

Note

If ϕ is unknown, a (consistent) estimator of ϕ is

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

Note

`R` terms the $D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})$ quantity the "residual deviance", $n - p$ is termed the "residual degrees

of freedom".[41]

Suppose ϕ is known, but two **nested** models are fitted. Recall

Definition 2.1 (Nested model)

A model M_A is nested inside model can be obtained from model M_B by the imposition of equality constraints at the interior of the parameter space.[42] 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

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

if M_A is an adequate simplification of M_B.

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

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

2.6.3. Pearson X^2 statistic

The statistic

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

is also a measure of goodness-of-fit or model adequacy.[43] Under the usual asymptotic arguments (namely Central Limit Theorem),

$$\frac{y_i - \mu_i}{\sqrt{\mathsf{Var}\,(\mu_i)}} \dot{\sim} \mathcal{N}(0, \phi)$$

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

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

[41] Analogous to the linear model, and since asymptotically the deviance is χ^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 ∞, then the asymptotic theory of LR breaks down.

[43] This statistic would work even for moment-based estimation and used for model adequacy assesment.

Note

To estimate ϕ, we may use $\widehat{\phi} = X^2/(n-p)$

2.7 Residuals

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

- **Pearson residual**

$$r_{p_i} = \frac{y_i - \widehat{\mu}_i}{\sqrt{\mathsf{Var}\,(\widehat{\mu}_i)}} \qquad \left(= \frac{\text{obs-fitted}}{\widehat{\text{se}}(\text{fitted})}\right)$$

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 $\dot{\ell}_n(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{x}_i^\top (y_i - \mu_i)$ a $(p \times 1)$ system as before. As $\dot{\ell}_n = \mathbf{0}_p$, we must have $\sum_{i=1}^n \boldsymbol{x}_i^\top (y_i - \widehat{\mu}_i) = 0$ *i.e.* we have $\underset{\sim}{\boldsymbol{x}}_j(\boldsymbol{y} - \widehat{\boldsymbol{\mu}}) = 0$ for $j = 1, \ldots, p$ *i.e.* a plot of $\underset{\sim}{\boldsymbol{x}}_j$ versus $\boldsymbol{y} - \widehat{\boldsymbol{\mu}}$ 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 $\boldsymbol{y} \to A(\boldsymbol{y})$ made to ensure that the distribution of $A(\boldsymbol{y})$ is approximately normal.[46] If $\boldsymbol{Y}$ follows and Exponential-Dispersion family distribution, then it can be shown that

$$A(t) = \int_{-\infty}^{t} \frac{\mathrm{d}\mu}{V^{\frac{1}{3}}(\mu)}$$

defines the appropriate transformation.[47]

Example 2.12

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

$$A(t) = \int_0^t \frac{\mathrm{d}\mu}{\mu^{\frac{1}{3}}} = \frac{3}{2} t^{\frac{2}{3}}$$

[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 $\widehat{\mu}_i$. 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 $Y \to \frac{3}{2}y^{\frac{2}{3}}$ and hence define the residual by

$$\frac{A(y_i) - A(\widehat{\mu}_i)}{\dot{A}(\widehat{\mu}_i)\sqrt{V(\widehat{\mu}_i)}}$$

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

$$\mathsf{Var}\,(Y) = V(\mu) \quad \Rightarrow \quad \mathsf{Var}\,(A(y)) \simeq \left(\dot{A}(\mu)\right)^2 V(\mu)$$

so that in the Poisson case,[48] we examine

$$r_{A_i} = \frac{\frac{3}{2}\left(y_i^{\frac{2}{3}} - \widehat{\mu}_i^{\frac{2}{3}}\right)}{\widehat{\mu}_i^{\frac{1}{6}}}.$$

◦ **Deviance residuals:** Recall that the deviance is defined via

$$\begin{aligned} 2(\ell_n(\widetilde{\boldsymbol{\theta}}) - \ell_n(\widehat{\boldsymbol{\theta}})) &= \frac{2}{\phi}\sum_{i=1}^{n} w_i \left(y_i(\widetilde{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_i) - (b(\widetilde{\boldsymbol{\theta}}_i) - b(\widehat{\boldsymbol{\theta}}_i))\right) \\ &= \frac{1}{\phi}\sum_{i=1}^{n} d_i \end{aligned}$$

where

$$d_i = 2w_i \left(y_i(\widetilde{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_i) - (b(\widetilde{\boldsymbol{\theta}}_i) - b(\widehat{\boldsymbol{\theta}}_i))\right)$$

The **deviance residual**[49] is defined by

$$r_{d_i} = \mathrm{sign}(y_i - \widehat{\mu}_i)\sqrt{d_i}$$

where

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

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

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 tymor 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 $Y_i \sim \mathcal{P}(\mu_i)$, with canonical link $\theta_i = \log(\mu_i) = \boldsymbol{x}_i\boldsymbol{\beta}$ 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) $Y \sim \mathcal{P}(\mu)$ implies $\frac{Y-\mu}{\sqrt{\mu}} \dot{\sim} \mathcal{N}(0,1)$ for large values of μ (*i.e.* large count rate) using the Central Limit theorem.

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

$$S \simeq h(\mu) + (Y-\mu)\dot{h}(\mu) + \frac{(Y-\mu)^2}{2}\ddot{h}(\mu)$$

If $h(t) = \sqrt{t}$,[50] then

$$S = Y^{\frac{1}{2}} \simeq \mu^{\frac{1}{2}} + \frac{(Y-\mu)}{2\mu^{\frac{1}{2}}}$$

yielding

$$\begin{aligned} \mathsf{E}(S) &\simeq \mu^{\frac{1}{2}} \\ \mathsf{Var}(S) &\simeq \frac{1}{4\mu}\mathsf{Var}(Y-\mu) = \frac{1}{4} \end{aligned}$$

and $\mathsf{Var}(S)$ does not depend on μ (for μ large). This is called a **variance stabilizing** transformation. In fact, McCullagh and Nelder go in detail on p. 196 and show

$$\begin{aligned} \mathsf{E}\left(Y^{\frac{1}{2}}\right) &\simeq \mu^{\frac{1}{2}}\left(1 - \frac{1}{8\mu}\right) \\ \mathsf{Var}\left(Y^{\frac{1}{2}}\right) &\simeq \mu^{\frac{1}{4}}\left(1 - \frac{3}{8\mu}\right) \end{aligned}$$

which work up to order $1/\mu$.

(c) Transform to symmetry: $h(t) = t^{\frac{2}{3}}$, an Anscombe transform, and

$$\begin{aligned} \mathsf{E}\left(Y^{\frac{2}{3}}\right) &\simeq \mu^{\frac{2}{3}}\left(1 - \frac{1}{9\mu}\right) \\ \mathsf{Var}\left(Y^{\frac{2}{3}}\right) &\simeq \frac{4\mu^{\frac{1}{3}}}{9}\left(1 + \frac{1}{6\mu}\right) \end{aligned}$$

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

$$D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n}\left(y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) - (y_i - \widehat{\mu}_i)\right)$$

[50] Transforming the original $\boldsymbol{y}$ and not the mean

If $g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$, then

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

where

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

Under the canonical log link,

$$\begin{aligned} \frac{\partial \mu_i}{\partial \beta_0} &= \frac{\partial}{\partial \beta_0} \left(\exp \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right) \right) \\ &= \exp \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right) \\ &= \mu_i \end{aligned}$$

which implies that

$$\frac{\partial \ell_n}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \mu_i = 0$$

to yield estimates. Therefore, at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$, we have

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

and therefore the deviance reduces to

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

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^{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 $\mathsf{E}(Y_i)$ which is not dependent on the predictors.

In the Poisson case, with log link, we have

$$\begin{aligned} \log(\mu_i) &= \log(E_i) + \eta_i \\ &= \log(E_i) + \boldsymbol{x}_i\boldsymbol{\beta} \end{aligned}$$

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 α_1, β_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.

$$\begin{aligned} M_0: &\quad 1 \\ M_1: &\quad 1 + (4-1) = 4 \\ M_2: &\quad 1 + (3-1) = 4 \\ M_3: &\quad 1 + (4-1) + (3-1) = 6 \end{aligned}$$

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 χ^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

$$\begin{aligned} \Delta\text{Deviance} &: 196.9 - 51.795 = 145.105 \\ \Delta\text{rdf} &: 6 - 3 = 3 \end{aligned}$$

To test H_0: M_2 is an adequate simplification of M_3, compare Δ 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

`tumor` + `site` + `tumor:site` or `tumor*site`

fits the data **exactly** as $\widehat{\mu}_{jk} = y_{jk}$ 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 `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 $X \sim \mathcal{B}(m, \pi)$ we can consider for X the Poisson approximation $X \dot{\sim} \mathcal{P}(\lambda)$ 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, π_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 `~offset()`. Since in this case `pop` is the population size, we need the offset to be on the log scale, so we get in this case `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 `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 $\Delta\text{Deviance}/\phi \dot{\sim} \chi^2_{\Delta\text{r.d.f}}$, *i.e.* if $\Delta\text{Deviance}/\phi$ is 'not too large', we do not reject te simpler model as an adequate simplification of the more complex model.

Note

If ϕ is unknown, it can be replaced by an estimated value $\widehat{\phi}$ in the analysis of deviance.[53] `R` uses the **Pearson-based** estimate

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

– other estimates may be used, and model-based (ML) estimates of ϕ may also be used. In the Analysis of Deviance, we compute $\widehat{\phi}$ under the **more complex** model.

3. Test the 'final' model for adequacy, looking at Deviance$/\phi \dot{\sim} \chi^2_{n-p}$ 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 $-2\ell_n(\widehat{\boldsymbol{\beta}}) + 2p$ and the BIC to be $-2\ell_n(\widehat{\boldsymbol{\beta}}) + p\log(n)$. 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 χ^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 $p(y^*|y) \approx p(y^*|\widehat{\boldsymbol{\beta}})$ 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 improvement would result. The assumption of the Poisson model of mean-variance equality relationship may not be valid. We could use the extension to the quasi-Poisson `quasipoisson`. Under this model, the parameter estimates stay the same (we are using the same link function, and as such estimates for μ_i yields identical estimates to `poisson`), but the standard errors change and as a result the P-values (standard errors are scaled by $\widehat{\phi}^{\frac{1}{2}}$). Note the dispersion parameter is 2.95, as opposed to 1 with the Poisson family. We still have $\mathsf{E}(Y_i) = \mu_i$ and $\mathsf{Var}(Y_i) = \phi\mu_i$, but where ϕ 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 improve 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 fctorial 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 χ^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: $A * C$). We should not entertain any simplification from the full factorial model $A * C = A + C + A.C$.

The analysis of deviance comparing one factor against the additive model yields strong significance of both factors. Only the "saturated" model $A * C$ 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 β's using invariance of the δ-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 accross 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.[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 Δ 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 Δ Deviance=10.378 and Δrdf=$(2-1)(3-1)=2$, 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] You would need to maximize the likelihood yourself.

drop other two-way interactions. For this, `update` could be used. See Exercise 2, Q1.[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 $\widehat{\phi} \approx 0.47$. 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,

$$\frac{D(\boldsymbol{y}; \widehat{\boldsymbol{\mu}})}{\phi} \dot{\sim} \chi^2_{n-p}$$

What if ϕ is unknown? In this case, we may replace ϕ by $\widehat{\phi}$ from the Pearson estimator, *i.e.*

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

Similarly, for the **model comparison** analysis of deviance test,

$$\frac{D_0 - D_1}{\phi} \dot{\sim} \chi^2_{p_1 - p_0}$$

wher 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 ϕ by $\widehat{\phi}$ to complete the test.

This follows as the estimation of $\widehat{\phi}$ 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

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

where $\widehat{\phi}$ computed from the "adequate model", which corresponds to the D quantity in the denominator above.[58] The two χ^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

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

[57] Note that in `R`, you can use the `step` function to do backward and forward selection based on deviance, AIC or BIC, with option `direction ="both"`.

[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 $\nu \to \infty, \chi^2_\nu/\nu \xrightarrow{\mathsf{P}} 1$, the above is therefore approximately distributed asymptotically to

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

as required.

Pearson X^2

For the Poisson model,

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

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

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

Asymptotically (in the μ_{jk}),

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

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

Here, $n = J \times K$. For the main effects only model, $p = (J-1) + (K-1) + 1$, *i.e.* $p = J + K - 1$.[59] We thus have $n - p = JK - (J + K - 1) = (J-1)(K-1)$ *i.e.* if the main model is adequate,

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

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

$$\log(\mu_{jk}) = \beta^{(A)}_j + \beta^{(B)}_k$$

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

[59] Corresponding to the main factors plus the baseline

Note

Recall that the deviance in models with an intercept takes the form

$$D(\boldsymbol{y}, \widehat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\widehat{\mu}_i}\right) = G^2$$

– this form always pertains when we deal exclusively with factor predictors.

For the change in deviance,

$$D_0(\boldsymbol{y}, \widehat{\mu}^{(0)}) - D_1(\boldsymbol{y}, \widehat{\mu}^{(1)}) = 2\sum_{i=1}^{n} y_i \log\left(\frac{\widehat{\mu}_i^{(1)}}{\widehat{\mu}_i^{(0)}}\right)$$

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

$$\log(\mu_{jkl}) = \beta_j^{(A)} + \beta_k^{(B)} + \beta + l^{(C)} + \gamma_{kl}^{(BC)}$$

– for fixed j the contribution to the expected response (on the log transformed scale) from factors (B, C) is **always**

$$\beta_k^{(B)} + \beta_l^{(C)} + \gamma_{kl}^{(BC)}$$

whatever value of j is chosen. We also have **conditional independence** of the form $A + B + C + A.B + B.C \equiv (A * B) + (B * C)$. 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 $A \perp C|B$ [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,

$$A \longrightarrow B \longrightarrow C$$
$$C \longrightarrow B \longrightarrow A$$

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

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

–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

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**:

$$\log(\mu_{jk}) = \log(\mu_{kj}) \qquad (i.e.\ \gamma_{jk} = \gamma_{kj})$$

In this model, the number of parameters is given by the cells on the diagonal and the lower triangular,

$$1 + (J-1) + \frac{J(J-1)}{2}$$

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 accross levels of B. For *e.g.*, if $J = 3$,

μ_{11}	μ_{12}	μ_{13}
μ_{21}	μ_{22}	μ_{23}
μ_{31}	μ_{32}	μ_{33}

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

$$\begin{aligned} Y_{j\bullet} = Y_{j1} + Y_{j2} + Y_{j3} &\sim \mathcal{P}(\mu_{j1} + \mu_{j2} + \mu_{j3}) \\ &\sim \mathcal{P}(\mu_{1j} + \mu_{2j} + \mu_{3j}) \end{aligned}$$

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

$$\begin{aligned} \beta_j^{(A)} &= \beta_k^{(B)} \quad \text{if } j = k \\ \gamma_{jk}^{(AB)} &= \gamma_{kj}^{(BA)} \quad \text{if } j \neq k \end{aligned}$$

◦ **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

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

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 $A + B$, [61] 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 $\beta_1^{(B)}, \ldots, \beta_K^{(B)}$

$$\widehat{\beta}_0 = \frac{1}{J} \sum_{j=1}^{J} \log\left(\frac{y_{j\bullet}}{\omega_\bullet^{(B)}}\right)$$

$$\widehat{\beta}_j^{(A)} = \log\left(\frac{y_{j\bullet}}{\omega_\bullet^{(B)}}\right) - \widehat{\beta}_0$$

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 β_0 as $\widehat{\beta}_1^{(B)}, \ldots, \widehat{\beta}_J^{(A)}, \widehat{\beta}_0$.
- 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 ϕ 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.* $\mathsf{Var}(Y|X) > \mathsf{E}(Y|X)$) or underdispersion (*i.e.* $\mathsf{Var}(Y|X) < \mathsf{E}(Y|X)$). 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 $\mathsf{P}(Z > 0) = 1$ for $\mathsf{E}(Y) > 0$. Using the iterated expectation and iterated variance formulas, we can easily verify that

$$\mathsf{E}(Y) = \mathsf{E}_Z\left(\mathsf{E}_{Y|Z}(Y|Z = z)\right) = \mu \mathsf{E}_Z(Z)$$

and

$$\begin{aligned}\mathsf{Var}(Y) &= \mathsf{E}_Z\left(\mathsf{Var}_{Y|Z}(Y|Z = z)\right) + \mathsf{Var}_Z\left(\mathsf{E}_{Y|Z}(Y|Z = z)\right) \\ &= \mu \mathsf{E}_Z(Z) + \mu^2 \mathsf{Var}_Z(Z).\end{aligned}$$

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

The choice of the Gamma mixing distribution is partly encouraged by the fact that this makes $Y \sim \mathcal{NB}(\mu, \theta_z)$, 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

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

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 $p \equiv \frac{\mu}{\mu + \theta_z}$.

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

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

where g is some link function, in the Poisson case for *e.g.* $\log(\mu_i) = \boldsymbol{x}_i \boldsymbol{\beta}$. 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 $\boldsymbol{\beta}$ and θ_z. For example

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

and if θ_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.[62]

Note

If θ_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-**

[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.

regular; the asymptotic distribution of the test statistic is no longer a simple χ^2_1.[63] It can be shown that the likelihood ratio statistic, when appropriately transformed

$$-2\left(\ell_n^{(\mathcal{P})}(\widehat{\boldsymbol{\beta}}) - \ell_n^{(\mathcal{NB})}(\widehat{\boldsymbol{\beta}}, \widehat{\varphi})\right) \xrightarrow{d} V$$

where V has a mixture distribution.

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

where $Q \sim \chi^2_1$, namely a mixture of a point mass at zero and χ^2_1 with equal probability. We have thus $\mathsf{P}(V > v) = \frac{1}{2}\int_v^\infty f_Q(t)dt$, 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

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

i.e.

$$\begin{aligned}\mathsf{P}(Y = 0) &= \pi + (1-\pi)e^{-\mu}\\ \mathsf{P}(Y = y) &= (1-\pi)\frac{e^{-\mu}\mu^y}{y!} \quad \text{for } y = 1, 2, \ldots\end{aligned}$$

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

[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).

Section 4
Models for Binomial and Multinomial Data

Binomial data arise as counts of "positive" responses when two possible outcomes may be observed in a sequence of independent trials. We assume

$$Y \sim \mathcal{B}(m, \pi)$$

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

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

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

$$\eta_i = \boldsymbol{x}_i\boldsymbol{\beta} = g(\pi)$$

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 $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iK})$ – (a K components vector) and write

$$\boldsymbol{Y}_i \sim \mathcal{M}(m_i, \pi_1, \ldots, \pi_K)$$

with m_i a fixed, known positive integer. In this model, the joint mass function for the vector random vector $\boldsymbol{Y}_i$ is

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

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

$$\begin{aligned} \mathsf{E}(Y_{ik}) &= m_i \pi_k, && \text{for } k = 1, \ldots, K \\ \mathsf{Var}(Y_{ik}) &= m_i \pi_k (1 - \pi_k) && \text{for } k = 1, \ldots, K \end{aligned}$$

Indeed, marginally $Y_{ik} \sim \mathcal{B}(m_i, \pi_k)$. 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

$$\mathsf{Cov}(Y_{ij}, Y_{ik}) = -m_i \pi_j \pi_k.$$

There is thus negative correlation between the variables. [64]

4.1 Binomial Model and Logistic Regression

Suppose $Y_i \sim \mathcal{B}(m_i, \pi_i)$, 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

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

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, $\mu_i = \mathsf{E}(Y_i) = \pi_i$ therefore the canonical link is given by $\theta_i = \text{logit}(\pi_i)$ 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 $\eta_i = \boldsymbol{x}_i \boldsymbol{\beta}$ yields the log likelihood

$$\begin{aligned} \ell_n(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(\frac{y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)}{1/m_i} + \text{ constant} \right) \\ &= \sum_{i=1}^{n} w_i \left(y_i \boldsymbol{x}_i \boldsymbol{\beta} - \log(1 + e^{\boldsymbol{x}_i \boldsymbol{\beta}}) \right) + \text{ constant} \end{aligned}$$

[64] Indeed, if Y_1 is large, since the variables sum to one, this entails Y_2 has to be small.

thus

$$\begin{aligned}\frac{\partial \ell_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} m_i \left(y_i \boldsymbol{x}_i^\top - \frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}} \boldsymbol{x}_i^\top \right) \\ &= \sum_{i=1}^{n} m_i (y_i - \pi_i) \boldsymbol{x}_i^\top\end{aligned}$$

a $(p \times 1)$ system usually. This requires solving numerically $\partial \ell_n(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = \mathbf{0}$, yielding $\widehat{\boldsymbol{\beta}}$.

Saturated model

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

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

Note

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

$$D(y; \widehat{\boldsymbol{\pi}}) = 2 \sum_{i=1}^{n} \left(y_i \log y_i + (1-y_i) \log(1-y_i) - y_i \text{logit}(\widehat{\pi}) - \log(1-\widehat{\pi}) \right)$$

Now $y_i = 0$ or $y_i = 1$, so

$$y_i \log(y_i) = (1-y_i) \log(1-y_i) = 0$$

Therefore

$$\begin{aligned}D(y; \widehat{\boldsymbol{\pi}}) &= -2 \sum_{i=1}^{n} y_i \boldsymbol{x}_i \widehat{\boldsymbol{\beta}} - 2 \sum_{i=1}^{n} \log(1-\widehat{\pi}_i) \\ &= -2 \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \boldsymbol{y} - 2 \sum_{i=1}^{n} \log(1-\widehat{\pi}_i).\end{aligned}$$

But

$$\frac{\partial \ell_n(\widehat{\boldsymbol{\beta}})}{\partial \widehat{\boldsymbol{\beta}}} = \sum_{i=1}^{n} (y_i - \pi_i) \boldsymbol{x}_i^\top = \mathbf{0}_p$$

which entails that $\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\top \widehat{\boldsymbol{\pi}}$ where $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, \ldots, \widehat{\pi}_n)^\top$.[65] Replacing this equality into

[65] This is a simple moment equation.

the deviance term, we have

$$D(y;\widehat{\boldsymbol{\pi}}) - 2\widehat{\boldsymbol{\eta}}^\top\widehat{\boldsymbol{\pi}} - 2\sum_{i=1}^{n}\log(1-\widehat{\pi}_i)$$

where $\widehat{\boldsymbol{\eta}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, therefore D is a function of $\widehat{\boldsymbol{\beta}}$ (that is, given $\widehat{\boldsymbol{\beta}}$, the quantity D is fixed). Now as $n \to \infty$, $\widehat{\boldsymbol{\beta}} \xrightarrow{\mathsf{p}} \boldsymbol{\beta}_0$, the true value, by usual consistency properties of MLE. Asymptotically, D is a function of $\boldsymbol{\beta}_0$.[66] That is, we don't have $D(y;\widehat{\boldsymbol{\pi}}) \dot{\sim} \chi^2_{n-p}$ for large n uniformly across $\boldsymbol{\beta}_0$ 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 $Y_i \sim \mathcal{B}(1,\pi)$, 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

$$-2n\left(\bar{y}\log(\bar{y}) + (1-\bar{y})\log(1-\bar{y})\right)$$

– asymptotics depend on the **true value** of π.

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 $\beta \xrightarrow{\mathsf{p}} \beta_0$, and the uncertainty corresponding to the inverse of the Fisher infomation, which depends on values of β usually. The deviance plot however is not χ^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 β is large; the asymptotic distribution must be derived for any particular value of β. 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 $Y_i \sim \mathcal{B}(1,\pi_i)$, with $M_1 : \text{logit}(\pi_i) = \beta_0 + \beta_1\boldsymbol{x}_i$ against $M_0 : \text{logit}(\pi_i) = \beta_0$ *i.e.* to test $H_0 : \beta_1 = 0$, we can use χ^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 $\boldsymbol{\beta}_0$, thus the asymptotics derived earlier do not hold

Figure 2: Deviation from true parameter β_{true} as a function of the parameter estimate

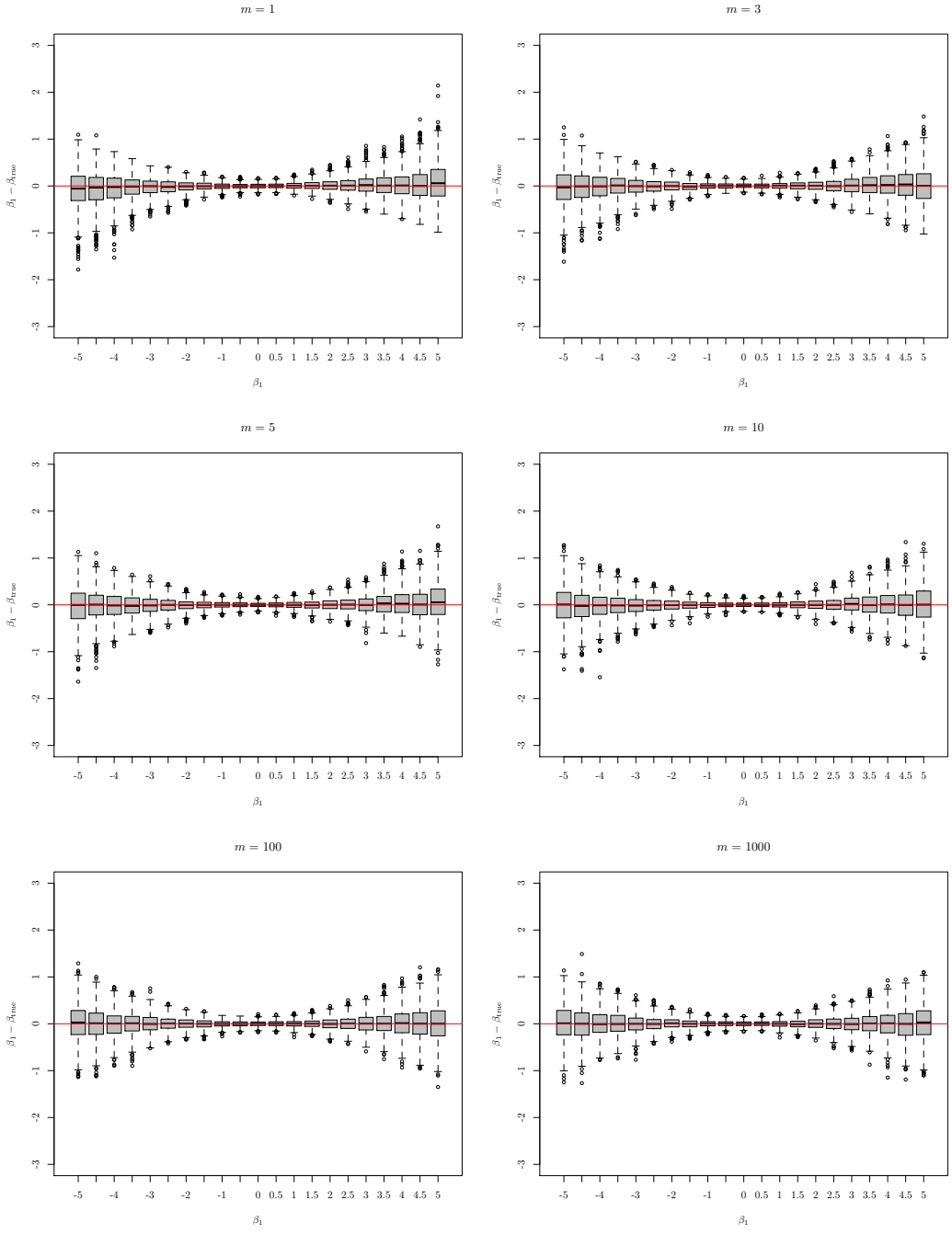


Figure 3: Deviance plot as function of number of observations and value of β_1

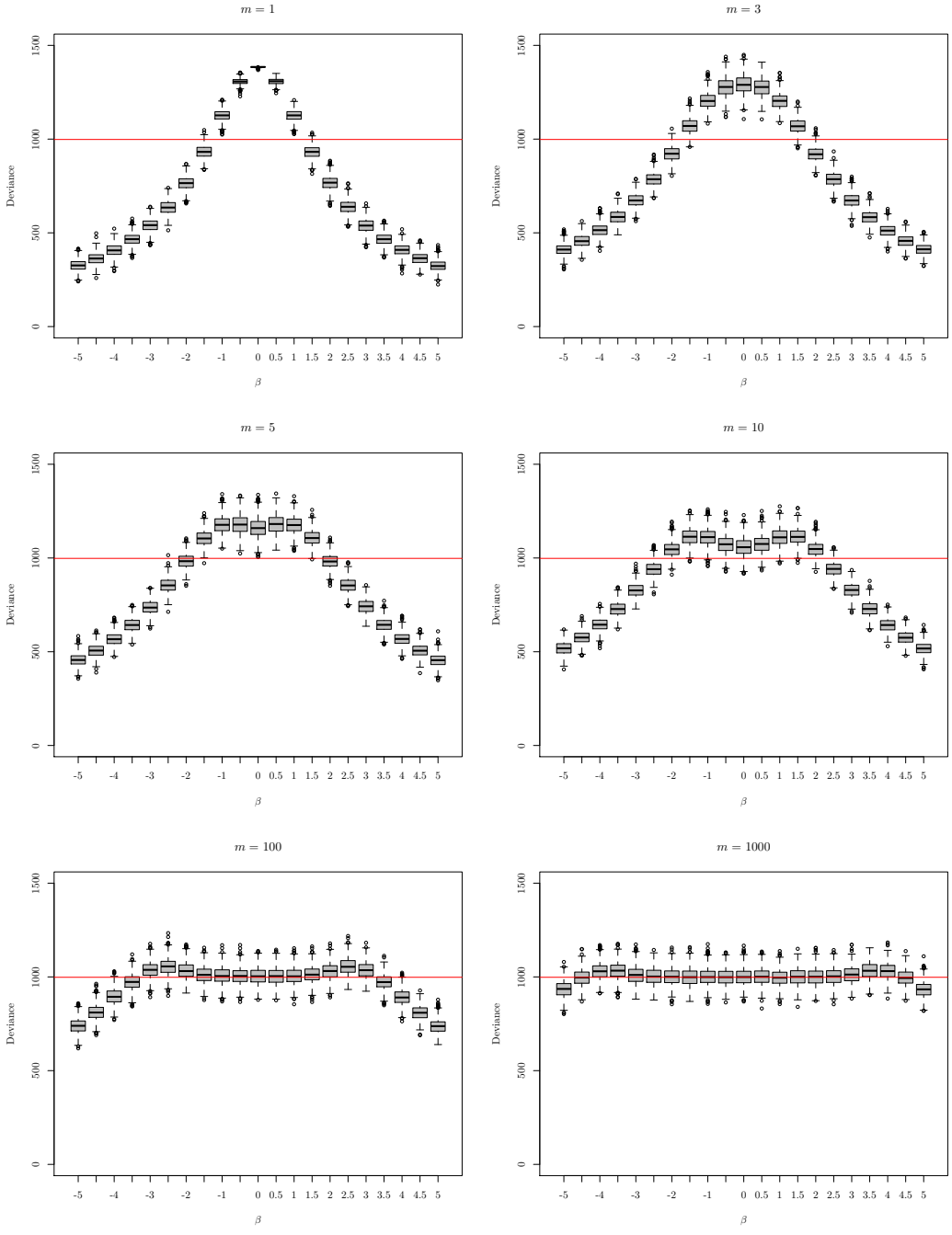
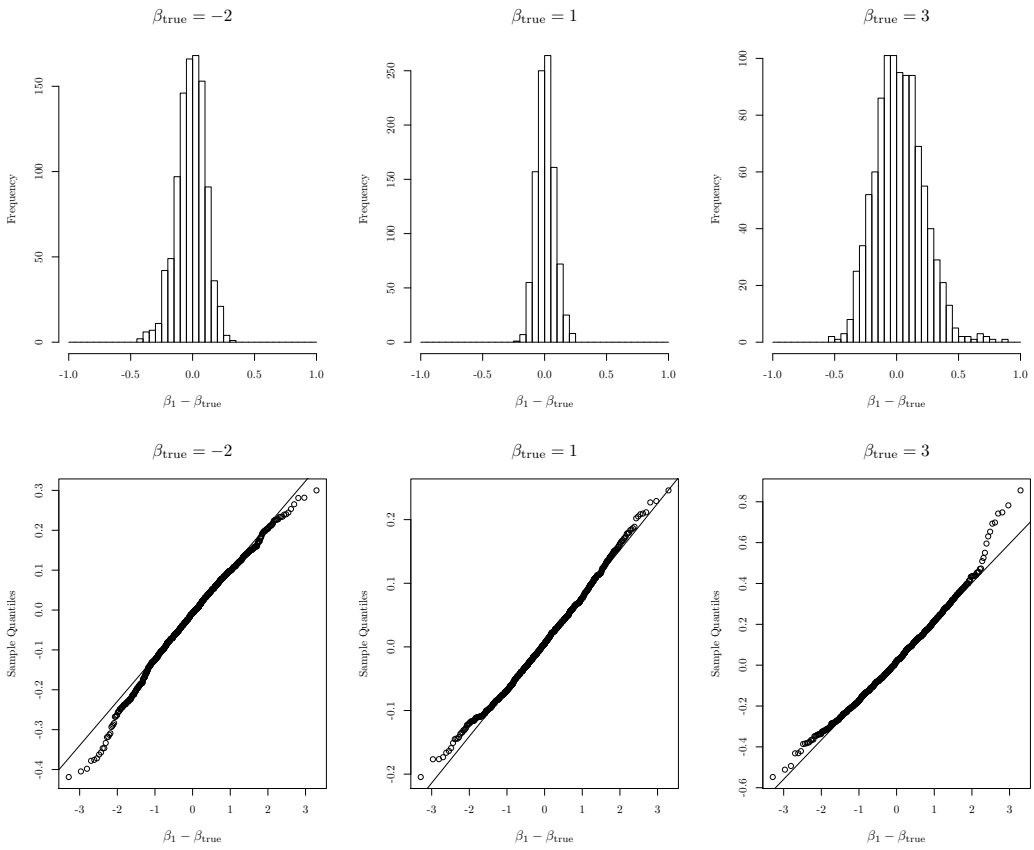


Figure 4: Densities estimates and QQ plot for $\beta_1 - \beta_{\text{true}}$



Plotting the chi-square empirical quantiles against the χ^2_1 alleged hypothetical distribution. The plot is good, with some deviances 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 covariates. 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(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

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

The comparison of the analysis of deviance can be though of as a multivariate version of the Wald test, and asymptotically equivalent. Here $W_n^2 \dot{\sim} \chi_1^2$, and we can assess using the test. Since we cannot use the model assessment criterion that $\texttt{Dev}/(n-p) \dot{\sim} \chi^2_{n-p}$, 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

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

so for control, we get β_0 and the treated groups effect is $\beta_0 + \beta_Y^{(\text{AZT})}$ on the logit scale. Here, the `Intercept` is -1.0361 and $\beta_Y^{(\text{AZT})} = -0.7218$. The odds are decreased when AZT is administered. The log-odds ratio gives

$$\log\left(\frac{\pi_{21}/(1-\pi_{21})}{\pi_{11}/(1-\pi_{11})}\right) = \beta_Y^{(\text{AZT})}$$

π_{jk}– j indexes `AZT=yes` ($j = 2$), for $k = 1$ and 2. Recall that an **odd** is

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

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

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

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

$$\begin{aligned} \log(\mu_1) &= \beta_0 \\ \log(\mu_2) &= \beta_0 + \beta_1 \end{aligned}$$

Then

$$\ell(\beta_0, \beta_1) = y_1\beta_0 + y_2(\beta_0 + \beta_1) - \left(e^{\beta_0} + e^{\beta_0+\beta_1}\right) + \text{ const.}$$

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,

$$\begin{aligned}\ell(\beta_1, \psi) &= (y_1+y_2)\log(\psi) - \psi + y_2\beta_1 - (y_1+y_2)\log(1+e^{\beta_1}) \\ &= \ell_M(\psi; m) + \ell_{Y_2|M}(\beta_1; y_2, m)\end{aligned}$$

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 $\pi \equiv \pi(\beta_1)$.

To estimate β_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

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

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

Suppose we model

$$\log(\mu_k) = \beta_0 + \boldsymbol{\beta}_1 \boldsymbol{x}_k$$

for some continuous covariate $\boldsymbol{x}$ taking values $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K$ for the K random variables. Let

$$\psi = \sum_{k=1}^{K} \exp\left(\beta_0 + \boldsymbol{\beta}_1 \boldsymbol{x}_k\right) = \exp(\beta_0) \sum_{k=1}^{K} \exp\left(\boldsymbol{\beta}_1 \boldsymbol{x}_k\right).$$

Reparametrize to $\psi, \boldsymbol{\beta}_1$; we have for the log-likelihood

$$\ell_n(\psi, \boldsymbol{\beta}_1) = \ell_M(\psi; m) + \ell_{Y|M}(\boldsymbol{\beta}_1; y_1, \ldots, y_K, m)$$

where

$$\ell_m(\psi; m) = m\log(\psi) - \psi \qquad \text{is the Poisson log-likelihood}$$

$$\ell_{Y|M}(\boldsymbol{\beta}_1; \boldsymbol{y}, m) = \beta_1 \sum_{k=1}^{K} \boldsymbol{x}_k y_j - m\log\left(\sum_{k=1}^{K} e^{\boldsymbol{\beta}_1 \boldsymbol{x}_k}\right) \qquad \text{is the multinomial log-likelihood}$$

and

$$\pi_k = \frac{\exp(\boldsymbol{\beta}_1 \boldsymbol{x}_k)}{\sum_{l=1}^{K} \exp(\boldsymbol{\beta}_1 \boldsymbol{x}_l)}$$

Therefore all the information concerning $\boldsymbol{\beta}_1$ lies in the conditional likelihood denoted $\ell_{Y|M}$ for $Y_1, \ldots, Y_{K-1} \left| \sum_{k=1}^{K} Y_k = m \right.$

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

$$\ell_n(\psi, \beta_2, \ldots, \beta_K) = \underbrace{\ell_M(\psi; m)}_{\textbf{(1)}} + \underbrace{\ell_{Y|M}(y_1, \ldots, y_k; m, \beta_2, \ldots, \beta_K)}_{\textbf{(2)}}$$

with

$$\textbf{(1)}: \quad m\log(\psi) - \psi$$

$$\textbf{(2)}: \quad \sum_{k=1}^{K} y_k \beta_k - m\log\left(\sum_{k=1}^{K} \exp(\beta_k)\right)$$

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

$$\pi(x) = \mathsf{P}\left(Y = 1 | X = x\right) = \frac{\exp(x\beta)}{1 + \exp(x\beta)} \qquad = \text{expit}(x\beta)$$

that is

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

the **log-odds** on $Y = 1$ at $X = x$. For continuous predictors,

$$\text{logit}(\pi(x)) = \beta_0 + \beta_1 x$$
$$\therefore \quad \log\left(\frac{\pi(x_1)/(1 - \pi(x_1))}{\pi(x_0)/(1 - \pi(x_0))}\right) = \beta_1(x_1 - x_0).$$

Here β_1 measures the change in log-odds for unit change in x. For a factor predictor taking two levels,

		X	
		1	2
Y	0	π_{01}	π_{02}
	1	π_{11}	π_{12}

and

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

and therefore

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

and β_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 β_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 2×2 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 2×2 table where we have

		X	
		0	1
Y	0	Y_{00}	Y_{01}
	1	Y_{10}	Y_{11}

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

These assumptions only allow us to study certain conditional probabilities.

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

Scientifically, we are interested in $\mathsf{P}(Y = 1|X = 0)$ versus $\mathsf{P}(Y = 1|X = 1)$; 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 $\mathsf{P}(X = 1|Y = 0)$ versus $\mathsf{P}(X = 1|Y = 1)$.

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

$$\pi_1 = \mathsf{P}(Y = 1|X = 1) = \frac{\mathsf{P}(X = 1|Y = 1)\,\mathsf{P}(Y = 1)}{\mathsf{P}(X = 1)} = \varpi_1 \frac{\mathsf{P}(Y = 1)}{\mathsf{P}(X = 1)}$$

therefore

$$\frac{\pi_1}{1 - \pi_1} = \frac{\mathsf{P}(X = 1|Y = 1)}{\mathsf{P}(X = 1|Y = 0)} \frac{\mathsf{P}(Y = 1)}{\mathsf{P}(Y = 0)} = \frac{\varpi_1}{\varpi_0} \frac{\mathsf{P}(Y = 1)}{\mathsf{P}(Y = 0)}$$

and

$$\frac{\pi_0}{1 - \pi_0} = \frac{\mathsf{P}(X = 0|Y = 1)}{\mathsf{P}(X = 0|Y = 0)} \frac{\mathsf{P}(Y = 1)}{\mathsf{P}(Y = 0)} = \frac{1 - \varpi_1}{1 - \varpi_0} \frac{\mathsf{P}(Y = 1)}{\mathsf{P}(Y = 0)}$$

Therefore, dividing $\frac{\pi_1}{1-\pi_1}$ by $\frac{\pi_0}{1-\pi_0}$ yields

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

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 π's, roughly approximating the denominator $\frac{1-\pi_0}{1-\pi_1} \approx 1$ for very rare disease and get π_1/π_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, π 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 $\widehat{\pi}_j$ quantities such as $\text{logit}(\widehat{\pi}_i)$ are needed to derive the properties of the estimators.

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

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

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 $\log(\widehat{\psi}) \dot{\sim} \mathcal{N}\left(\log(\psi), y_{00}^{-1} + y_{01}^{-1} + y_{10}^{-1} + y_{11}^{-1}\right)$. 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(·) for the log-scale. We have PE/(1-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

$$\begin{aligned} \mathsf{P}\left(Y = 1 | X = 1, Z = k\right) &= \pi_{11k} \\ \mathsf{P}\left(Y = 1 | X = 0, Z = k\right) &= \pi_{10k} \end{aligned}$$

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} 2×2 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~X+Z` or `Y~X*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 $Y_i \sim \mathcal{B}(m_i, \pi_i)$ for $i = 1, \ldots, n$. Now we can write $Y_i \overset{d}{=} \sum_{j=1}^{m_i} Y_{ij}$ where $Y_{ij} \sim \mathcal{B}(\pi_i)$ are independent. If we allow π_i to be a **random** quantity,

$$\begin{aligned} \mathsf{E}\left(\pi_i\right) &= \lambda_i & (0 < \lambda_i < 1) \\ \mathsf{Var}\left(\pi_i\right) &= \kappa\lambda_i(1 - \lambda_i) & (\kappa > 0) \end{aligned}$$

then

$$\begin{aligned} \mathsf{E}\left(Y_i\right) &= m_i\lambda_i \\ \mathsf{Var}\left(Y_i\right) &= m_i\lambda_i(1 - \lambda_i)[1 + (m_i - 1)\kappa] \end{aligned}$$

by repeated application of the iterated variance formula and remark that $[1+(m_i-1)\kappa] \geq 1$. Indeed, since $\mathsf{E}_\pi\left(\pi^2\right) = \mathsf{Var}_\pi(\pi) + (\mathsf{E}_\pi(\pi))^2 = \lambda(\kappa(1-\lambda)+\lambda)$

$$\begin{aligned}
\mathsf{Var}_Y(Y) &= \mathsf{E}_\pi\left(\mathsf{Var}_{Y|\pi}(Y)\right) + \mathsf{Var}_\pi\left(\mathsf{E}_{Y|\pi}(Y)\right) \\
&= m\mathsf{E}_\pi(\pi(1-\pi)) + m^2\mathsf{Var}_\pi(\pi) \\
&= m\lambda(1-[\kappa(1-\lambda)+\lambda]) + m^2\kappa\lambda(1-\lambda) \\
&= m\lambda(1-\lambda)[1+(m-\kappa)]
\end{aligned}$$

as claimed.

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

$$\begin{aligned}
\mathsf{E}(Y_i) &= m_i \\
\mathsf{Var}(Y_i) &= \phi m_i \lambda_i (1-\lambda_i)
\end{aligned}$$

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

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

The attention went from modelling π to modelling its expected value λ_i and use the estimating function/equation

$$\sum_{i=1}^{n}(y_i - m_i\lambda_i(\boldsymbol{\beta}))\boldsymbol{x}_i^\top = 0$$

to estimate $\boldsymbol{\beta}$. Finally, we may use

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i - m_i\lambda_i(\widehat{\boldsymbol{\beta}}))^2}{m_i\lambda_i(\widehat{\boldsymbol{\beta}})(1-\lambda_i(\widehat{\boldsymbol{\beta}}))}$$

to estimate ϕ. This formulation represents 'clustering' amongst the $Y_{ij}, j = 1, \ldots, m_i$ around π_i.

This is what is used in `R` under the `quasibinomial` family. For purely **parametric** approach, take $\pi_i \sim \mathcal{B}(\alpha_\pi, \beta_\pi)$. 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, $\boldsymbol{Y}_i \equiv (Y_{i1}, \ldots, Y_{iK})$ 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 π_{ik} is defined for an individual by $\pi_{ik} = \mathsf{P}\left(Y_i^{(R)} = k | X_i = x_i\right)$ 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

$$g(\pi_{ik}) = \boldsymbol{x}_i\boldsymbol{\beta}$$

(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 $Y_i \overset{\mathrm{d}}{=} \sum_{j=1}^{m_i} Y_{ij}$ 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 $\pi_k(\boldsymbol{x}_i) = \mathsf{P}\left(Y_{ij} = k | \boldsymbol{X}_i = \boldsymbol{x}_i\right)$

for $j = 1, \ldots, m_i$ and $k = 1, \ldots, K$ where

$$\log\left(\frac{\pi_k(\boldsymbol{x}_i)}{\pi_1(\boldsymbol{x}_i)}\right) = \beta_{0k} + \boldsymbol{x}_i\beta_{1k}$$

That is,

$$\pi_k(\boldsymbol{x}_i) = \frac{\exp\left(\beta_{0k} + \boldsymbol{x}_i\beta_{1k}\right)}{1 + \sum_{l=2}^{K}\exp\left(\beta_{0l} + \boldsymbol{x}_i\beta_{1l}\right)}$$

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 π 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, $\mathsf{P}\left(Y_i^{(R)} \leq k \mid \boldsymbol{X}_i = \boldsymbol{x}_i\right) = \pi_1(\boldsymbol{x}_i) + \cdots + \pi_k(\boldsymbol{x}_i)$ is the probability of observing a response in categories $\{1, \ldots, k\}$ and we model

$$\log\left(\frac{\mathsf{P}\left(Y_i^{(R)} \leq k \mid \boldsymbol{X}_i = \boldsymbol{x}_i\right)}{1 - \mathsf{P}\left(Y_i^{(R)} \leq k \mid \boldsymbol{X}_i = \boldsymbol{x}_i\right)}\right) = \log\left(\frac{\pi_1(\boldsymbol{x}_i) + \cdots + \pi_k(\boldsymbol{x}_i)}{\pi_{k+1}(\boldsymbol{x}_i) + \cdots + \pi_K(\boldsymbol{x}_i)}\right)$$

as a function of the covariates. To construct the likelihood for this model, we need explicit forms for $\pi_k(\boldsymbol{x}_i)$ by inverting this more complicated link function. Note that

$$\pi_k(\boldsymbol{x}_i) = \mathsf{P}\left(Y_i^{(R)} = k \mid \boldsymbol{X}_i = \boldsymbol{x}_i\right) = \mathsf{P}\left(Y_i^{(R)} \leq k \mid \boldsymbol{X}_i = \boldsymbol{x}_i\right) - \mathsf{P}\left(Y_i^{(R)} \leq k - 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i\right)$$

- **Proportional odds model**: in this case, we look at

$$\text{logit}\left(\mathsf{P}\left(Y_i^{(R)} \leq k \mid \boldsymbol{X}_i = \boldsymbol{x}_i\right)\right) = \beta_{0k} + \boldsymbol{x}_i\beta_1$$

where β_1 is **common** for all k and $\beta_{01} < \beta_{02} < \cdots < \beta_{0(K-1)}$.[67] For scalar x_i,

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

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.[68] 1|2 corresponds in our notation to β_{01} (*i.e.* the probability left of k). Note that the non-intercept terms are reported on the negative scale, *i.e.* $-\widehat{\beta}_1$. 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 $\text{logit}(\mathsf{P}\left(Y = k|\boldsymbol{X} = \boldsymbol{x}\right), \mathsf{P}\left(Y = k+1\right)/\mathsf{P}\left(Y = k\right)$ or $\mathsf{P}\left(Y = k+1\right)/\mathsf{P}\left(Y = k \text{ or } Y = k+1\right)$ 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 $\mathsf{P}\left(Y_i^{(R)} \geq k \mid \boldsymbol{X}_i = \boldsymbol{x}_i\right)$ 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

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

is approximately χ^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

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

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

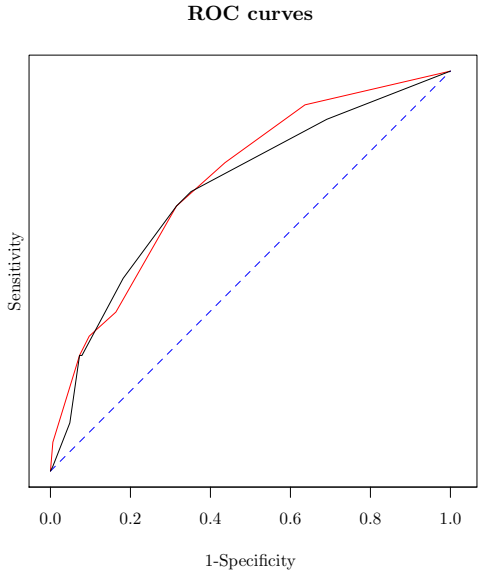
Table 1: **Misclassification table (or confusion matrix)**

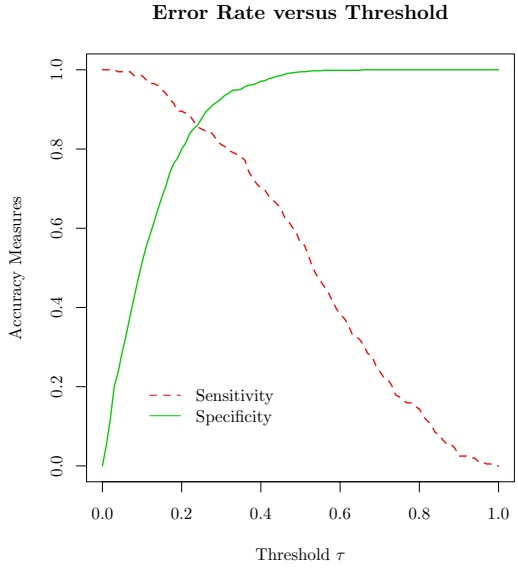
		Predicted $\widehat{Y}_i$	
		0	1
True Y_i	0	a	b
	1	c	d

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 τ 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 τ is optimally chosen: such as model has an ROC curve that, for some τ, 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 τ 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 $Y_i \sim \mathcal{B}(\pi_i(\boldsymbol{x}_i))$ where $\text{logit}(\pi_i(\boldsymbol{x}_i)) = \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}$. In this setup, $\boldsymbol{\beta}$ is the parameter of interest, while β_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

$$\mathcal{L}_n(\beta_0, \boldsymbol{\beta}) = \frac{\exp\left(T_1(\boldsymbol{y})\beta_0 + \boldsymbol{T}_2(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{\beta}\right)}{\prod_{i=1}^n \left(1 + \exp\left(\beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}\right)\right)}$$

where

$$T_1(\boldsymbol{y}) = \sum_{i=1}^n y_i \qquad \boldsymbol{T}_2(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^n y_i\boldsymbol{x}_i$$

are sufficient statistics for $\beta_0, \boldsymbol{\beta}$ respectively.

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

$$\begin{aligned}\mathcal{L}_n^{(T_1)}(\boldsymbol{\beta}, t) &= \mathsf{P}\left(Y_1 = y_1, \ldots, Y_n = y_n \mid T_1(\boldsymbol{y}) = t\right) \\ &= \frac{\exp\left(\boldsymbol{T}_2(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{\beta}\right)}{\sum_{\boldsymbol{z}\in\mathcal{Y}_1(t_1)} \exp\left(\boldsymbol{T}_2(\boldsymbol{x}, \boldsymbol{z})\boldsymbol{\beta}\right)}\end{aligned}$$

where

$$\mathcal{Y}_1(t_1) = \left\{\boldsymbol{z} = (z_1, \ldots, z_n) \ : \ \sum_{i=1}^n z_i = t\right\}$$

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 $\boldsymbol{z}$ 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 2×2 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 $j = 2$.[69] The cell counts in the resulting 2×2 table can be modeled using $\mathcal{M}(n; \pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$, assuming n individuals are sampled independently from the population. Inference about $\pi_{00}, \ldots, \pi_{11}$ is straightforward.[70] – we use the ML estimates

$$(\widehat{\pi}_{00}, \widehat{\pi}_{01}, \widehat{\pi}_{10}, \widehat{\pi}_{11}) = \frac{1}{n}(n_{00}, n_{01}, n_{10}, n_{11})$$

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

$$\mathsf{Cov}(\widehat{\pi}_{1\star}, \widehat{\pi}_{\star 1}) = \frac{\pi_{00}\pi_{11} - \pi_{01}\pi_{10}}{n}$$

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

$$\sqrt{n}(\widehat{\delta} - \delta) \xrightarrow{d} \mathcal{N}(0, V(\pi))$$

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

[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.

[70] Giving $2n$ observations with independent pairs

[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 $W_n^2 \sim \chi_1^2$.

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_{i1}, Y_{i2}) be the responses for study $j = 1, 2, \ldots$ for individual i; $\pi_{ij} = \mathsf{P}(Y_{ij} = 1)$ is the probability of "$Y = 1$" response for individual i on study j. Consider the model

$$\operatorname{logit}(\pi_{ij}) = \begin{cases} \beta_{0i} & \text{if } j = 1 \\ \beta_{0i} + \zeta & \text{if } j = 2 \end{cases}$$

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, β_{0i} measures an individual-level (stratum-specific) component of the response probabilities. Thus ζ 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 $\mathcal{L}_n(\beta_{01}, \ldots, \beta_{0n}, \zeta)$ for the data is

$$\prod_{i=1}^{n} \left(\operatorname{expit}(\beta_{0i})^{y_{i1}} (1 - \operatorname{expit}(\beta_{0i}))^{1-y_{i1}} \operatorname{expit}(\beta_{0i} + \zeta)^{y_{i2}} (1 - \operatorname{expit}(\beta_{0i} + \zeta))^{1-y_{i2}}\right)$$

We may thus consider the β_{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 ζ. We have

$$\mathsf{P}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2} \mid S_i = s) = \begin{cases} 1 & \text{if } (s, y_{i1}, y_{i2}) = (0,0,0) \text{ or } (2,1,1) \\ \frac{e^{y_{i2}\zeta}}{1+e^{\zeta}} & \text{if } (s, y_{i1}, y_{i2}) = (1,0,1) \text{ or } (1,1,0) \end{cases}$$

The parameter $\widehat{\zeta}$ has a different form, given by $\widehat{\zeta} = \log\left(\frac{n_{01}}{n_{10}}\right)$ and $\widehat{\text{se}} = \sqrt{n_{01}^{-1} + n_{10}^{-1}}$.

The conditional likelihood approach is very appealing when the number of parameter is very large or grows with the data.

Section 5
Special Topics

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 $\mathcal{G}(\mu, \nu)$ where $\mathsf{E}(Y) = \mu, \mathsf{Var}(Y) = \mu/\nu$ (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

$$\eta = \beta_0 + \boldsymbol{\beta}_1 \boldsymbol{x}$$
$$\eta = \beta_0 + \boldsymbol{\beta}_1 / \boldsymbol{x}$$
$$\eta = \beta_0 + \boldsymbol{\beta}_1 / \boldsymbol{x} + \boldsymbol{\beta}_2 \boldsymbol{x}$$

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

$$\frac{\Delta\text{Deviance}}{\widehat{\phi}} \overset{\cdot}{\sim} \chi^2_{\Delta\text{df}}$$

where $\widehat{\phi}$ 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 $Y_i \sim \mathcal{G}$ and $\mathsf{E}(Y_i) = \mu_i$, $\mathsf{Var}(Y_i) = \frac{\mu_i^2}{\nu_i}$ 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.

$$D(y;\widehat{\mu}) = -2\sum_{i=1}^{n} w_i \left[\log\left(\frac{y_i}{\widehat{\mu}_i}\right) - \left(\frac{y_i - \widehat{\mu}_i}{\widehat{\mu}_i}\right)\right]$$

–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

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

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

$$\sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{V(\mu_i)\dot{g}(\mu_i)} x_{ij} = 0, \quad j = 1, \ldots, p$$

where μ_i is a function of the unknown parameters β and in this case g is the link function:

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

$g(t)$ defines the relationship between η_i and μ_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 $\mathsf{E}(Y_i) = \mu_i$ and $\mathsf{Var}(Y_i) = \sigma^2 V(\mu_i)$ for an arbitrary $V(\cdot)$ returning a positive quantity.

Let $U_i = \frac{Y_i - \mu_i}{\sigma^2 V(\mu_i)}$. Then $\mathsf{E}(U_i) = 0$ and $\mathsf{Var}(U_i) = (\sigma^2 V(\mu_i))^{-1}$ and also $-\mathsf{E}\left(\frac{\partial U_i}{\partial \mu_i}\right) =$

$\frac{1}{\sigma^2 V(\mu_i)}$ thus U_i has the same properties as a **score** random variable.

$$S(Y_i, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_T(Y_i; \boldsymbol{\theta}).$$

Therefore, we could use

$$\sum_{i=1}^{n} U_i(\boldsymbol{\beta}) = 0 \qquad (p \times 1)$$

to estimate $\boldsymbol{\beta}$ 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

$$Q(\mu_i, y_i) = \int_{y_i}^{\mu_i} \left(\frac{y_i - t}{\sigma^2 V(t)} \right) \mathrm{d}t$$

and the **quasi-deviance** $D_i(\mu_i; y_i) = -2\sigma^2 Q(\mu_i, y_i)$.

Asymptotic variance

If $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ $(n \times 1)$, $\boldsymbol{V}(\boldsymbol{\mu}) = \text{diag}\,(V(\mu_1), \ldots, V(\mu_n))$ an $(n \times n)$ covariance matrix, $\boldsymbol{B}(\boldsymbol{\mu})$ is the $n \times p$ matrix with $(i,j)^{\text{th}}$ term $[\boldsymbol{B}]_{ij} = \frac{\partial \mu_i}{\partial \beta_j}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, p$ and

$$\boldsymbol{U}(\boldsymbol{\beta}) = \frac{1}{\sigma^2} \boldsymbol{B}^\top \left(\boldsymbol{V}(\boldsymbol{\mu}) \right)^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})$$

an $(n \times 1)$ system. Then $\mathsf{E}\left(\boldsymbol{U}(\boldsymbol{\beta})\right) = \mathbf{0}_n$ and $\mathsf{Var}\left(\boldsymbol{U}(\boldsymbol{\beta})\right) = \frac{1}{\sigma^2} \boldsymbol{B}^\top \boldsymbol{V}^{-1} \boldsymbol{B}$ and $\widehat{\boldsymbol{\beta}}$ satisfies $\boldsymbol{U}(\widehat{\boldsymbol{\beta}}) = \mathbf{0}_n$ (an $(n \times 1)$ system).

Thus

$$\begin{aligned} \mathsf{E}\left(\widehat{\boldsymbol{\beta}}\right) &\simeq \boldsymbol{\beta} \\ \mathsf{Var}\left(\widehat{\boldsymbol{\beta}}\right) &\simeq \sigma^2 \left(\boldsymbol{B}^\top \boldsymbol{V}^{-1} \boldsymbol{B} \right)^{-1}. \end{aligned}$$

When n is large, by analogy with weighted least squares, $\widehat{\boldsymbol{\beta}}$ is computed using recursive methods *e.g.* Fisher scoring

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} + \left(\widehat{\boldsymbol{B}}^{(t)^\top} \widehat{\boldsymbol{V}}^{(t)^{-1}} \widehat{\boldsymbol{B}}^{(t)} \right)^{-1} \widehat{\boldsymbol{B}}^{(t)} \widehat{\boldsymbol{V}}^{(t)^{-1}} (\boldsymbol{y} - \boldsymbol{\mu}^{(t)})$$

yields $\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{V}}$ and $\widehat{\sigma^2}$ where

$$\widehat{\sigma^2} = \frac{1}{n-p} \sum_{i=1}^{n} \left(\frac{Y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}} \right)^2$$

Index

License