Statistical modelling

Rajen D. Shah
r.shah@statslab.cam.ac.uk

Course webpage: `http://www.statslab.cam.ac.uk/~rds37/statistical_modelling.html`

Introduction

This course is largely about analysing data composed of observations that come in the form of pairs

$$(y_1, x_1), \dots, (y_n, x_n). \tag{0.0.1}$$

Our aim will be to infer an unknown *regression function* relating the values y_i, to the x_i, which may be p-dimensional vectors $x_i = (x_{i1}, \dots, x_{ip})^T$. The y_i are often called the response, target or dependent variable; the x_i are known as predictors, covariates, independent variables or explanatory variables. Below are some examples of possible responses and covariates.

Response	Covariates
House price	Numbers of bedrooms, bathrooms; Plot area; Year built; Location
Weight loss	Type of diet plan; type of exercise regime
Short-sightedness	Parents' short-sightedness; Hours spent watching TV or reading books

First note that in each of the examples above, it would be hopeless to attempt to find a deterministic function that gives the response for every possible set of values of the covariates. Instead, it makes sense to think of the data-generating mechanism as being inherently random, with perhaps a deterministic function relating *average* values of the responses to values of the covariates.

We model the responses y_i as realisations of random variables Y_i. Depending on how the data were collected, it may seem appropriate to also treat the x_i as random. However, in such cases we usually condition on the observed values of the explanatory variables. To aid intuition, it may help to imagine a hypothetical sequence of repetitions of the 'experiment' that was conducted to produce the data with the $x_i, i = 1, \dots, n$ held fixed, and think of the dataset at hand as being one of the many elements of such a sequence.

In the course Principles of Statistics, theory was developed for data that were i.i.d. In our setting here, this assumption is not appropriate: the distributions of Y_i and Y_j may well be different is $x_i \neq x_j$. In fact what we are interested in is *how* the distributions of the Y_i differ. However, we will still usually assume that the data are at least independent. It turns out that with this assumption of independence, much of the theory from Principles of Statistics can be applied, with little modification.

In this course we will study some of the most popular and important statistical models for data of the form (0.0.1). We begin with the linear model, which you will have met in Statistics IB.

Contents

1 Linear models **2**

1.1 Ordinary least squares (OLS) . . . 2

1.1.1 Orthogonal projections . . . 3

1.1.2 Analysis of OLS . . . 4

1.2 Normal errors . . . 5

1.2.1 Maximum likelihood estimation . . . 5

1.2.2 The multivariate normal distribution and related distributions . . . 6

1.2.3 Inference for the normal linear model . . . 8

1.2.4 ANOVA and ANCOVA . . . 13

1.2.5 Model selection . . . 14

1.2.6 Model checking . . . 18

2 Exponential families and generalised linear models **22**

2.1 Non-normal responses . . . 22

2.2 Exponential families . . . 22

2.3 Exponential dispersion families . . . 25

2.4 Generalised linear models . . . 27

2.4.1 Choice of link function . . . 27

2.4.2 Likelihood equations . . . 27

2.5 Inference . . . 28

2.5.1 The score function . . . 29

2.5.2 Fisher information . . . 29

2.5.3 Two key asymptotic results . . . 30

2.5.4 Inference in generalised linear models . . . 33

2.6 Computation . . . 35

3 Specific regression problems **37**

3.1 Binomial regression . . . 37

3.1.1 Link functions . . . 37

3.1.2 *A classification view of logistic regression* . . . 39

3.1.3 *Logistic regression and linear discriminant analysis* . . . 39

3.1.4 Model checking . . . 40

3.2 Poisson regression . . . 41

3.2.1 Likelihood equations . . . 42

3.2.2 Modelling rates with an offset . . . 43

3.2.3 Contingency tables . . . 43

3.2.4 Test for independence of columns and rows . . . 46

3.2.5 Test for homogeneity of rows . . . 47

Chapter 1

Linear models

1.1 Ordinary least squares (OLS)

The linear regression model assumes that

$$Y = X\beta + \varepsilon,$$

where

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \; X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix}, \; \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

and the ε_i are to be considered as random errors that satisfy

(A1) $\mathbb{E}(\varepsilon_i) = 0$,

(A2) $\mathrm{Var}(\epsilon_i) = \sigma^2$,

(A3) $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.

A word on models. It is important to recognise that this, or any statistical model is a mathematical object and cannot really be thought of as a 'true' representation of reality. Nevertheless statistical models can nevertheless be a *useful* representation of reality. Though the model may be wrong, it can still be used to answer questions of interest, and help inform decisions.

The design matrix X

If we want to include an intercept term in the linear model, we can simply take our design matrix X as

$$X = \begin{pmatrix} 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_n^T \end{pmatrix}.$$

To include quadratic terms, we may take

$$X = \begin{pmatrix} 1 & x_1^T & x_{11}^2 & \cdots & x_{1p}^2 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n^T & x_{n1}^2 & \cdots & x_{np}^2 \end{pmatrix}.$$

The resulting model will not be linear in the x_i, but it is still a linear model because it is linear in β.

Least squares

Under assumptions A1–A3, a sensible way to estimate β is using OLS. This gives an estimate $\hat{\beta}$ that satisfies

$$\begin{aligned}\hat{\beta} :=& \arg\min_{b \in \mathbb{R}^p} \|Y - Xb\|^2 \\ =& (X^T X)^{-1} X^T Y,\end{aligned}$$

provided the n by p matrix X has full column rank (i.e. $\mathrm{r}(X) = p$) so $X^T X$ is invertible (see example sheet). The fitted values, $\hat{Y} := X\hat{\beta}$ are then given by $X(X^T X)^{-1} X^T Y$. Let $P := X(X^T X)^{-1} X^T$. Then P known as the 'hat' matrix because it puts the hat on Y. In fact it is an *orthogonal projection* on to the column space of X. To discuss this further, we recall some facts about projections from linear algebra.

1.1.1 Orthogonal projections

Let V be a subspace of $\mathbb{R}^n$ and define

$$V^\perp := \{w \in \mathbb{R}^n : w^T v = 0 \text{ for all } v \in V\}.$$

$V^\perp$ is known as the *orthogonal complement* of V. **Fact:** Then $\mathbb{R}^n = V \oplus V^\perp$, so each $x \in \mathbb{R}^n$ may be written uniquely as $x = v + w$ with $v \in V$ and $w \in V^\perp$. This follows because we can pick an orthonormal basis for V, $v_1, \ldots, v_m$, and then extend it to an orthonormal basis $v_1, \ldots, v_m, v_{m+1}, \ldots, v_n$ for $\mathbb{R}^n$. $V^\perp$ is then the span of $v_{m+1}, \ldots, v_n$.

Definition 1. A matrix $\Pi \in \mathbb{R}^{n \times n}$ is called an *orthogonal projection on to* $V \leq \mathbb{R}^n$ if $\Pi x = v$ when $x = v + w$ with $v \in V$, $w \in V^\perp$. Thus Π acts as the identity on V and sends everything orthogonal to V to 0. We will say that Π is an *orthogonal projection* if it is an orthogonal projection on to its column space.

Let Π be an orthogonal projection on to V. Here are some important properties.

(i) The column space (a.k.a. range, image) of Π is V.

(ii) $I - \Pi$ is an orthogonal projection on to $V^\perp$, so $I - \Pi$ fixes everything in $V^\perp$ and sends everything in V to 0. Indeed, $(I - \Pi)(v + w) = v + w - \Pi(v + w) = v + w - v = w$.

(iii) $\Pi^2 = \Pi = \Pi^T$, so Π is idempotent and symmetric. The former is clear from the definition. To see that Π is symmetric observe that for all $u_1, u_2 \in \mathbb{R}^n$,

$$0 = (\Pi u_1)^T (I - \Pi) u_2 = u_1^T (\Pi^T - \Pi^T \Pi) u_2 = 0 \Leftrightarrow \Pi^T = \Pi^T \Pi.$$

In fact, we can see that $\Pi^2 = \Pi = \Pi^T$ is an alternative definition for Π being an orthogonal projection. Indeed, if v is in the column space of Π, then $v = \Pi u$, for some $u \in \mathbb{R}^n$. But then $\Pi v = \Pi^2 u = \Pi u = v$, so Π fixes everything in its column space. Now if v is orthogonal to the column space of Π, then $\Pi v = \Pi^T v = 0$.

(iv) Orthonormal bases of V and $V^\perp$ are eigenvectors of Π with eigenvalues 1 and 0 respectively. Therefore we can from the eigendecomposition $\Pi = UDU^T$ where U is an orthogonal matrix with columns as eigenvectors of Π and D is a diagonal matrix of corresponding eigenvalues.

(v) $\mathrm{r}(\Pi) = \dim(V)$. Also, by the eigendecomposition above,

$$\mathrm{r}(\Pi) = \mathrm{tr}(D) = \mathrm{tr}(U^T U D) = \mathrm{tr}(U D U^T) = \mathrm{tr}(\Pi),$$

where we have used the cyclic property of the trace.

Note that the matrix $P = X(X^T X)^{-1} X^T$ defined earlier is the orthogonal projection on to the column space of X. Indeed, $PXb = Xb$ and if w is orthogonal to the column space of X, so $X^T w = 0$, then $Pw = 0$. Also, our derivation of PY as the linear combination of columns of X that is closest in Euclidean distance to Y reveals another property of orthogonal projections: if Π is an orthogonal projection on to V, then for any $v \in \mathbb{R}^n$, Πv is the closest point on V to the vector v—in other words

$$\Pi v = \arg\min_{u \in V} \|v - u\|^2.$$

1.1.2 Analysis of OLS

Back to Statistics: the fitted values of OLS are given by the projection of the vector of responses, Y on to the column space of the matrix of predictors X.

Recall that the covariance between two random vectors $Z_1 \in \mathbb{R}^{n_1}$ and $Z_2 \in \mathbb{R}^{n_2}$ is defined by

$$\mathrm{Cov}(Z_1, Z_2) := \mathbb{E}[\{Z_1 - \mathbb{E}(Z_1)\}\{Z_1 - \mathbb{E}(Z_1)\}^T].$$

The correlation matrix between Z_1 and Z_2 is the n_1 by n_2 matrix with entries given by

$$\mathrm{Corr}(Z_1, Z_2) := \frac{\mathrm{Cov}(Z_1, Z_2)_{ij}}{\sqrt{\mathrm{Var}(Z_{1,i})\mathrm{Var}(Z_{2,j})}}.$$

For any constants $a_1 \in \mathbb{R}^{n_1}$ and $a_2 \in \mathbb{R}^{n_2}$, $\mathrm{Cov}(Z_1 + a_1, Z_2 + a_2) = \mathrm{Cov}(Z_1, Z_2)$. Also recall that for any d by n_1 matrix A and any constant vector $m \in \mathbb{R}^{n_1}$, as expectation is a linear operator, $\mathbb{E}(m + AZ_1) = m + A\mathbb{E}(Z_1)$.

We can show that the vector of *residuals*, $\hat{\varepsilon} := Y - \hat{Y} = (I - P)Y$ is uncorrelated with the fitted values $\hat{Y}$:

$$\begin{aligned}
\mathrm{Cov}(PY, (I-P)Y) &= \mathrm{Cov}(P\varepsilon, (I-P)\varepsilon) \\
&= \mathbb{E}(P\varepsilon\varepsilon^T (I-P)^T) \\
&= P \underbrace{\mathbb{E}(\varepsilon\varepsilon^T)}_{\sigma^2 I}(I-P) \\
&= \sigma^2 P(I-P) = 0.
\end{aligned}$$

Here is another way to think of the OLS coefficients that can offer further insight. Let us write X_j for the j^{th} column of X, and X_{-j} for the $n \times (p-1)$ matrix formed by removing the j^{th} column from X. Define P_{-j} as the orthogonal projection on to the column space of X_{-j}.

Proposition 1. *Let $X_j^\perp := (I - P_{-j})X_j$, so $X_j^\perp$ is the orthogonal projection of X_j on to the orthogonal complement of the column space of X_{-j}. Then*

$$\hat{\beta}_j = \frac{(X_j^\perp)^T Y}{\|X_j^\perp\|^2}.$$

Proof. Note that $Y = PY + (I - P)Y$ and

$$X_j^T(I - P_{-j})(I - P)Y = X_j^T(I - P)Y = 0,$$

so

$$\frac{(X_j^\perp)^T Y}{\|X_j^\perp\|^2} = \frac{(X_j^\perp)^T X(X^TX)^{-1}X^TY}{\|X_j^\perp\|^2}.$$

Since $X_j^\perp$ is orthogonal to the column space of X_{-j}, we have

$$(X_j^\perp)^T X = (0 \cdots 0 \ \underset{\substack{\uparrow \\ j^{\text{th}} \text{ position}}}{(X_j^\perp)^T X_j} \ 0 \cdots 0)$$

and $(X_j^\perp)^T X_j = X_j^T(I - P_{-j})X_j = \|(I - P_{-j})X_j\|^2$. □

We see that $\text{Var}(\hat\beta_j) = \sigma^2\|X_j^\perp\|^{-2}$. Thus if X_j is closely aligned to the column space of X_{-j}, the variance of $\hat\beta_j$ will be large. In particular, if a pair of variables are highly correlated with each other, the variances of the estimates of the corresponding coefficients will be large.

We can measure the quality of a regression procedure by its mean-squared prediction error (MSPE). This is defined here as

$$\frac{1}{n}\mathbb{E}(\|X\beta - X\hat\beta\|^2).$$

Note that $X\hat\beta = PY = X\beta + P\varepsilon$, so

$$\mathbb{E}(\|X\beta - X\hat\beta\|^2) = \mathbb{E}(\varepsilon^T P^T P\varepsilon) = \mathbb{E}\{\text{tr}(\varepsilon^T P\varepsilon)\} = \text{tr}\{\mathbb{E}(\varepsilon\varepsilon^T)P\} = \sigma^2\text{tr}(P) = \sigma^2 p.$$

Thus

$$\frac{1}{n}\mathbb{E}(\|X\beta - X\hat\beta\|^2) = \sigma^2\frac{p}{n}.$$

More is true. Note that $\hat\beta$ is unbiased, as

$$\mathbb{E}_\beta(\hat\beta) = \mathbb{E}_\beta\{(X^TX)^{-1}X^TX\beta\} = \beta. \tag{1.1.1}$$

Further,

$$\text{Var}(\hat\beta) = (X^TX)^{-1}X^T\text{Var}(\varepsilon)\{(X^TX)^{-1}X^T\}^T = \sigma^2(X^TX)^{-1}. \tag{1.1.2}$$

In fact it is the best linear unbiased estimator (BLUE), that is for any other estimator $\tilde\beta$ that is linear in Y, we have $\text{Var}(\tilde\beta) - \text{Var}(\hat\beta)$ is positive semi-definite. In particular this means that given a new observation $x^* \in \mathbb{R}^p$, we can estimate the regression function at x^* optimally in the sense that $\mathbb{E}\{(x^{*T}\beta - x^{*T}\hat\beta)^2\} \le \mathbb{E}\{(x^{*T}\beta - x^{*T}\tilde\beta)^2\}$.

Theorem 2 (Gauss–Markov). *Under (A1)–(A3) OLS is BLUE.*

1.2 Normal errors

1.2.1 Maximum likelihood estimation

The method of least squares is just one way to construct as estimator. A more general technique is that of maximum likelihood estimation. Here given data $y \in \mathbb{R}^n$ that we take as a realisation of a random variable Y, we specify its density $f(y; \theta)$ up to some unknown vector of parameters

$\theta \in \Theta \subseteq \mathbb{R}^d$, where Θ is the parameter space. The likelihood function is a function of θ for each fixed y given by

$$L(\theta) := L(\theta; y) = c(y) f(y; \theta),$$

where $c(y)$ is an arbitrary constant of proportionality. We form an estimate $\hat{\theta}$ by choosing that θ which maximises the likelihood. Often it is easier to work with the log-likelihood defined by

$$\ell(\theta) := \ell(\theta; y) = \log f(y; \theta) + \log(c(y)).$$

If we assume that the errors ε_i in our linear model have $N(0, \sigma^2)$ distributions, we see that the log-likelihood for (β, σ^2) is

$$\ell(\beta, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2.$$

The maximiser of this over β is precisely the least squares estimator $(X^T X)^{-1} X^T Y$. Maximum likelihood does much more than simply give us another interpretation of OLS here. It allows us to perform *inference*: that is construct confidence interval for parameters and perform hypothesis tests. Before moving on to this topic, we review some facts about the multivariate normal distribution.

1.2.2 The multivariate normal distribution and related distributions

Multivariate normal distribution

We say a random variable $Z \in \mathbb{R}^d$ has a d-variate normal distribution if for every $t \in \mathbb{R}^d$, $t^T Z$ has a univariate normal distribution. Thus linear combinations of Z are also normal: for any $m \in \mathbb{R}^k$ and $A \in \mathbb{R}^{k \times d}$, $m + AZ$ is multivariate normal. **Fact:** the multivariate normal distribution is uniquely characterised by its mean and variance. Thus we write write $Z \sim N_d(\mu, \Sigma)$ when $\mathbb{E}(Z) = \mu$ and $\mathrm{Var}(Z) = \Sigma$. Note that $m + AZ \sim N_k(m + \mu, A\Sigma A^T)$.

When Σ is invertible, the density of Z is given by

$$f(z; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (z - \mu)^T \Sigma^{-1} (z - \mu) \right\}, \quad z \in \mathbb{R}^d.$$

Note that for example, the vector of residuals from the normal linear model $(I - P)\varepsilon \sim N_n(0, \sigma^2(I - P))$ but it does not have a density of the form given above as $I - P$ is not invertible.

Proposition 3. *If Z_1 and Z_2 are jointly normal (i.e. (Z_1, Z_2) has a multivariate normal distribution), then if* $\mathrm{Cov}(Z_1, Z_2) := \mathbb{E}[\{Z_1 - \mathbb{E}(Z_1)\}\{Z_2 - \mathbb{E}(Z_2\}^T] = 0$, *we have that Z_1 and Z_2 are independent.*

Proof. Let $\tilde{Z}_1$ and $\tilde{Z}_2$ be independent and have the same distributions as Z_1 and Z_2 respectively. Then the mean and variance of the random variables $(\tilde{Z}_1, \tilde{Z}_2)$ and (Z_1, Z_2) are identical and they are both multivariate normal (the former is multivariate normal because sums of independent normal random variables are normal). Since a multivariate normal distribution is uniquely determined by its mean and variance, we must have $(\tilde{Z}_1, \tilde{Z}_2) \overset{d}{=} (Z_1, Z_2)$. □

χ^2 distribution

We say Z has a χ^2 distribution on k degrees of freedom, and write $Z \sim \chi^2_k$ if $Z \overset{d}{=} Z_1^2 + \cdots + Z_k^2$ where $Z_1, \ldots, Z_k \overset{\text{i.i.d.}}{\sim} N(0,1)$.

Proposition 4. *Let Π be an n by n orthogonal projection with rank k, and let $\varepsilon \sim N_n(0, \sigma^2 I)$. Then $\|\Pi\varepsilon\|^2 \sim \sigma^2\chi^2_k$.*

Proof. As Π is an orthogonal projection, we may form its eigendecomposition UDU^T where U is an orthogonal matrix and D is diagonal with entries in $\{0,1\}$. Then

$$\|\Pi\varepsilon\|^2 = \|DU^T\varepsilon\|^2 \quad \text{and} \quad \|D\varepsilon\|^2$$

have the same distribution. But

$$\frac{1}{\sigma^2}\|D\varepsilon\|^2 = \frac{1}{\sigma^2}\sum_{i:D_{ii}\neq 0} \varepsilon_i^2 \sim \chi^2_k. \qquad \square$$

Student's t distribution

We say Z has a t distribution on k degrees of freedom, and write $Z \sim t_k$ if

$$Z \overset{d}{=} \frac{Z_1}{\sqrt{Z_2/k}}$$

where Z_1 and Z_2 are independent $N(0,1)$ and χ^2_k random variables respectively.

Multivariate t distribution

This is a generalisation of the Student's t distribution above. We say Z has a p-dimensional multivariate t distribution on k degrees of freedom, and write $Z \sim t_k(\mu, \Sigma)$ if

$$Z \overset{d}{=} \mu + \frac{Z_1}{\sqrt{Z_2/k}}$$

where Z_1 and Z_2 are independent $N_p(0,\Sigma)$ and χ^2_k random variables respectively. It can be shown that when Σ is invertible, Z has density

$$f(z) := \frac{\Gamma((k+p)/2)}{\Gamma(k/2)(k\pi)^{p/2}}|\Sigma|^{-1/2}\left(1 + \frac{1}{k}(z-\mu)^T\Sigma^{-1}(z-\mu)\right)^{-(k+p)/2},$$

where the gamma function Γ satisfies $\Gamma(m) = (m-1)!$ for $m \geq 1$.

F distribution

We say Z has an F distribution on k and l degrees of freedom, and write $Z \sim F_{k,l}$ if

$$Z \overset{d}{=} \frac{Z_1/k}{Z_2/l}$$

where Z_1 and Z_2 are independent and follow χ^2_k and χ^2_l distributions respectively.

Notation. We will denote the upper α-points of the χ^2_k, t_k and $F_{k,l}$ distributions by $\chi^2_k(\alpha)$, $t_k(\alpha)$ and $F_{k,l}(\alpha)$ respectively. (So, for example, if $Z \sim \chi^2_k$ then $\mathbb{P}\{Z \geq \chi^2_k(\alpha)\} = \alpha$. As the t_k distribution is symmetric, if $Z \sim t_k$, then $\mathbb{P}\{-t_k(\alpha/2) \leq Z \leq t_k(\alpha/2)\} = 1 - \alpha$.)

Informal summary

$$\begin{aligned}\chi^2_k &= \underbrace{N(0,1)^2 + \cdots + N(0,1)^2}_{k \text{ times}} \\ t_k &= \frac{N(0,1)}{\sqrt{\chi^2_k/k}} \\ F_{k,l} &= \frac{\chi^2_k/k}{\chi^2_l/l}, \\ \text{so } t_l^2 &= F_{1,l}\end{aligned}$$

with appropriate independence between relevant random variables.

1.2.3 Inference for the normal linear model

Distribution of $\hat\beta$

We already know the mean and variance of $\hat\beta$ (equations (1.1.1) and (1.1.2)). As it is a linear combination of ε, we know it must be normally distributed: $\hat\beta \sim N_p(\beta, \sigma^2(X^TX)^{-1})$.

Distribution of $\hat\sigma^2$

The maximum likelihood estimate for the σ^2 is

$$\frac{1}{n}\|Y - X\hat\beta\|^2 = \frac{1}{n}\|(I-P)Y\|^2 = \frac{1}{n}\|(I-P)\varepsilon\|^2.$$

We already know that the fitted values PY and residuals $(I-P)Y$ are uncorrelated. But $(PY, (I-P)Y)$ is a linear transformation of the multivariate normal Y, so PY and $(I-P)Y$ must be independent. Therefore $\hat\beta = (X^TX)^{-1}X^TPY$ and $\hat\sigma^2$ are independent. Proposition 4 shows that $\hat\sigma^2 \sim \sigma^2\chi^2_{n-p}/n$. Note that $\mathbb{E}(\hat\sigma^2) = (n-p)\sigma^2/n$, so $\hat\sigma^2$ is a biased estimator of σ^2. Let

$$\tilde\sigma^2 := \frac{n}{n-p}\hat\sigma^2 = \frac{1}{n-p}\|Y - X\hat\beta\|^2 \sim \frac{\sigma^2}{n-p}\chi^2_{n-p},$$

so $\tilde\sigma^2$ is now an unbiased estimator of σ^2.

Now that we know the joint distribution of $(\hat\beta, \tilde\sigma^2)$, it is rather easy to construct confidence sets for β.

Confidence statements for β

We can obtain confidence sets for β by using the fact that the quantity

$$\frac{\hat\beta - \beta}{\tilde\sigma} \qquad \left[= \frac{N_p(0, (X^TX)^{-1})}{\sqrt{\frac{1}{n-p}\chi^2_{n-p}}} \right\} \text{independent} \Bigg]$$

is a *pivot*, that is its distribution does not depend on β or σ^2. In fact it has a $t^{(p)}_{n-p}(0, (X^TX)^{-1})$ distribution. For example, observe that

$$\frac{\hat\beta_j - \beta_j}{\sqrt{\tilde\sigma^2 (X^TX)^{-1}_{jj}}} \sim t_{n-p},$$

so a $(1-\alpha)$-confidence interval for β_j is given by

$$\left[\hat{\beta}_j - \sqrt{\tilde{\sigma}^2 (X^T X)^{-1}_{jj}}\, t_{n-p}(\alpha/2),\ \hat{\beta}_j + \sqrt{\tilde{\sigma}^2 (X^T X)^{-1}_{jj}}\, t_{n-p}(\alpha/2)\right] =: C_j(\alpha).$$

Note that $\prod_{j=1}^p C_j(\alpha)$ does not constitute a $1-\alpha$ confidence cuboid for the entire parameter vector β, though $\prod_{j=1}^p C_j(\alpha/p)$ does have coverage at least $1-\alpha$ (see example sheet). However, the latter can have very large volume.

A confidence ellipsoid with much lower volume can be constructed by considering

$$\|X(\beta - \hat{\beta})\|^2 = \|P\varepsilon\|^2,$$

which has a $\sigma^2 \chi^2_p$ distribution (by Proposition 4) and is independent of $\tilde{\sigma}^2$. Thus

$$1 - \alpha = \mathbb{P}_{\beta,\sigma^2}\left(\frac{\frac{1}{p}\|X(\beta - \hat{\beta})\|^2}{\tilde{\sigma}^2} \leq F_{p,n-p}(\alpha) \right),$$

so

$$\left\{ b \in \mathbb{R}^p : \frac{\frac{1}{p}\|X(b - \hat{\beta})\|^2}{\tilde{\sigma}^2} \leq F_{p,n-p}(\alpha) \right\}$$

is a $(1-\alpha)$-level confidence set for β. One disadvantage of this confidence set is that it might be harder to interpret.

Of course the arguments used to arrive at the confidence intervals above can also be used to perform hypothesis tests of the form

$$\begin{aligned} H_0 &: \beta_j = \beta_{0,j} \\ H_1 &: \beta_j \neq \beta_{0,j}. \end{aligned}$$

and

$$\begin{aligned} H_0 &: \beta = \beta_0 \\ H_1 &: \beta \neq \beta_0. \end{aligned}$$

Prediction intervals

Given a new observation x^*, we can easily form a confidence interval for ${x^*}^T\beta$, the regression function at x^*, by noting that

$${x^*}^T(\hat{\beta} - \beta) \sim N(0, \sigma^2 {x^*}^T (X^T X)^{-1} x^*),$$

so

$$\frac{{x^*}^T(\hat{\beta} - \beta)}{\sqrt{\tilde{\sigma}^2 {x^*}^T (X^T X)^{-1} x^*}} \sim t_{n-p}.$$

A $(1-\alpha)$-level *prediction interval* for x^* is a random interval I depending only on Y such that $\mathbb{P}_{\beta,\sigma^2}(Y^* \in I) = 1 - \alpha$ where $Y^* := {x^*}^T\beta + \varepsilon^*$ and $\varepsilon^* \sim N(0, \sigma^2)$ independently of $\varepsilon_1, \ldots, \varepsilon_n$. This will be wider than the confidence interval for ${x^*}^T\beta$ as it must take into account the additional variability of ε^*. Indeed

$$Y^* - {x^*}^T\hat{\beta} = \varepsilon^* + {x^*}^T(\beta - \hat{\beta}) \sim N(0, \sigma^2\{1 + {x^*}^T (X^T X)^{-1} x^*\}),$$

so

$$\frac{Y^* - {x^*}^T\hat{\beta}}{\sqrt{\tilde{\sigma}^2\{1 + {x^*}^T (X^T X)^{-1} x^*\}}} \sim t_{n-p}.$$

The Bayesian normal linear model

So far we have treated β and σ^2 as unknown but fixed quantities. We have constructed estimators of these quantities and tried to understand how we expect them to vary under hypothetical repetitions of the experiment used to generate the data (with the design matrix fixed). This is a *frequentist* approach to inference.

A Bayesian approach instead treats unknown parameters as random variables, and examines their distribution conditional on the data observed. To fix ideas, suppose we have posited that the density of the r.v. representing our data $Y \in \mathbb{R}^n$ conditional on a parameter vector θ is $p(y|\theta)$. In addition to this statistical model, the Bayesian method requires that we agree on a marginal distribution for θ, $p(\theta)$. This can represent prior information about the parameters that is known before any of the data has been analysed, and hence it is called the *prior distribution.*

Inference about θ is based on the *posterior distribution*, $p(\theta|y)$, which satisfies

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{p(y)}.$$

Taking the mean or mode of the posterior distribution gives point estimates for θ. Note that in order to determine the posterior $p(\theta|y)$, we only need knowledge of the right-hand side up to multiplication by an arbitrary function of y, in particular it suffices to consider $p(y|\theta)p(\theta)$. To recover $p(\theta|y)$ we simply multiply by

$$\left(\int p(y|\theta')p(\theta')d\theta'\right)^{-1},$$

or alternatively we may be able to spot the form of the density for θ and find the normalising constant that way.

In contrast to frequentist confidence sets, using $p(\theta|y)$, we can construct sets S such that the posterior probability of $\{\theta \in S\}$ is at least $1-\alpha$. These are known as *credible sets.*

In the context of the Bayesian linear model, it is convenient to work with the precision $\omega := \sigma^{-2}$ rather than the variance. A commonly used prior for the parameters (β, ω) is $p(\beta, \omega) = \omega^{-1}$. This is not a density since it does not have a finite integral. Nevertheless, the posterior resulting from this prior is a genuine density, and inference based on this posterior has many similarities with inference in the frequentist context. To see this we first recall the gamma distribution.

The gamma distribution. If a random variable Z has density

$$f(z; a, b) = \frac{b^a z^{a-1} e^{-bz}}{\Gamma(a)} \text{ for } z \geq 0 \text{ and } a, b > 0,$$

we write $Z \sim \Gamma(a, b)$ and say Z has a gamma distribution with shape a and rate b. We note, for future use, that since the gamma density integrates to 1, we must have that

$$\int_{z=0}^{\infty} z^{a-1} e^{-bz} dz = \frac{\Gamma(a)}{b^a}. \tag{1.2.1}$$

Let us write the likelihood as

$$p(y|\beta, \omega) \propto \underbrace{\omega^{(n-p)/2} \exp\{-\omega\|(I-P)y\|^2/2\}}_{\propto\, \Gamma((n-p)/2+1,\, \|(I-P)y\|^2/2) \text{ density}} \; \underbrace{\omega^{p/2} \exp\{-\omega(\beta-\hat{\beta})^T X^T X(\beta-\hat{\beta})/2\}}_{\text{For fixed } \omega\, \propto\, N_p(\hat{\beta},\, \omega^{-1}(X^T X)^{-1}) \text{ density}}.$$

Then, multiplying by the prior, we see that the posterior is a product of gamma and normal densities: informally

$$p(\beta, \omega|y) = \underbrace{\Gamma((n-p)/2, \|(I-P)y\|^2)}_{p(\omega|y)} \times \underbrace{N_p(\hat{\beta}, \omega^{-1}(X^TX)^{-1})}_{p(\beta|\omega,y)}.$$

Thus $\beta|\omega, Y \sim N_p(\hat{\beta}, \omega^{-1}(X^TX)^{-1})$. Compare this to the distribution of $\hat{\beta}$ in the frequentist setting. The marginal posterior for β can be obtained by integrating out ω in the joint posterior above. Rather than performing the integration directly, we note that as a function of ω alone, the joint posterior is of the form

$$\omega^{A-1} \times \exp(-\omega B),$$

where

$$\begin{aligned} A &= \frac{n}{2} \\ B &= \frac{1}{2}\{\|(I-P)y\|^2 + \|X(\beta - \hat{\beta})\|^2\}. \end{aligned}$$

Thus by (1.2.1), we have that the marginal posterior for β satistfies

$$\begin{aligned} p(\beta|y) &\propto \int_{\omega=0}^{\infty} \omega^{A-1} \times \exp(-\omega B) \\ &\propto B^{-A} \\ &\propto \left(1 + \frac{\|X(\beta-\hat{\beta})\|^2}{\|(I-P)y\|^2}\right)^{-\{(n-p)/2+p/2\}} \\ &\propto \left(1 + \frac{1}{n-p}(\beta-\hat{\beta})^T(\tilde{\sigma}^{-2}X^TX)(\beta-\hat{\beta})\right)^{-\{(n-p)/2+p/2\}}, \end{aligned}$$

which we recognise as proportional to the density of a $t_{n-p}^{(p)}(\hat{\beta}, \tilde{\sigma}^2(X^TX)^{-1})$ distribution. Thus

$$\left.\frac{\beta - \hat{\beta}}{\tilde{\sigma}}\right| Y \;\sim\; t_{n-p}^{(p)}(0, (X^TX)^{-1}),$$

similarly to the frequentist case, though here it is β rather than $\hat{\beta}$ that is random. From this we see that

$$\begin{aligned} \left.\frac{\beta_j - \hat{\beta}_j}{\sqrt{\tilde{\sigma}^2(X^TX)^{-1}_{jj}}}\right| Y &\sim t_{n-p}, \\ \left.\frac{\|X(\beta-\hat{\beta})\|^2}{\tilde{\sigma}^2}\right| Y &\sim F_{p,n-p}, \end{aligned}$$

so the frequentist confidence regions described in earlier sections can also be thought of as Bayesian credible regions, when the prior $p(\beta, \omega) \propto \omega^{-1}$ is used.

Testing significance of groups of variables

Often we want to test whether a given group of variables is significant. Consider partitioning

$$X = (X_0 \; X_1) \qquad \text{and} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix},$$

where X_0 is n by p_0 and X_1 is n by $p-p_0$, and correspondingly $\beta_0 \in \mathbb{R}^{p_0}$ and $\beta_1 \in \mathbb{R}^{p-p_0}$. We are interesting in testing

$$\begin{aligned} H_0 &: \beta_1 = 0 \qquad \text{against} \\ H_1 &: \beta_1 \neq 0. \end{aligned}$$

One sensible way of proceeding is to construct a generalised likelihood ratio test. Recall that given an n-vector Y, assumed to have density $f(y;\theta)$ for some unknown $\theta \in \Theta$, the likelihood ratio test for testing

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \qquad \text{against} \\ H_1 &: \theta \notin \Theta_0, \end{aligned}$$

where $\Theta_0 \subset \Theta$, rejects the null hypothesis for large values of w_{LR} defined by

$$w_{\mathrm{LR}}(H_0) = 2\log\left\{\frac{\sup_{\theta'\in\Theta} L(\theta')}{\sup_{\theta'\in\Theta_0} L(\theta')}\right\} = 2\{\sup_{\theta'\in\Theta} \ell(\theta') - \sup_{\theta'\in\Theta_0} \ell(\theta')\}.$$

Let us apply the generalised likelihood ratio test to the problem of assigning significance to groups of variables in the linear model. Write $\check{\beta}_0$ and $\check{\sigma}^2$ for the MLEs of the vector of regression coefficients and the variance respectively under the null hypothesis (i.e. when the model is $Y = X_0\beta_0 + \varepsilon$ with $\varepsilon \sim N_n(0, \sigma^2 I)$).

We have

$$\begin{aligned} w_{\mathrm{LR}}(H_0) &= -n\log(\hat{\sigma}^2) - \frac{1}{\hat{\sigma}^2}\|Y - X\hat{\beta}\|^2 + n\log(\check{\sigma}^2) + \frac{1}{\check{\sigma}^2}\|Y - X_0\check{\beta}_0\|^2 \\ &= -n\log\left\{\frac{\|(I-P)Y\|^2}{\|(I-P_0)Y\|^2}\right\}. \end{aligned}$$

To determine the right cutoff for an α-level test, we need to obtain the distribution of (a monotone function of the) argument of the logarithm under the null hypothesis, that is, the distribution of

$$\frac{\|(I-P_0)\varepsilon\|^2}{\|(I-P)\varepsilon\|^2}.$$

By dividing top and bottom by σ^2, we see that the distribution of the quantity above doesn't depend on any unknown parameters. To find its distribution we argue as follows. Write

$$I - P_0 = (I - P) + (P - P_0).$$

Now since the columns of P and P_0 are in the column space of X, $(I-P)(P-P_0) = 0$, so

$$\|(I-P_0)\varepsilon\|^2 = \|(I-P)\varepsilon\|^2 + \|(P-P_0)\varepsilon\|^2,$$

whence

$$\frac{\|(I-P_0)\varepsilon\|^2}{\|(I-P)\varepsilon\|^2} = 1 + \frac{\|(P-P_0)\varepsilon\|^2}{\|(I-P)\varepsilon\|^2}.$$

Also

$$\mathrm{Cov}((I-P)\varepsilon, (P-P_0)\varepsilon) = \mathbb{E}\{(I-P)\varepsilon\varepsilon^T(P-P_0)^T\} = (I-P)(P-P_0) = 0.$$

As the random vector

$$\begin{pmatrix} (I-P)\varepsilon \\ (P-P_0)\varepsilon \end{pmatrix}$$

is multivariate normal (being the image of a multivariate normal vector under a linear map), we know that $(I-P)\varepsilon$ and $(P-P_0)\varepsilon$ are independent. Hence $\|(I-P)\varepsilon\|^2$ and $\|(P-P_0)\varepsilon\|^2$ are independent. We know that $\|(I-P)\varepsilon\|^2/\sigma^2 \sim \chi^2_{n-p}$. It turns out that $\|(P-P_0)\varepsilon\|^2/\sigma^2 \sim \chi^2_{p-p_0}$.

This follows from Proposition 4 and the fact that $P-P_0$ is an orthogonal projection with rank $p-p_0$. Indeed, it is certainly symmetric, and

$$(P-P_0)^2 = P - PP_0 - P_0P + P_0 = P - P_0,$$

the final equality following from $P_0P = P_0^TP^T = (PP_0)^T = P_0^T = P_0$. Thus $P-P_0$ is an orthogonal projection, so we know

$$\mathrm{r}(P-P_0) = \mathrm{tr}(P-P_0) = \mathrm{tr}(P) - \mathrm{tr}(P_0) = \mathrm{r}(P) - \mathrm{r}(P_0) = p - p_0.$$

Finally, we may conclude that

$$\frac{\frac{1}{p-p_0}\|(P-P_0)\varepsilon\|^2}{\frac{1}{n-p}\|(I-P)\varepsilon\|^2} \sim F_{p-p_0,n-p}.$$

In summary, we can perform a generalised likelihood ratio test for

$$\begin{aligned} &H_0 : \beta_1 = 0 \qquad \text{against} \\ &H_1 : \beta_1 \neq 0 \end{aligned}$$

at level α by comparing the test statistic

$$\frac{\frac{1}{p-p_0}\|(P-P_0)Y\|^2}{\frac{1}{n-p}\|(I-P)Y\|^2}$$

to $F_{p-p_0,n-p}(\alpha)$ and rejecting for large values of the test statistic.

1.2.4 ANOVA and ANCOVA

Although so far we have thought of our covariates as being real-valued (i.e. things like age, time, height, volume etc.), *categorical* predictors (also known as *factors*) can also be dealt with. These can arise in situations such as the following. Consider measuring the weight loss of people each participating in one of J different exercise regimes, the first regime being no exercise (the control). Let the weight loss of the k^{th} participant of regime j be Y_{jk}. The model that the responses are independent with

$$Y_{jk} \sim N(\mu_j, \sigma^2), \qquad j = 1, \dots, J;\ k = 1, \dots n_j$$

can be cast within the framework of the normal linear model by writing

$$Y = \begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ Y_{J1} \\ \vdots \\ Y_{Jn_J} \end{pmatrix} \; ; \; X = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ & & \vdots & & \\ 0 & \cdots & \cdots & 0 & 1 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix} \; ; \; \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_J \end{pmatrix}.$$

This type of model is known as a one-way **an**alysis **o**f variance (ANOVA). If all the n_j were equal, it would be called a balanced one-way ANOVA.

An alternative parametrisation is

$$Y_{jk} = \mu + \alpha_j + \varepsilon_{jk}, \qquad \varepsilon_{jk} \sim N(0, \sigma^2);\ j = 1, \ldots, J;\ k = 1, \ldots, n_j,$$

where μ is the baseline or mean effect and α_j is the effect of the j^{th} regime in relation to the baseline.

Notice that the parameter vector (μ, β) is not *identifiable* since, for example, replacing μ with $\mu + c$ and each α_j with $\alpha_j - c$ gives the same model for every $c \in \mathbb{R}$. To make the model identifiable, one option is to constrain $\alpha_1 = 0$. This is known as a corner point constraint and is the default in `R`. This makes it easier to test for differences from the control. Another option is to use a sum-to-zero constraint: $\sum_{j=1}^{J} n_j \alpha_j = 0$. Note that the particular constraints used do not affect the fitted values in any way.

If each of the subjects in our hypothetical experiment also went on one of I different diets, then writing Y_{ijk} now to mean the weight loss of the k^{th} participant of exercise regime j and diet i, we might model the Y_{ijk} as independent with

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}, \qquad \varepsilon_{ijk} \sim N(0, \sigma^2);\ i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, n_{ij}.$$

This model is called an additive two-way ANOVA because it assumes that the effects of the different factors are additive. The model is over-parametrised and as before, constraints must be imposed on the parameters to ensure identifiability. By default, `R` uses the corner point constraints $\alpha_1 = \beta_1 = 0$.

If the contribution of one of the exercise regimes to the response was not the same for all the different types of diets, it may be more appropriate to use the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}.$$

The γ_{ij} are known as *interaction* terms.

One might also have information about the subjects in the form of continuous variables, e.g. blood pressure, BMI. Since all the models above are normal linear models, these variables can simply be appended to the design matrix to include them in the model. A linear model that contains both factors and continuous variables is known as an **an**alysis of **cov**ariance (ANCOVA).

1.2.5 Model selection

Recall that the MSPE of $\hat{\beta}$ is $\sigma^2 p/n$. If only p_0 of the components of β were non-zero, say the first p_0, then we could perform regression on just X_0, the matrix formed from the first p_0 columns of X, and the resulting estimator of the non-zero coefficients, $\hat{\beta}_0$, would have reduced MSPE $\sigma^2 p_0/n$, rather than $\sigma^2 p/n$. Moreover

$$\mathrm{Var}(\hat{\beta}_{0,j}) = \frac{\sigma^2}{\|(I - P_{0,-j})X_j\|^2} \leq \frac{\sigma^2}{\|(I - P_{-j})X_j\|^2} = \mathrm{Var}(\hat{\beta}_j),$$

for $j = 1, \ldots, p_0$ (see example sheet). Here $P_{0,-j}$ is the orthogonal projection on to the column space of $X_{0,-j}$, the matrix formed by removing the j^{th} column from X_0.

It is thus useful to check whether a model formed from a smaller set of variables can adequately explain the data observed. Another advantage of selecting the right model is that it allows one to focus on variables of interest.

Coefficient of determination

One popular measure of the goodness of fit of a linear model is the *coefficient of determination* or R^2. It compares the residual sum of squares (RSS) under the model in question to a minimal model containing just an intercept, and is defined by

$$R^2 := \frac{\|Y - \bar{Y}1_n\|^2 - \|(I-P)Y\|^2}{\|Y - \bar{Y}1_n\|^2},$$

where 1_n is an n-vector of 1's. The interpretation of R^2 is as the proportion of the total variation in the data explained by the model. It takes values between 0 and 1 with higher values indicating a better fit. The R^2 will always increase if variables are added to the model. The adjusted R^2, $\tilde{R}^2$ defined by

$$\tilde{R}^2 := 1 - \frac{n-1}{n-p}(1-R^2)$$

can be motivated by analogy with the F statistic, and takes account of the number of parameters.

AIC

Another approach to measuring the fit of a model is Akaike's Information Criterion (AIC). We will describe AIC in a more general setting than the normal linear model, since it will be used when assessing the fit of generalised linear models which will be introduced in the next chapter.

Suppose that our data $(Y_1, x_1^T), \dots, (Y_n, x_n^T)$ are generated with Y_i independent conditional on the design matrix X whose rows are the x_i^T. Suppose that given x_i, the true pdf of Y_i is g_{x_i} and from a model $\mathcal{F} := \{(f_{x_i}(\cdot;\theta))_{i=1}^n, \theta \in \Theta \subseteq \mathbb{R}^p\}$ the corresponding maximum likelihood fitted pdf is $f_{x_i}(\cdot;\hat{\theta})$. One measure of the quality of $\hat{f}_{x_i}(\cdot) := f_{x_i}(\cdot;\hat{\theta})$ as an estimate of the true density g_{x_i} is the Kullback–Leibler divergence, $K(g_{x_i}, \hat{f}_{x_i})$ defined as

$$K(g_{x_i}, \hat{f}_{x_i}) := \int_{-\infty}^{\infty} [\log\{g_{x_i}(y)\} - \log\{\hat{f}_{x_i}(y)\}] g_{x_i}(y) dy.$$

For an overall measure of fit, we can consider

$$\bar{K} := \frac{1}{n}\sum_{i=1}^n K(g_{x_i}, \hat{f}_{x_i}).$$

One can show via Jensen's inequality that $\bar{K} \geq 0$ with equality if and only if each $g_{x_i} = \hat{f}_{x_i}$ (almost surely). Thus if $\bar{K}$ is low, we have a good fit. Given a collection of different fitted densities for the data, it is therefore desirable to select that which minimises $\bar{K}$. This is equivalent to minimising

$$\tilde{K} := -\frac{1}{n}\sum_{i=1}^n \int_{-\infty}^{\infty} \log\{\hat{f}_{x_i}(y)\} g_{x_i}(y) dy = -\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{Y_i^* \sim g_{x_i}}[\log\{\hat{f}_{x_i}(Y_i^*)\}|Y_i].$$

Of course, we cannot compute $\bar{K}$ or $\tilde{K}$ from the data since this requires knowledge of g_{x_i} for $i = 1, \dots, n$. However, it can be shown that it is possible to estimate $\mathbb{E}(\tilde{K})$ (where the expectation is over the randomness in the $\hat{f}_{x_i}$). Akaike's information criterion (AIC) defined as

$$\begin{aligned} \text{AIC} &:= -2\ell(\hat{\theta}) + 2p, \\ &= 2 \times \text{ (-maximised loglikelihood } + \text{ number of parameters in the model)} \end{aligned}$$

satisfies $\mathbb{E}(\mathrm{AIC})/n \approx 2\mathbb{E}(\tilde{K})$ for large n, provided the true densities g_{x_i}, $i = 1, \ldots, n$ are contained in the model $\mathcal{F}$.

In the normal linear model where X is n by p with full column rank, AIC amounts to

$$n\{1 + \log(2\pi\hat{\sigma}^2)\} + 2(p+1),$$

thus the best set of variables to use according to the AIC method is determined by minimising $n\log(\hat{\sigma}) + p$ across all candidate models.

Corrected information criterion

In fact we may form an unbiased estimate of $2n\mathbb{E}(\tilde{K})$ in the normal linear model. Suppose we have computed $\hat{\beta}$ from data Y generated by $Y = X\beta + \varepsilon$ with $\varepsilon \sim N_n(0, \sigma^2 I)$. Now let $Y^* = X\beta + \varepsilon*$ where $\varepsilon^* \sim N_n(0, \sigma^2 I)$ and ε^* and ε are independent. Then

$$\begin{aligned} 2n\mathbb{E}(\tilde{K}) &= \mathbb{E}\left\{\mathbb{E}\left(n\log(2\pi\hat{\sigma}^2) + \frac{\|Y^* - X\hat{\beta}\|^2}{\hat{\sigma}^2}\right)\middle| Y\right\} \\ &= \mathbb{E}\{n\log(2\pi\hat{\sigma}^2)\} + \mathbb{E}\left(\frac{n\sigma^2 + \|X\beta - X\hat{\beta}\|^2}{\hat{\sigma}^2}\right). \end{aligned}$$

Fact: If $Z \sim \chi^2_k$ with $k > 2$ then $\mathbb{E}(Z^{-1}) = (k-2)^{-1}$. Since $\hat{\sigma}^2$ and $\|X\beta - X\hat{\beta}\|^2 = \|P\varepsilon\|^2$ are independent, the second expectation in the display above equals

$$\frac{n(n+p)}{n-p-2},$$

provided $n > p + 2$. Thus an unbiased estimator of $2n\mathbb{E}(\tilde{K})$ is

$$n\log(2\pi\hat{\sigma}^2) + \frac{n(n+p)}{n-p-2}.$$

The corrected information criterion, AIC_c, is given by

$$\mathrm{AIC}_c = n\log(2\pi\hat{\sigma}^2) + n\frac{1+p/n}{1-(p+2)/n}.$$

Note that

$$\begin{aligned} n\frac{1+p/n}{1-(p+2)/n} &= n\left(1 + 2\frac{\frac{p+1}{n}}{1-\frac{p+2}{n}}\right) \\ &= n + 2(p+1)\frac{1}{1-\frac{p+2}{n}}. \end{aligned}$$

Thus when p/n is small, $\mathrm{AIC}_c \approx \mathrm{AIC}$ in the case of the normal linear model.

Orthogonality

One way to use the above model selection criteria is to fit each of the 2^{p-1} submodels that can be created using our design matrix (assuming we include an intercept every time and the first column of X is a column of 1's) and pick the one that seems best based on our criterion of choice. However, if p is reasonably large, this becomes a very computationally intensive task.

One situation where such an approach is feasible is when the columns of X are orthogonal. Indeed, more generally, if X can be partitioned as $X = (X_0 \; X_1)$ with the vector of coefficients correspondingly partitioned as $\beta = (\beta_0^T, \beta_1^T)^T$, we say that β_0 and β_1 are *orthogonal sets of parameters* if $X_0^T X_1 = 0$. Then

$$\begin{aligned}
\hat{\beta} &= \left(\begin{pmatrix} X_0^T \\ X_1^T \end{pmatrix} (X_0 \; X_1) \right)^{-1} \begin{pmatrix} X_0^T \\ X_1^T \end{pmatrix} Y \\
&= \begin{pmatrix} (X_0^T X_0)^{-1} & 0 \\ 0 & (X_1^T X_1)^{-1} \end{pmatrix} \begin{pmatrix} X_0^T \\ X_1^T \end{pmatrix} Y \\
&= \begin{pmatrix} (X_0^T X_0)^{-1} X_0^T Y \\ (X_1^T X_1)^{-1} X_1^T Y \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}.
\end{aligned}$$

If all the columns of X are orthogonal, we can easily find the best fitting model (in terms of the RSS) with p_0 variables. We simply order the $\|\hat{\beta}_j X_j\|^2 = X_j^T Y / \|X_j\|$ (excluding the intercept term) in decreasing order, and pick variables corresponding to the first p_0 terms. This works because letting X_S for $S \subseteq \{1, \ldots, p\}$ be the matrix formed from the columns of X indexed by S, and writing P_S for the projection on to the column space of X_S,

$$\|(I - P_S)Y\|^2 = \left\| Y - \sum_{j \in S} \hat{\beta}_j X_j \right\|^2 = \|Y\|^2 - \sum_{j \in S} \|\hat{\beta}_j X_j\|.$$

Exact orthogonality is of course unlikely to occur unless we have *designed* the design matrix X ourselves, either through choosing the values of the original covariates, or through transforming them in particular ways. A very common example of the latter is mean-centring each variable before adding an intercept term, so the intercept coefficient is then orthogonal to the rest of the coefficients.

Forward and backward selection

When the design matrix does not have orthogonal columns another strategy to avoid a search through all submodels is a *forward selection* approach.

Forward selection.

1. Start by fitting an intercept only model: call this S_0.
2. Add to the current model the predictor variable reduces the residual sum of squares the most.
3. Continue step 2 until all predictor variables have been chosen or until a large number of predictor variables has been selected. This produces a sequence of sub-models $S_0 \subset S_1 \subset S_2 \subset \cdots$.
4. Pick a model from the sequence of models created using either AIC or R^2 based criteria (or something better!).

An alternative is:

Backward selection.

1. Fit the largest model available (i.e. include all predictors) and call this S_0.
2. Exclude the predictor variable whose removal from the current model decreases the residual sum of squares the least.
3. Continue step 2 until all predictor variables have been removed (or a large number of predictor variables have been removed). This produces a sequence of submodels $S_0 \supset S_1 \supset S_2 \supset \cdots$.
4. Finally pick a model from the sequence as with forward selection.

Inference after model selection

Once a model has been selected, it is tempting to simply pretend that the variables in the submodel were the only ones that were ever collected and then proceed with constructing confidence intervals and using other inferential tools. *But this ignores the fact that the data has already been used to select the submodel.* Recall how we can imagine a $1-\alpha$ level confidence interval as being a particular construction of an interval that when applied to data generated through hypothetical repetitions of the "experiment" (keeping X fixed), gives intervals a proportion $1-\alpha$ of which we expect to contain the true parameter. However when the confidence intervals to be constructed are determined based on the response, we cannot interpret confidence intervals this way, because different responses would have led to different models being selected. The same issue arises for other inferential methods. This is a big problem in Statistics and currently the subject of a great deal of research in Statistics.

What can we do to combat this problem? One option is to divide the observations into two halves. One half can be used to pick the best model and then the other half to construct confidence intervals, p-values etc. However, because we are only using part of the data to perform inference, our procedures will lose power. Moreover different splits of the data will give different results (this is less of a problem since we can try to aggregate results in some way). An alternative is to try to perform model selection in a way such that for almost all datasets (i.e. realisations of the response Y), we expect the same submodel to be selected. In any case, inferences drawn after model selection must be reported with care: this is a tricky issue with no easy universally accepted solutions.

1.2.6 Model checking

The validity of the inferences drawn from the normal linear model rest on four assumptions.

(A1) $\mathbb{E}(\varepsilon_i) = 0$. If this is false, the coefficients in the linear model need to be interpreted with care. Furthermore, our estimate of σ^2 will tend to be inflated and F-tests may lose power though they will have the correct size (see example sheet).

(A2) $\mathrm{Var}(\varepsilon_i) = \sigma^2$. This assumption of constant variance is called *homoscedasticity*, and its violation (nonconstant variance) is called *heteroscedasticity*. A violation of this assumption means the least squares estimates are not as efficient as they could be, and furthermore hypothesis tests and confidence intervals need not have their nominal levels and coverages respectively. If the variances of the errors are known up to an unknown multiplicative constant, weighted least squares can be used (see example sheet).

(A3) $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$: the errors are uncorrelated. When data are ordered in time or space, this assumption is often violated. As with heteroscedasticity, the standard inferential techniques can give misleading results.

(A4) The errors ε_i are normally distributed. Though the confidence intervals and hypothesis tests we have studied rest on the assumption of normality, arguments based on the central limit theorem can be used to show that even when the errors are not normally distributed, provided (A1–A3) are satisfied, inferences are still asymptotically valid under reasonable conditions.

A useful way of assessing whether the assumptions above are satisfied is to analyse the residuals $\hat{\varepsilon} := (I - P)Y$ arising from the model fit. This is usually done graphically rather than through formal tests. An advantage of the graphical approach is that we can look for many different signs for departures from the assumptions simultaneously. One potential issue is that it may not always be clear what indicates a genuine violation of assumptions compared to the natural variation that one should expect even if the assumptions held.

Note that under (A1), $\mathbb{E}(\hat{\varepsilon}) = 0$. It is common to plot the residuals against the fitted values $\hat{Y}_i$, and also against each of the variables in the design matrix (including those not in the current model). If (A1) holds, there should not be an obvious trend in the mean of the residuals.

Under (A2) and (A3), $\text{Var}(\hat{\varepsilon}) = \sigma^2(I - P)$. Define the *studentised residuals* to be

$$\hat{\eta}_i := \frac{\hat{\varepsilon}_i}{\tilde{\sigma}\sqrt{1 - p_i}}, \qquad \text{where } p_i := P_{ii} \quad i = 1, \ldots, n.$$

Provided $\tilde{\sigma}$ is a good estimate of σ, the variance of $\hat{\eta}_i$ should be approximately 1. A standard check of the validity of (A2) involves plotting $\sqrt{|\hat{\eta}_i|}$ against the fitted values.

If (A1–A4) hold, then we'd expect the $\hat{\eta}_i$ to look roughly like an i.i.d. sample from a $N(0, 1)$ distribution since

$$\hat{\eta}_i \approx \frac{\hat{\varepsilon}_i}{\sigma\sqrt{1 - p_i}}$$

and so

$$\text{Cov}(\hat{\eta}_i, \hat{\eta}_j) \approx \frac{-P_{ij}}{\sqrt{(1 - p_i)(1 - p_j)}},$$

for $i \neq j$. When $n \gg p$ We expect this covariance to be close to 0 because

$$\frac{1}{n^2}\sum_{i,j} P_{ij}^2 = \frac{1}{n^2}\text{tr}(P^T P) = \frac{1}{n^2}\text{tr}(P) = \frac{1}{n^2}\text{r}(P) = \frac{p}{n^2},$$

(so the average of the squared entries of P is small).

A good way of checking that the $\hat{\eta}_i$ look roughly standard normal is to look at a *Quantile–Quantile* (Q–Q) plot. This involves plotting the order statistics of the sample of $\hat{\eta}_i$ against the expected order statistics of the normal distribution. Since the latter are rather complicated to compute, we often approximate the expected value of the i^{th} order statistic $Z_{(i)}$ from a sample of i.i.d. standard normal random variables $Z_1, \ldots, Z_n$, by

$$\mathbb{E}(Z_{(i)}) \approx \Phi^{-1}\left(\frac{i}{n + 1}\right).$$

In summary, we

1. sort the studentised residuals, $\hat{\eta}_1, \ldots, \hat{\eta}_n$ into into increasing order, and

2. plot them against $\{\Phi^{-1}(\frac{i}{n+1}) : i = 1, \ldots, n\}$.

We expect an approximately straight line through the origin with gradient 1 if our normality assumption is correct.

Variable transformations

We have already discussed how predictors may be transformed so that models that are nonlinear in the original data (but linear in the parameter β) still fall within the linear model framework. Sometimes it can also be helpful to transform the response so that it fits the linear model. Consider the following model

$$Y_i = \exp(x_i^T\beta + \varepsilon_i), \qquad i = 1, \ldots, n,\ \varepsilon \sim N_n(0, \sigma^2 I).$$

If we make the transformation $Y_i \mapsto \log(Y_i)$ we will have a linear model in the logged response.

The Box–Cox family of transformations is given by

$$y \mapsto y^{(\lambda)} := \begin{cases} \dfrac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \\ \log(y) & \text{if } \lambda = 0. \end{cases}$$

Typically one plots the log-likelihood of the transformed data $(y_1^{(\lambda)}, \ldots, y_n^{(\lambda)})$ as a function of λ and then selects a value of λ which lies close to the λ that maximises the log-likelihood, and still gives a model with interpretable parameters.

Unusual observations

Often we may find that though the bulk of our data satisfy the assumptions (A1–A4) and fit the model well, there are a few observations that do not. These are called outliers. It is important to detect these so that they can be excluded when fitting the model, if necessary. A more subtle way in which an observation can be unusual is if it is unusual in the predictor space i.e. it has an unusual x value; it is this we discuss first.

Leverage. Recall that the fitted values $\hat{Y}$ satisfy

$$\hat{Y}_i = (PY)_i = P_{i1}Y_1 + \cdots + P_{ii}Y_i + \cdots + P_{in}Y_n.$$

The value $p_i := P_{ii}$ is called the leverage of the i^{th} observation. It measures the contribution that Y_i makes to the fitted value $\hat{Y}_i$. It can be shown that $0 \leq p_i \leq 1$. Since $\text{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - p_i)$, values of p_i close to 1 force the regression line (or plane) to pass very close to Y_i.

The idea of leverage is about the *potential* for an observation to have a large effect on the fit; if the observation does not have an unusual response value, it is possible that removing the observation will change the estimated regression coefficients very little. However in this case, the R^2 and the results of an F-test with the null hypothesis as the intercept only model may still change a lot.

The relationship $\sum_{i=1}^n p_i = \text{tr}(P) = p$ motivates a rule of thumb that says the influence of the i^{th} observation may be of concern if $p_i > 3p/n$. When the design matrix consists of just a single variable and a column of 1's representing an intercept term (as the first column), it can be shown that

$$p_i = \frac{1}{n} + \frac{(X_{i2} - \bar{X}_2)^2}{\sum_{k=1}^n (X_{k2} - \bar{X}_2)^2},$$

where $\bar{X}_2 := \frac{1}{n}\sum_{i=1}^n X_{i2}$.

Cook's distance. The Cook's distance D_i of the observation (Y_i, x_i) is defined as

$$D_i := \frac{\frac{1}{p}\|X(\hat{\beta}_{(-i)} - \hat{\beta})\|^2}{\tilde{\sigma}^2},$$

where $\beta_{(-i)}$ is the OLS estimate of β when omitting observation (Y_i, x_i).

The interpretation of Cook's distance is that if $D_i = F_{p,n-p}(\alpha)$ then omitting the i^{th} data point moves the m.l.e. of β to the edge of the $(1-\alpha)$-level confidence set for β.

Note that we do not need to fit $n+1$ linear models to compute all of the Cook's distances, since in fact

$$D_i = \frac{1}{p}\frac{p_i}{1-p_i}\hat{\eta}_i^2 \qquad \text{(see example sheet).}$$

Thus Cook's distance combines the studentised fitted residuals with the leverage as a measure of influence. A rule of thumb is that we should be concerned about the influence of (Y_i, x_i) if $D_i > F_{p,n-p}(0.5)$.

Chapter 2

Exponential families and generalised linear models

2.1 Non-normal responses

Suppose we are interested in predicting the probability that an internet advert gets clicked by web surfers visiting the page where it is displayed, based on it's colour, size, position, font used and other information. Given a vector of responses $Y \in \{0,1\}^n$ (1 = 'clicked' and 0 = 'didn't click') and a design matrix X collecting together the relevant information, a linear model would attempt to find $\hat{\beta}$ such that Y and $X\hat{\beta}$ are close. However the fitted values do not relate well to probabilities that $Y_i = 1$: indeed there is no guarantee that we even have $X\hat{\beta} \in [0,1]^n$.

Really, we would like to model

$$Y_i \sim \text{Bin}(\mu_i, 1),$$

with μ_i related to some function of the predictors x_i whose range is contained in $[0,1]$.

Generalised linear models (GLMs) extend linear models to deal with situations such as that discussed above. We can think of a normal linear model as consisting of three components.

(i) The random component: $Y_1, \ldots, Y_n$ are independent normal random variables, with Y_i having mean μ_i and variance σ^2.

(ii) The systematic component: a *linear predictor* $\eta = (\eta_1, \ldots, \eta_n)^T$, where $\eta_i = x_i^T \beta$.

(iii) The *link* between the random and systematic components: $\mu_i = \eta_i$.

Of course this is an unnecessarily wasteful way to write out the linear model, but it is suggestive of generalisations.

GLMs extend linear models in (i) and (iii) above, allowing different classes of distributions for the response variables and allowing a more general link:

$$\eta_i = g(\mu_i)$$

where g is a strictly increasing, twice differentiable function.

2.2 Exponential families

We want to consider a class of distributions large enough to include the normal, binomial and other familiar distributions, but which is still relatively simple, both conceptually and computationally.

Why is this a useful endeavour? We could just work with a particular family of distributions for the response that is useful for our own purposes, and develop algorithms for estimating parameters and theory for the distributions of our estimates (just as we did for the normal linear model). However, if we work in a more general framework, there we may be able to formulate inference procedures and develop computational techniques that are applicable for a number of families of distributions.

We begin our quest for such a general framework with the concept of an *exponential family.* We motivate the idea by starting with a single density or probability mass function $f_0(y)$, $y \in \mathcal{Y} \subseteq \mathbb{R}$. Rather than always writing "density or probability mass function", we will use the term "model function" to mean either a density function or p.m.f. (Of course, those of you who attended Probability and Measure will know that p.m.f.'s are just densities with respect to counting measure, so we could equally well use "density" throughout).

We will require that f_0 be a non-degenerate model function, that is if Y has model function f_0 then $\mathrm{Var}(Y) > 0$. For example, $f_0(y)$ might be the uniform density on the unit interval $\mathcal{Y} = [0, 1]$, or might have the probability mass function $y(1 - y)$ on $\mathcal{Y} = \{0, 1\}$.

We can generate a whole family of model functions based on f_0 via *exponential tilting*:

$$f(y; \theta) = \frac{e^{y\theta} f_0(y)}{\int e^{y'\theta} f_0(y')dy'}, \qquad y \in \mathcal{Y}.$$

We can only consider values of θ for which the integral in the denominator is finite. Note that the denominator is precisely the moment generating function of f_0 evaluated at θ. Let us briefly recall some facts about moment generating functions before proceeding.

The moment and cumulant generating functions. The moment generating function (m.g.f.) of a random variable, or equivalently its model function, is $M(t) := \mathbb{E}(e^{tY})$. The cumulant generating function (c.g.f.) is the logarithm of the m.g.f.: $K(t) := \log(M(t))$. The set of values where these functions are finite is an interval containing 0. If this contains an open interval about 0, then we have the series expansions

$$\begin{aligned} M(t) &= \sum_{r=0}^{\infty} \mathbb{E}(Y^r) \frac{t^r}{r!}, \\ K(t) &= \sum_{r=0}^{\infty} \kappa_r \frac{t^r}{r!}, \end{aligned}$$

where κ_r is known as the r^{th} cumulant. Standard theory about power series tells us that

$$\begin{aligned} \mathbb{E}(Y^r) &= M^{(r)}(0), \\ \kappa_r &= K^{(r)}(0). \end{aligned}$$

Check that $\kappa_1 = \mathbb{E}(Y)$ and $\kappa_2 = \mathrm{Var}(Y)$.

Let K now be the c.g.f. of f_0 and suppose $\Theta := \{\theta : K(\theta) < \infty\}$ is an open interval containing 0. Then the class of model functions

$$\{f(y; \theta) : \theta \in \Theta\}$$

is called the *natural exponential family* (of order 1) generated by f_0, and is an example of an exponential family. With a different generating model function f_0, we can get a different exponential family.

The parameter θ is called the *natural parameter* and Θ is called the *natural parameter space*. Note that we may write $f(y;\theta) = e^{\theta y - K(\theta)} f_0(y)$ so $\int_{\mathcal{Y}} e^{\theta y - K(\theta)} f_0(y) = 1$ for all $\theta \in \Theta$.

The mean of $f(y;\theta)$ is of course related to the parameter θ, and it is often useful to reparametrise the family of model functions in terms of their means. To discuss this, let us first find the mean and variance of $f(y;\theta)$, i.e. the first and second cumulants.

The m.g.f. of $f(\cdot;\theta)$, $M(t;\theta)$ is

$$\begin{aligned} M(t;\theta) &= \int_{\mathcal{Y}} e^{ty} e^{\theta y - K(\theta)} f_0(y) dy \\ &= e^{K(\theta+t) - K(\theta)} \int_{\mathcal{Y}} e^{(\theta+t)y - K(\theta+t)} f_0(y) dy \\ &= e^{K(\theta+t)-K(\theta)}, \qquad \text{for } \theta, \theta + t \in \Theta. \end{aligned}$$

Thus if Y has a $f(y;\theta)$ model function, then

$$\mathbb{E}_\theta(Y) = \frac{d}{dt} K(t;\theta)\Big|_{t=0} = K'(\theta), \qquad \mathrm{Var}_\theta(Y) = \frac{d^2}{dt^2} K(t;\theta)\Big|_{t=o} = K''(\theta).$$

It can be shown that since f_0 was assumed to be non-degenerate, so must be every $f(y;\theta)$. Then

$$\begin{gathered} \mu(\theta) := \mathbb{E}_\theta(Y) = K'(\theta) \qquad \text{satisfies} \\ \mu'(\theta) = K''(\theta) > 0 \end{gathered}$$

so μ is a smooth, strictly increasing function from Θ to $\mathcal{M} := \{\mu(\theta) : \theta \in \Theta\}$ ($\mathcal{M}$ for 'mean space'), with inverse function $\theta := \theta(\mu)$. This leads to the *mean value parametrisation*:

$$f(y;\mu) = e^{\theta(\mu)y - K(\theta(\mu))} f_0(y), \qquad y \in \mathcal{Y},\ \mu \in \mathcal{M}.$$

The function $V : \mathcal{M} \to (0,\infty)$ defined by $V(\mu) = \mathrm{Var}_{\theta(\mu)}(Y) = K''(\theta(\mu))$ is called the variance function.

Examples.

1. Let $f_0 = \phi$, the standard normal density. Then $M(\theta) = e^{\theta^2/2}$, $\theta \in \mathbb{R}$ so $K(\theta) = \frac{1}{2}\theta^2$. Thus the natural exponential family generated by the standard normal density is

   $$f(y;\theta) = e^{\theta y - \theta^2/2} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} = \frac{1}{\sqrt{2\pi}} e^{-(y-\theta)^2/2}, \qquad y \in \mathbb{R},\ \theta \in \mathbb{R}.$$

   This is the $N(\theta,1)$ family. Clearly $\mu(\theta) = \theta$, $\theta(\mu) = \mu$, $\mathcal{M} = \mathbb{R}$ and $V(\mu) = 1$, as can be verified by taking derivatives of $K(\theta)$.

2. Let f_0 denote the Pois(1) p.m.f.:

   $$f_0(y) = e^{-1} \frac{1}{y!}, \qquad y \in \{0,1,\ldots\}.$$

   Then

   $$M(\theta) = e^{-1} \sum_{r=0}^{\infty} \frac{e^{\theta r}}{r!} = \exp(e^\theta - 1).$$

   Thus with exponential tilting, we get

   $$f(y;\theta) = e^{\theta y - \exp(\theta)} \frac{1}{y!} = \frac{(e^\theta)^y \exp(-e^\theta)}{y!}, \qquad y \in \{0,1,\ldots\},\ \theta \in \mathbb{R}.$$

   This is the Pois(e^θ) family of distributions. The mean function is $\mu = e^\theta$ with inverse $\theta = \log(\mu)$, and the variance function, $V(\mu) = \mu$; the mean space is $\mathcal{M} = (0,\infty)$.

Technical conditions. Why did we impose the technical conditions that the set of values where the c.g.f. of f_0 is finite, Θ, is an open interval containing 0? Note that then given any $\theta \in \Theta$, $\{t : \theta + t \in \Theta\} = \Theta - \theta$ is an open interval containing 0. Thus the result we have shown that

$$K(\theta + t) - K(\theta)$$

is the c.g.f. of $f(\cdot;\theta)$ is valid for all $t \in \Theta - \theta$, and as this has a power series expansion, we can recover the cumulants by taking derivatives an evaluating at 0.

2.3 Exponential dispersion families

The natural exponential families are not broad enough for our purposes. We should like more control over the variance. A family of model functions of the form

$$f(y;\theta,\sigma^2) = a(\sigma^2, y)\exp\left[\frac{1}{\sigma^2}\{\theta y - K(\theta)\}\right], \qquad y \in \mathcal{Y}, \theta \in \Theta, \sigma^2 \in \Phi \subseteq (0,\infty), \tag{2.3.1}$$

where

- $a(\sigma^2, y)$ is a known positive function (c.f. $f_0(y)$ that generated the exponential family),
- Θ is an open interval,

and in addition the model functions are non-degenerate, is called an *exponential dispersion family* (of order 1). The parameter σ^2 is called the *dispersion parameter.* (Note many authors simply call the family of model functions in (2.3.1) an example of an exponential family.)

Let $K(\cdot;\theta,\sigma^2)$ be the c.g.f. of the model function $f(y;\theta,\sigma^2)$ in (2.3.1). It can be shown (see example sheet) that the c.g.f. of the density in (2.3.1) is

$$K(t;\theta,\sigma^2) = \frac{1}{\sigma^2}\{K(\sigma^2 t + \theta) - K(\theta)\},$$

for $\theta + \sigma^2 t \in \Theta$. Since the set of values where $K(\cdot;\theta,\sigma^2)$ is finite contains an open interval about 0, if Y has model function (2.3.1) then

$$\mathbb{E}_{\theta,\sigma^2}(Y) = K'(\theta), \qquad \mathrm{Var}_{\theta,\sigma^2}(Y) = \sigma^2 K''(\theta).$$

As before, we may define $\mu(\theta) := K'(\theta)$. Since $\mathrm{Var}_{\theta,\sigma^2}(Y) > 0$ (by non-degeneracy of the model functions), $K''(\theta) > 0$, so we can define an inverse function to μ, $\theta(\mu)$. Further define $\mathcal{M} := \{\mu(\theta) : \theta \in \Theta\}$ and variance function $V : \mathcal{M} \to (0,\infty)$ given by $V(\mu) := K''(\theta(\mu))$ (though now the variance of the model function is actually $\sigma^2 V(\mu)$).

Examples.

1. Consider the family $N(\nu, \tau^2)$ where $\nu \in \mathbb{R}$ and $\tau^2 \in (0,\infty)$. We may write the densities as
$$f(y;\nu,\tau^2) = \frac{1}{\sqrt{2\pi\tau^2}}\exp\left(-\frac{y^2}{2\tau^2}\right)\exp\left\{\frac{1}{\tau^2}\left(\nu y - \frac{1}{2}\nu^2\right)\right\},$$
showing that this family is an exponential dispersion family with $\theta = \nu$, $\mu(\theta) = \theta$, $\sigma^2 = \tau^2$. Of course we know that $\mu(\theta) = \theta$ and $V(\mu) = 1$, but we can also check this by differentiating $K(\theta) = \theta^2/2$.

2. Let $Z \sim \text{Bin}(n,p)$. Then $Y := Z/n \sim \frac{1}{n}\text{Bin}(n,p)$ has p.m.f.

$$f(y;p) = \binom{n}{ny} p^{ny}(1-p)^{n(1-y)}, \qquad y \in \{0, 1/n, 2/n, \dots, 1\}$$

Consider the family of p.m.f.'s of the form above with $p \in (0,1)$ and $n \in \mathbb{N}$. To show this is an exponential dispersion family, we write

$$\begin{aligned} f(y;p) &= \exp\left\{ny\log\left(\frac{p}{1-p}\right) + n\log(1-p)\right\}\binom{n}{ny} \\ &= \exp\left\{\frac{y\theta - \log(1+e^\theta)}{\sigma^2}\right\}\binom{1/\sigma^2}{y/\sigma^2}, \end{aligned}$$

with $\sigma^2 = 1/n$, $\theta = \log\{p/(1-p)\}$ and $K(\theta) = \log(1+e^\theta)$. To find the mean function $\mu(\theta)$, we differentiate K

$$\mu(\theta) = \frac{d}{d\theta}\log(1+e^\theta) = \frac{e^\theta}{1+e^\theta} \quad (= p),$$

with inverse $\theta(\mu) = \log\{\mu/(1-\mu)\}$. Differentiating once more we see that

$$\begin{aligned} V(\mu) &= \frac{(1+e^{\theta(\mu)})e^{\theta(\mu)} - (e^{\theta(\mu)})^2}{(1+e^{\theta(\mu)})^2} \\ &= \frac{e^{\theta(\mu)}}{1+e^{\theta(\mu)}}\left(1 - \frac{e^{\theta(\mu)}}{1+e^{\theta(\mu)}}\right) \\ &= \mu(1-\mu). \end{aligned}$$

Here $\mathcal{M} = (0,1)$ and $\Phi = \mathbb{N}$.

3. Consider the gamma family of densities,

$$f(y;\alpha,\lambda) = \frac{\lambda^\alpha y^{\alpha-1}e^{-\lambda y}}{\Gamma(\alpha)} \text{ for } y > 0 \text{ and } \alpha, \lambda > 0.$$

It is not immediately clear how to write this in exponential dispersion family form, so let us take advantage of the fact that we know the mean and variance of a gamma distribution. If Y has the gamma density then $\mathbb{E}_{\alpha,\lambda}(Y) = \alpha/\lambda$ and $\text{Var}_{\alpha,\lambda}(Y) = \alpha/\lambda^2$. If this family were an exponential dispersion family then $\mu = \alpha/\lambda$ and $\sigma^2 V(\mu) = \alpha/\lambda^2$. It is not clear what we should take as σ^2. However, the $y^{\alpha-1}$ term would need to be absorbed by the $a(y,\sigma^2)$ in the definition of the EDF. Thus we can try taking σ^2 as a function of α alone. What function must this be? Imagine that $\alpha = \lambda$, so $\sigma^2 V(\mu) = \sigma^2 \times \text{constant} \propto 1/\lambda = 1/\alpha$. Thus we must have $\sigma^2 = \alpha^{-1}$ (or some constant multiple of it). In the new parametrisation where $\alpha = \sigma^{-2}$ and $\lambda = (\mu\sigma^2)^{-1}$

$$\begin{aligned} f(y;\mu,\sigma^2) &= \frac{y^{\sigma^{-2}-1}\exp(-\frac{y}{\sigma^2\mu})}{(\sigma^2\mu)^{\sigma^{-2}}\Gamma(\sigma^{-2})} \\ &= \frac{y^{\sigma^{-2}-1}}{(\sigma^2)^{\sigma^{-2}}\Gamma(\sigma^{-2})}\exp\left\{\frac{1}{\sigma^2}\left(-\frac{y}{\mu} - \log\mu\right)\right\} \\ &= \frac{y^{\sigma^{-2}-1}}{(\sigma^2)^{\sigma^{-2}}\Gamma(\sigma^{-2})}\exp\left[\frac{1}{\sigma^2}\{y\theta - K(\theta)\}\right], \end{aligned}$$

where $\theta(\mu) = -\mu^{-1}$ and $K(\theta) = \log(-\theta^{-1})$. We found the variance function to be $V(\mu) = \mu^2$ and both $\mathcal{M}$ and Φ and $(0,\infty)$.

2.4 Generalised linear models

Having finally defined the concept of an exponential dispersion family, we can now define what a generalised linear model is. A generalised linear model for observations $(Y_1, x_1), \dots, (Y_n, x_n)$ is defined by the following properties.

1. $Y_1, \dots, Y_n$ are independent, each Y_i having model function in the same exponential dispersion family of the form

$$f(y; \theta_i, \sigma_i^2) = a(\sigma_i^2, y) \exp\left[\frac{1}{\sigma_i^2}\{\theta_i y - K(\theta_i)\}\right], \qquad y \in \mathcal{Y}, \theta_i \in \Theta, \sigma_i^2 \in \Phi \subseteq (0, \infty),$$

   with $\sigma_i^2 = \sigma^2 a_i$ where $a_1, \dots, a_n$ are known and $a_i > 0$, though σ^2 may be unknown. Note that the functions a and K must be fixed for all i.

2. The mean μ_i of the i^{th} observation and the i^{th} component of the linear predictor $\eta_i := x_i^T \beta$ are linked by the equation

$$g(\mu_i) = \eta_i, \qquad i = 1, \dots, n,$$

   where g is a strictly increasing, twice differentiable function called the *link function*.

2.4.1 Choice of link function

Note that the only allowable values of β are those such that $g^{-1}(x_i^T \beta)$ is in the mean space $\mathcal{M}$ of the exponential dispersion family. Allowing the non-identity link function is particularly useful when $\mathcal{M}$ does not coincide with $\mathbb{R}$, as for the Poisson, gamma and binomial model functions. This is because if we choose g to map $\mathcal{M}$ to the whole real line, then no restriction needs to be placed on β.

For example, if we had

$$Y_i \sim \frac{1}{n_i} \text{Bin}(n_i, \mu_i),$$

then we know that $\mathcal{M} = (0, 1)$. A popular choice for g in this situation is the logit function: $g(\mu) = \log\{\mu/(1-\mu)\}$.

Recall the function $\theta(\mu)$ from the definition of the exponential dispersion family (the inverse of the mean function). The choice

$$g(\mu) = \theta(\mu)$$

is called the *canonical link function*. In view of this, we may also refer to the function $\theta(\mu)$ as the canonical link function. The logit function is the canonical link when the Y_i have scaled binomial distributions as above.

2.4.2 Likelihood equations

A sensible way to estimate β in a generalised linear model is using maximum likelihood. The likelihood for a generalised linear model is

$$L(\beta, \sigma^2; y_1, \dots, y_n) = \exp\left\{\sum_{i=1}^n \frac{1}{\sigma^2 a_i}[\theta(\mu_i) y_i - K(\theta(\mu_i))]\right\} \prod_{i=1}^n a(\sigma^2 a_i, y_i),$$

where $\theta(\mu_i) = \theta(g^{-1}(x_i^T \beta))$.

Using the canonical link function can simplify some calculations. With g the canonical link function, $\theta(\mu_i) = x_i^T\beta$, so we have log-likelihood

$$\ell(\beta, \sigma^2; y_1, \ldots, y_n) = \sum_{i=1}^n \frac{1}{\sigma^2 a_i}\{y_i x_i^T\beta - K(x_i^T\beta)\} + \sum_{i=1}^n \log\{a(\sigma^2 a_i, y_i)\}.$$

One feature of the log-likelihood above that makes it particularly easy to maximise over β is that the Hessian is negative semi-definite so the log-likelihood is is a concave function of β (for any fixed σ^2). We have

$$\begin{aligned}\frac{\partial \ell(\beta,\sigma^2)}{\partial\beta} &= \sum_{i=1}^n \frac{x_i}{\sigma^2 a_i}\{y_i - K'(x_i^T\beta)\} \\ \frac{\partial^2 \ell(\beta,\sigma^2)}{\partial\beta\partial\beta^T} &= -\sum_{i=1}^n \frac{x_i x_i^T}{\sigma^2 a_i} K''(x_i^T\beta),\end{aligned}$$

and $K'' > 0$. This in particular means that as a function of β, the log-likelihood cannot have multiple local maxima. Indeed, we know that if a maximiser of the log-likelihood, $\hat{\beta}$ exists, it must satisfy

$$\frac{\partial}{\partial\beta}\ell(\beta,\sigma^2)\bigg|_{\beta=\hat{\beta}} = 0. \tag{2.4.1}$$

However due to concavity of the log-likelihood, the converse is also true: if $\hat{\beta}$ satisfies (2.4.1) then it must maximise the log-likelihood. Indeed, for any β_0, consider the function

$$f(t) := \ell(\hat{\beta} + t(\beta_0 - \hat{\beta}), \sigma^2).$$

Note that $f(0) = \ell(\hat{\beta}, \sigma^2)$ and $f(1) = \ell(\beta_0, \sigma^2)$. A Taylor expansion of f about 0 gives us

$$f(1) = f(0) + f'(0) + \frac{1}{2}f''(t)$$

for some $t \in [0, 1]$ (note this is a Taylor expansion with a "mean-value" form of the remainder). Noting that $f'(0) = 0$ by assumption,

$$f(1) - f(0) = \frac{1}{2}(\beta_0 - \hat{\beta})^T \frac{\partial^2 \ell(\beta,\sigma^2)}{\partial\beta\partial\beta^T}\bigg|_{\beta=\tilde{\beta}} (\beta_0 - \hat{\beta}) \leq 0,$$

where $\tilde{\beta} := \hat{\beta} + t(\beta_0 - \hat{\beta})$.

2.5 Inference

Having generalised the normal linear model, how do we compute maximum likelihood estimators and how can we perform inference (i.e. construct confidence sets, perform hypothesis test)? These tasks were fairly simple in the normal linear model setting since the maximum likelihood estimator had an explicit form. In our more general setting, this will not (necessarily) be the case. Despite this, we can still perform inference and compute m.l.e.'s, but approximations must be involved in both of these tasks. We first turn to the problem of inference.

2.5.1 The score function

Consider data $(Y_1, x_1^T), \ldots, (Y_n, x_n^T)$ with the Y_i independent given the x_i, and suppose $Y = (Y_1, \ldots, Y_n)^T$ has density in

$$\{f(y, \theta),\ y \in \mathcal{Y}^n : \theta \in \Theta \subseteq \mathbb{R}^d\} = \left\{ \prod_{i=1}^n f_{x_i}(y_i; \theta),\ y_i \in \mathcal{Y} : \theta \in \Theta \subseteq \mathbb{R}^d \right\}.$$

We will review some theory associated with maximum likelihood estimators in this setting. Here we simply aim to sketch out the main results; for a rigorous treatment see your Principles of Statistics notes (or borrow someone's). In particular, we do not state all the conditions required for the results to be true (broadly known as "regularity conditions"), but they will all be satisfied for the generalised linear model setting to which we wish to apply the results.

Let $\hat{\theta}$ be the maximum likelihood estimator of θ (assuming it exists and is unique). If we cannot write down the explicit form of $\hat{\theta}$ as a function of the data, in order to study its properties, we must argue from what we do know about the m.l.e.—the fact that it maximises the likelihood, or equivalently the log-likelihood. This means $\hat{\theta}$ satisfies

$$\left. \frac{\partial}{\partial \theta} \ell(\theta; Y) \right|_{\theta = \hat{\theta}} = 0,$$

where

$$\ell(\theta; Y) = \log f(Y; \theta) = \sum_{i=1}^n \log f_{x_i}(Y_i; \theta).$$

We call the vector of partial derivatives of the likelihood the *score function*, $U(\theta; Y)$:

$$U_r(\theta; Y) := \frac{\partial}{\partial \theta_r} \ell(\theta; Y).$$

Two key features of the score function are that provided the order of differentiation w.r.t. a component of θ and integration over the sample space $\mathcal{Y}^n$ may be interchanged,

1. $\mathbb{E}_\theta\{U(\theta; Y)\} = 0$,
2. $\mathrm{Var}_\theta\{U(\theta; Y)\} = -\mathbb{E}_\theta\left(\frac{\partial^2}{\partial \theta \partial \theta^T} \ell(\theta; Y) \right)$.

To see the first property, note that for $r = 1, \ldots, d$,

$$\begin{aligned} \mathbb{E}_\theta\{U_r(\theta; Y)\} &= \int_{\mathcal{Y}^n} \frac{\partial}{\partial \theta_r} \log\{f(y; \theta)\} f(y; \theta) dy \\ &= \int_{\mathcal{Y}^n} \frac{\partial}{\partial \theta_r} f(y; \theta) dy \\ &= \frac{\partial}{\partial \theta_r} \int_{\mathcal{Y}^n} f(y; \theta) dy = \frac{\partial}{\partial \theta_r}(1) = 0. \end{aligned}$$

We leave property 2 as an exercise.

2.5.2 Fisher information

The quantity

$$i(\theta) := \mathrm{Var}_\theta\{U(\theta; Y)\}$$

is known as the *Fisher information.* It can be thought of as a measure of how hard it is to estimate θ when it is the true parameter value. A related quantity is the *observed information matrix*, $j(\theta)$ defined by

$$j(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta^T}\ell(\theta; Y).$$

Note that $i(\theta) = \mathbb{E}_\theta(j(\theta))$.

Example. Consider our friend the normal linear model: $Y = X\beta + \varepsilon$, $\varepsilon \sim N_n(0, \sigma^2)$. Then

$$i(\beta, \sigma^2) = \begin{pmatrix} \sigma^{-2}X^TX & 0 \\ 0 & n\sigma^{-4}/2 \end{pmatrix}.$$

Note that writing $i^{-1}(\beta)$ for the top left $p \times p$ sub-matrix of $i^{-1}(\beta, \sigma^2)$ (the matrix inverse of $i(\beta, \sigma^2)$), we have that $\mathrm{Var}(\hat{\beta}) = i^{-1}(\beta)$.

In fact we have the following result.

Theorem 5 (Cramér–Rao lower bound). *Let $\tilde{\theta}$ be an unbiased estimator of θ. Then under regularity conditions,*

$$\mathrm{Var}_\theta(\tilde{\theta}) - i^{-1}(\theta)$$

is positive semi-definite.

Proof. We only sketch the proof when $d = 1$. By the Cauchy–Schwarz inequality,

$$i(\theta)\mathrm{Var}(\tilde{\theta}) = \mathrm{Var}(U(\theta))\mathrm{Var}(\tilde{\theta}) \geq \{\mathrm{Cov}(\tilde{\theta}, U(\theta))\}^2.$$

As $\mathbb{E}\{U(\theta)\} = 0$,

$$\begin{aligned}
\mathrm{Cov}(\tilde{\theta}, U(\theta)) &= \mathbb{E}(\tilde{\theta}U(\theta)) \\
&= \int_{\mathcal{Y}^n} \tilde{\theta}(y)\left(\frac{\partial}{\partial\theta}\log f(y;\theta)\right) f(y;\theta)dy \\
&= \int_{\mathcal{Y}^n} \frac{\partial}{\partial\theta} f(y;\theta)\tilde{\theta}(y)dy \\
&= \frac{\partial}{\partial\theta}\int_{\mathcal{Y}^n} \tilde{\theta}(y)dy = \frac{\partial}{\partial\theta}\mathbb{E}_\theta\tilde{\theta}.
\end{aligned}$$

But as $\tilde{\theta}$ is unbiased we finally get

$$\mathrm{Cov}(\tilde{\theta}, U(\theta)) = \frac{\partial}{\partial\theta}\theta = 1. \qquad \square$$

Since the m.l.e. of β in the normal linear model, $\hat{\beta} := (X^TX)^{-1}X^TY$ is unbiased, we conclude that $\hat{\beta}$ has the minimum variance among all unbiased estimators of β (not just the *linear* unbiased estimators as the Gauss–Markov theorem yields).

It turns out that this is, to a certain extent, a general feature of maximum likelihood estimators (in finite dimensional models), as we now discuss.

2.5.3 Two key asymptotic results

A feature of maximum likelihood estimators that *asymptotically* they are normally distributed with mean the true parameter value θ and variance the inverse of the Fisher information matrix evaluated at θ. Thus asymptotically they achieve the Cramér–Rao lower bound. To make this a little more precise, let us recall some definitions to do with convergence of random variables.

***Convergence of random variables*.** We say a sequence of random variables $Z_1, Z_2, \dots$ with corresponding distribution functions $F_1, F_2, \dots$ converges in distribution to a random variable Z with distribution function F, and write $Z_n \xrightarrow{d} Z$ if $F_n(x) \to F(x)$ at all x where F is continuous.

A sequence of random vectors $Z_n \in \mathbb{R}^k$ converges in distribution to a continuous random vector $Z \in \mathbb{R}^k$ when

$$\mathbb{P}(Z_n \in B) \to \mathbb{P}(Z \in B)$$

for all (Borel) sets B for which $\delta B := \mathrm{cl}(B) \setminus \mathrm{int}(B)$ has $\mathbb{P}(Z \in \delta B) = 0$.

For example, the multidimensional central limit theorem (CLT) states that if $Z_1, Z_2, \dots$ are i.i.d. random vectors in $\mathbb{R}^k$ with positive definite variance Σ and mean $\mu \in \mathbb{R}^k$, then writing $\bar{Z^{(n)}}$ for $\frac{1}{n}\sum_{i=1}^n Z_i$, we have

$$\sqrt{n}(\bar{Z^{(n)}} - \mu) \xrightarrow{d} N_k(0, \Sigma).$$

A stronger mode of convergence is convergence in probability. We say $Z_n \in \mathbb{R}^k$ tends to $Z \in \mathbb{R}^k$ in probability if for every $\epsilon > 0$

$$\mathbb{P}(\|Z_n - Z\| > \epsilon) \to 0,$$

as $n \to \infty$. If $Z_n \xrightarrow{p} Z$ then $Z_n \xrightarrow{d} Z$.

For example, the weak law of large numbers (WLLN) states that if $Z_1, Z_2, \dots$ are i.i.d. with mean $\mu \in \mathbb{R}^k$ then $\bar{Z^{(n)}} \xrightarrow{p} \mu$. Some useful results concerning convergence of random variables are:

Proposition 6 (Continuous mapping theorem). *If $Z_n \xrightarrow{d} Z$ and $h : \mathbb{R} \to \mathbb{R}$ is continuous, then $h(Z_n) \xrightarrow{d} h(Z)$.*

Proposition 7 (Slutsky's lemma). *If $Y_n \xrightarrow{d} Y$ and $Z_n \xrightarrow{p} c$, where c is a constant, then*

1. $Y_n + Z_n \xrightarrow{d} Y + c$,
2. $Y_n Z_n \xrightarrow{d} cY$,
3. $\dfrac{Y_n}{Z_n} \xrightarrow{d} \dfrac{Y}{c}$.

Asymptotic distribution of maximum likelihood estimators.

Theorem 8. *Assume that the Fisher information matrix when there are n observations, $i^{(n)}(\theta)$ (where we have made the dependence on n explicit) satisfies $i^{(n)}(\theta)/n \to I(\theta)$ for some positive definite matrix I. Then denoting the maximum likelihood estimator of θ when there are n observations by $\hat{\theta}^{(n)}$, under regularity conditions we have*

$$\sqrt{n}(\hat{\theta}^{(n)} - \theta) \xrightarrow{d} N_d(0, I^{-1}(\theta)).$$

A short-hand and informal version of writing this (which is fine for this course) is that

$$\hat{\theta} \sim AN_d(\theta, i^{-1}(\theta)),$$

to be read "$\hat{\theta}$ is asymptotically normal with mean θ and variance $i^{-1}(\theta)$".

***Sketch of proof*.** Here is a sketch of the proof when $d = 1$ and our data are i.i.d. rather than simply independent.

A Taylor expansion of the score function about the true parameter value θ gives

$$0 = U(\hat{\theta}) = U(\theta) - (\hat{\theta} - \theta)j(\theta) + \mathrm{Rem}_n(\theta).$$

Since $\mathbb{E}(U(\theta)) = 0$ and $\mathrm{Var}(U(\theta)) = i(\theta) = ni_1(\theta)$, where $i_1(\theta)$ is the Fisher information of the first observation, by the CLT we have

$$\frac{U(\theta)}{\sqrt{n}} \xrightarrow{d} N(0, i_1(\theta)).$$

By Slutsky's lemma (no. 1 additive), provided that

$$\frac{\mathrm{Rem}_n(\theta)}{\sqrt{n}} \xrightarrow{p} 0,$$

we have

$$\begin{aligned}\sqrt{n}(\hat{\theta} - \theta)\frac{j(\theta)}{n} &= \frac{U(\theta)}{\sqrt{n}} + \frac{\mathrm{Rem}_n(\theta)}{\sqrt{n}} \\ &\xrightarrow{d} N(0, i_1(\theta)).\end{aligned}$$

But by WLLN, $j(\theta)/n \xrightarrow{p} i_1(\theta)$ as $n \to \infty$, so by Slutsky's lemma (no. 3),

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, i_1^{-1}(\theta)).$$

Relevance of the result. How are we to use this result? The first issue is that as the true parameter θ is unknown, so is $i^{-1}(\theta)$. However, provided that $i^{-1}(\theta)$ is a continuous function of θ, we may estimate this well with $i^{-1}(\hat{\theta})$, and we can show that, for example

$$\frac{\hat{\theta}_j - \theta_j}{\sqrt{(i^{-1}(\hat{\theta}))_{jj}}} \xrightarrow{d} N(0, 1)$$

Thus we can create an approximate $1 - \alpha$ level confidence interval for θ_j with

$$\left[\hat{\theta}_j - z_{\alpha/2}\sqrt{(i^{-1}(\hat{\theta}))_{jj}},\ \hat{\theta}_j + z_{\alpha/2}\sqrt{(i^{-1}(\hat{\theta}))_{jj}}\right],$$

where z_α is the upper α-point of $N(0, 1)$. The coverage of this confidence interval tends to $1 - \alpha$ as $n \to \infty$. Similarly, an asymptotic $1 - \alpha$ level confidence set for θ is given by

$$\{\theta' : (\hat{\theta} - \theta')^T i(\hat{\theta})(\hat{\theta} - \theta') \leq \chi^2_d(\alpha)\}.$$

Another issue is that we never have an infinite amount of data. What does the asymptotic result have to say when we have maybe 100 observations? From a purely logical point of view, it says absolutely nothing. You will have had it drilled into you long ago in Analysis I that even the first trillion terms of a sequence have nothing to do with its limiting behaviour. On the other hand, we can be more optimistic and hope that $n = 100$ is large enough for the finite sample distribution of $\hat{\theta}$ to be close to the limiting distribution. Performing simulations can help justify this optimism and give us values of n for which we can expect the limiting arguments to apply.

Wilks' theorem. The result on asymptotic normality of maximum likelihood estimators allows us to construct confidence intervals for individual components of θ and hence perform hypothesis tests of the form $H_0 : \theta_j = 0$, $H_1 : \theta_j \neq 0$. Now suppose we wish to test

$$\begin{aligned} &H_0 : \theta \in \Theta_0 \qquad \text{against} \\ &H_1 : \theta \notin \Theta_0 \end{aligned}$$

where $\Theta_0 \subset \Theta$, the full parameter space, and Θ_0 is of lower dimension than Θ. The precise meaning of dimension when Θ_0 and Θ are not affine spaces (i.e. a translation of a subspace) but rather general manifolds would require us to go into the realm of differential geometry, which we won't do here. Perhaps the most important case of interest is when $\theta = (\theta_0^T, \theta_1^T)^T$ and $\theta_0 \in \mathbb{R}^{d_0}$ with $\Theta = \mathbb{R}^d$, and we are testing

$$\begin{aligned} &H_0 : \theta_0 = 0 \qquad \text{against} \\ &H_1 : \theta_0 \neq 0. \end{aligned}$$

Wilks' theorem gives the asymptotic distribution of of the likelihood ratio statistic

$$w_{\text{LR}}(H_0) = 2 \log \left\{ \frac{\sup_{\theta' \in \Theta} L(\theta')}{\sup_{\theta' \in \Theta_0} L(\theta')} \right\} = 2\{\sup_{\theta' \in \Theta} \ell(\theta') - \sup_{\theta' \in \Theta_0} \ell(\theta')\}.$$

Theorem 9 (Wilks' theorem)**.** *Suppose that H_0 is true. Then, under regularity conditions*

$$w_{\text{LR}}(H_0) \xrightarrow{d} \chi_k^2$$

where $k = \dim(\Theta) - \dim(\Theta_0)$.

Sketch of proof We only sketch the proof when the null hypothesis is simple so $\Theta_0 = \{\theta_0\}$, and when the data $Y_1, Y_2, \ldots$ are i.i.d. rather than just independent. A Taylor expansion of $\ell(\theta_0)$ centred at the (unrestricted) maximum likelihood estimate $\hat{\theta}$ gives

$$\ell(\theta_0) = \ell(\hat{\theta}) + (\hat{\theta} - \theta_0)^T U(\hat{\theta}) - \frac{1}{2}(\hat{\theta} - \theta_0)^T j(\hat{\theta})(\hat{\theta} - \theta_0) + \text{Rem}_n(\hat{\theta}).$$

Using $U(\hat{\theta}) = 0$ and provided that $\text{Rem}_n(\hat{\theta}) \xrightarrow{p} 0$,

$$2\{\ell(\hat{\theta}) - \ell(\theta_0)\} = (\hat{\theta} - \theta_0)^T j(\hat{\theta})(\hat{\theta} - \theta_0) - 2\text{Rem}_n(\hat{\theta}) \xrightarrow{d} \chi_k^2$$

under H_0 by Slutsky's theorem, provided $(\hat{\theta} - \theta_0)^T j(\hat{\theta})(\hat{\theta} - \theta_0) \xrightarrow{d} \chi_k^2$.

Note that the likelihood ratio test in conjunction with Wilks' theorem can also be used to test whether individual components of θ are 0. Unlike the analogous situation in the normal linear model where the F-test for an individual variable is equivalent to the t-test, here tests based on asymptotic normality of $\hat{\theta}$ and the likelihood ratio test will in general be different—usually the likelihood ratio test is to be preferred, though it may be require more computation to calculate the test statistic.

2.5.4 Inference in generalised linear models

Let $i(\beta, \sigma^2)$ be the Fisher information in a generalised linear model. It can be shown that this matrix is block diagonal, so writing $i(\beta)$ for the $p \times p$ top left submatrix of $i(\beta, \sigma^2)$ and $i(\sigma^2)$ for the bottom right entry, we have

$$i(\beta, \sigma^2) = \begin{pmatrix} i(\beta) & 0 \\ 0 & i(\sigma^2) \end{pmatrix} \qquad \text{and} \qquad i^{-1}(\beta, \sigma^2) = \begin{pmatrix} i^{-1}(\beta) & 0 \\ 0 & i^{-1}(\sigma^2) \end{pmatrix}.$$

The asymptotic results we have studied then show that

$$\hat{\beta} \sim AN_p(\beta, i^{-1}(\beta)).$$

This (along with continuity of $i^{-1}(\beta)$) justifies the following asymptotic $(1-\alpha)$-level confidence set for β_j

$$\left[\hat{\beta}_j - z_{\alpha/2}\sqrt{\{i^{-1}(\hat{\beta})\}_{jj}}, \hat{\beta}_j + z_{\alpha/2}\sqrt{\{i^{-1}(\hat{\beta})\}_{jj}}\right],$$

where z_α is the upper α-point of $N(0,1)$. To test $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$, we can reject H_0 if the confidence interval above excludes 0 i.e. if

$$\frac{|\hat{\beta}_j|}{\sqrt{\{i^{-1}(\hat{\beta})\}_{jj}}} > z_{\alpha/2}.$$

Now suppose β is partitioned as $\beta = (\beta_0^T, \beta_1^T)^T$ where $\beta_0 \in \mathbb{R}^{p_0}$ and we wish to test $H_0 : \beta_1 = 0$ against $\beta_1 \neq 0$. Write $\check{\beta}_0$ for the m.l.e. of β_0 under the null model, and assume for the moment that the dispersion parameter σ^2 is known (as is the case for the Poisson and binomial models—the two most important generalised linear models). Write $\tilde{\ell}(\mu, \sigma^2)$ for $\ell(\beta, \sigma^2)$ so that

$$\tilde{\ell}(\mu, \sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^n \frac{1}{a_i}[y_i\theta(\mu_i) - K\{\theta(\mu_i)\}] + \sum_{i=1}^n \log\{a(\sigma^2, y_i)\}$$

with $\mu_i = g^{-1}(x_i^T\beta)$. Further write

- $\tilde{\ell}(y, \sigma^2)$ for $\ell(\mu, \sigma^2)$ with each μ_i replaced by its m.l.e. under the model where $\mu_1, \ldots, \mu_n$ are unrestricted (i.e. μ_i is replaced by y_i),
- $\tilde{\ell}(\hat{\mu}, \sigma^2)$ for $\ell(\hat{\beta}, \sigma^2)$, and
- $\tilde{\ell}(\check{\mu}, \sigma^2)$ for $\ell(\check{\beta}_0, \sigma^2)$ (so $\check{\mu}_i = g^{-1}(x_i^T\check{\beta}_0)$).

The *deviance* of the model in H_1 is

$$D(y; \hat{\mu}) := 2\sigma^2\{\tilde{\ell}(y, \sigma^2) - \tilde{\ell}(\hat{\mu}, \sigma^2)\};$$

the deviance of the null model is $D(y; \check{\mu}) := 2\sigma^2\{\tilde{\ell}(y, \sigma^2) - \tilde{\ell}(\check{\mu}, \sigma^2)\}$. The deviance may be thought of as the appropriate generalisation to GLMs of the residual sum of squares from the linear model. Note that the deviance in reduced in the larger model.

Notice that

$$w_{\mathrm{LR}}(H_0) = \frac{D(y; \check{\mu}) - D(y; \hat{\mu})}{\sigma^2},$$

so by Wilks' theorem we may test if $\beta_1 = 0$ at level α by rejecting the null hypothesis when the value of this test statistic is larger than $\chi^2_{p-p_0}(\alpha)$.

If σ^2 is unknown it must be replaced with an estimate such as

$$\tilde{\sigma}^2 := \frac{1}{n-p}\sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)^2}{a_i V(\hat{\mu}_i)}.$$

Provided that $\tilde{\sigma}^2 \xrightarrow{p} \sigma^2$ Slutsky's theorem ensures that the asymptotic distribution of $w_{\mathrm{LR}}(H_0)$ remains unchanged, though it is often better to use $F_{p-p_0,n-p}(\alpha)$ as the critical value for the test statistic

$$\frac{\frac{1}{p-p_0}\{D(y; \check{\mu}) - D(y; \hat{\mu})\}}{\tilde{\sigma}^2}$$

in this case.

2.6 Computation

We have seen how despite the maximum likelihood estimator $\hat{\beta}$ of β in a generalised linear model not having an explicit form (except in special cases such as the normal linear model), we can show that asymptotically the m.l.e. has rather attractive properties and we can still perform inference that is asymptotically valid. How are we to compute $\hat{\beta}$ when all we know about it is the fact that it satisfies

$$0 = \left.\frac{\partial \ell(\beta, \sigma^2)}{\partial \beta}\right|_{\beta=\hat{\beta}} =: U(\hat{\beta})? \tag{2.6.1}$$

Here, with a slight abuse of notation, we have written $U(\beta)$ for the first p components of $U(\beta, \sigma^2)$; similarly let us write $j(\beta)$ and $i(\beta)$ for the top left $p \times p$ submatrix of $j(\beta, \sigma^2)$ and $i(\beta, \sigma^2)$ respectively.

If U were linear in β, we should be able to solve the system of linear equations in (2.6.1) to find $\hat{\beta}$. Though in general U won't be a linear function, given that it is differentiable (recall that the link function g is required to be twice differentiable), an application of Taylor's theorem shows that it is at least locally linear, so

$$U(\beta) \approx U(\beta_0) - j(\beta_0)(\beta - \beta_0)$$

for β close to β_0. If we managed to find a β_0 close to $\hat{\beta}$, the fact that $U(\hat{\beta}) = 0$ suggests approximating $\hat{\beta}$ by the solution of

$$U(\beta_0) - j(\beta_0)(\beta - \beta_0) = 0$$

in β, i.e.

$$\beta_0 + j^{-1}(\beta_0)U(\beta_0),$$

where we have assumed that $j(\beta_0)$ is invertible. This motivates the following iterative algorithm (the Newton–Raphson algorithm): starting with an initial guess at $\hat{\beta}$, $\hat{\beta}_0$, at the m^{th} iteration we update

$$\hat{\beta}_m = \hat{\beta}_{m-1} + j^{-1}(\hat{\beta}_{m-1})U(\hat{\beta}_{m-1}). \tag{2.6.2}$$

A potential issue with this algorithm is that $j(\hat{\beta}_{m-1})$ may be singular or close to singular and thus make the algorithm unstable. The method of *Fisher scoring* replaces $j(\hat{\beta}_{m-1})$ with $i(\hat{\beta}_{m-1})$ which is always positive definite (subject to regularity conditions) and generally better behaved.

Fisher scoring may not necessarily converge to $\hat{\beta}$ but almost always does. We terminate the algorithm when successive iterations produce negligible difference.

Let us examine this procedure in more detail. It can be shown (see example sheet) that the score function and Fisher information matrix have entries

$$\begin{aligned} U_j(\beta) &= \sum_{i=1}^n \frac{(y_i - \mu_i)X_{ij}}{\sigma_i^2 V(\mu_i) g'(\mu_i)} \qquad j = 1, \ldots, p, \\ i_{jk}(\beta) &= \sum_{i=1}^n \frac{X_{ij}X_{ik}}{\sigma_i^2 V(\mu_i)\{g'(\mu_i)\}^2} \qquad k = 1, \ldots, p. \end{aligned}$$

Choosing the canonical link $g(\mu) = \theta(\mu)$ simplifies $U_j(\beta)$ and $i_{jk}(\beta)$ since $g'(\mu) = 1/V(\mu)$. Let $W(\mu)$ be the $n \times n$ diagonal matrix with i^{th} diagonal entry

$$W_{ii}(\mu) := \frac{1}{a_i V(\mu_i)\{g'(\mu_i)\}^2}.$$

Further let $\tilde{Y}(\mu) \in \mathbb{R}^n$ be the vector with i^{th} component

$$\tilde{Y}_i(\mu) = g'(\mu_i)(y_i - \mu_i).$$

Then we may write

$$\begin{aligned} U(\beta) &= \sigma^{-2} X^T W \tilde{Y} \\ i(\beta) &= \sigma^{-2} X^T W X. \end{aligned}$$

Let us set

$$\begin{aligned} W_m &:= W(\hat{\mu}_m) \\ \tilde{Y}_m &:= \tilde{Y}(\hat{\mu}_m). \end{aligned}$$

(Note here the subscript m is not indexing different components of a single vector $\tilde{Y}$ but different vectors $\tilde{Y}_m$. Then we see that

$$\hat{\beta}_m = \hat{\beta}_{m-1} + (X^T W_{m-1} X)^{-1} X^T W_{m-1} \tilde{Y}_{m-1}.$$

If we define the *adjusted dependent variable* Z_m by

$$Z_m := \tilde{Y}_m + \hat{\eta}_m,$$

where $\eta_m = X\hat{\beta}_m$, then

$$\hat{\beta}_m = (X^T W_{m-1} X)^{-1} X^T W_{m-1} Z_{m-1} = \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \left\{ \sum_{i=1}^n W_{m-1,ii} (Z_{m-1,i} - x_i^T b)^2 \right\}.$$

See example sheet 1 for the final equality. Thus the sequence of approximations to $\hat{\beta}$ are given by *iterative weighted least squares* (IWLS) of the adjusted dependent variable $Z_{m-1,i}$ on X with weights given by the diagonal entries of W_{m-1}.

With this formulation, we can start with an initial guess of $\hat{\mu}$ rather than one of $\hat{\beta}$. An obvious choice for this initial guess $\hat{\mu}_0$ is the response y, although a small adjustment such as $\hat{\mu}_{0,i} = \max\{y_i, \epsilon\}$ for $\epsilon > 0$ may be necessary if $g(\mu) = \log(\mu)$ for example, to avoid problems when $y_i = 0$.

Chapter 3

Specific regression problems

3.1 Binomial regression

Suppose we have data $(y_1, x_1^T), \dots, (y_n, x_n^T) \in \mathbb{R} \times \mathbb{R}^p$ where it seems reasonable to assume the y_i are realisations of random variables Y_i that are independent for $i = 1, \dots, n$ and

$$Y_i \sim \frac{1}{n_i}\text{Bin}(n_i, \mu_i), \qquad \mu_i \in (0,1)$$

with the n_i known positive integers. An example of such data could be the proportion Y_i of n_i organisms to have been killed by concentrations of various drugs / temperature level etc. collected together in a vector x_i. Often the $n_i = 1$ so $Y_i \in \{0,1\}$—we could have 1 representing spam and 0 representing ham for example. If we assume that $\mu_i = \mathbb{E}(Y_i)$ is related to the covariates x_i through $g(\mu_i) = x_i^T\beta$ for some link function g and unknown vector of coefficients $\beta \in \mathbb{R}^p$, then this model falls within the framework of the generalised linear model. Indeed,

$$\begin{aligned} f(y_i;\mu_i) &= \binom{n_i}{n_i y_i}\mu_i^{n_i y_i}(1-\mu_i)^{n_i - n_i y_i} \\ &= \underbrace{\binom{n_i}{n_i y_i}}_{a(a_i,y_i)} \exp\left[\frac{1}{n_i^{-1}}\left\{y_i \underbrace{\log\left(\frac{\mu_i}{1-\mu_i}\right)}_{\theta_i=\theta(\mu_i)} + \underbrace{\log(1-\mu_i)}_{-K(\theta_i)}\right\}\right]. \end{aligned}$$

We can take the dispersion parameter as 1 and let $a_i = n_i^{-1}$.

Once we have chosen a link function, we can obtain the m.l.e. of β using the IWLS algorithm and then perform hypothesis tests or construct confidence intervals that are asymptotically valid using the general theory of maximum likelihood estimators.

3.1.1 Link functions

In order to avoid having to place restrictions on the values β can take, we can choose a link function g such that the image $g((0,1)) = g(\mathcal{M}) = \mathbb{R}$. Three commonly used link functions are given below in increasing order of their popularity (coincidentally this is also the order in which they were introduced). Their graphs are plotted in Figure 3.1.

1. $g(\mu) = \log(-\log(1-\mu))$ gives the *complementary log–log* link.
2. $g(\mu) = \Phi^{-1}(\mu)$ where Φ is the c.d.f. of the standard normal distribution (so Φ^{-1} is the quantile function of the standard normal) gives the *probit* link.

3. $g(\mu) = \log\left(\dfrac{\mu}{1-\mu}\right)$ is the logit link. This is the canonical link function for the GLM.

The probit link gives an interesting latent variable interpretation of the model. Consider the case where $n_i = 1$. Imagine that there exists a $Y^* \in \mathbb{R}^n$ such that

$$Y^* = X\beta^* + \varepsilon$$

where $\varepsilon \sim N_n(0, \sigma^2 I)$. Suppose we do not observe Y^* but instead, only see $Y \in \{0,1\}^n$ with i^{th} component given by

$$Y_i = \mathbb{1}_{\{Y_i^* > 0\}}.$$

Then we see that

$$\begin{aligned}\mathbb{P}(Y_i = 1) = \mathbb{P}(Y_i^* > 0) &= \mathbb{P}(x_i^T\beta^* > -\varepsilon_i) \\ &= \mathbb{P}(x_i^T\beta^*/\sigma > Z_i) \qquad \text{where } Z_i \sim N(0,1) \\ &= \Phi(x_i^T\beta) \qquad \text{where } \beta := \beta^*/\sigma.\end{aligned}$$

The models generated by the other two link functions also have latent variable interpretations.

Of the three link functions, by far the most popular is the logit link. This is partly because it is the canonical link, and so simplifies some calculations, but perhaps more importantly, the coefficents from a model with logit link (a *logistic regression* model) are easy to interpret. The value e^{β_j} gives the multiplicative change in the odds $\mu_i/(1-\mu_i)$ for a unit increase in the value of the j^{th} variable, keeping the values of all other variables fixed. To see this note that

$$\frac{\mu_i}{1-\mu_i} = \exp\left(\sum_{j=1}^p X_{ij}\beta_j\right) = \prod_{j=1}^p (e^{\beta_j})^{X_{ij}}.$$

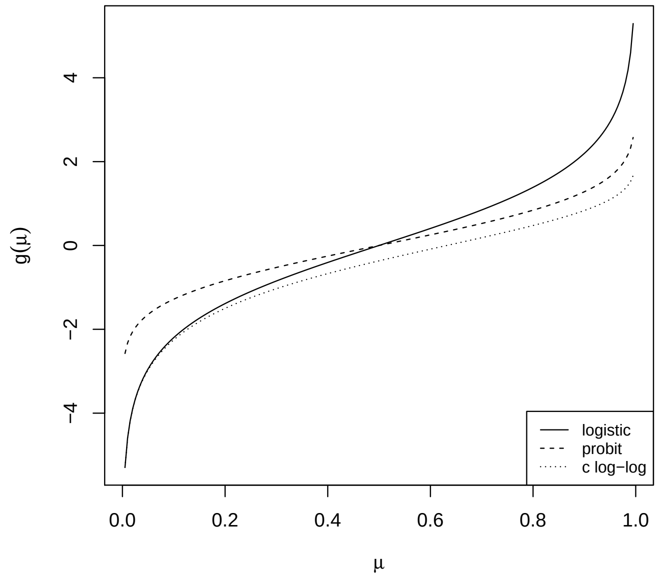

Figure 3.1: The graphs of three commonly used link functions for binomial regression.

3.1.2 *A classification view of logistic regression*

In the case where $n_i = 1$ for all i, logistic regression can be thought of as a classification procedure. The response value of each observation is then either 0 or 1, and so divides the observations into two classes. Having fit a logistic regression to some data which we shall call the *training data*, we can then predict responses (class labels) for new data for which we only have the covariate values. We can do this by applying the function $\hat{C}_\tau$ below to each new observation:

$$\hat{C}_\tau(x) := \mathbb{1}_{\{\hat{\pi}(x) \geq \tau\}},$$

where

$$\hat{\pi}(x) := \frac{\exp(x^T\hat{\beta})}{1 + \exp(x^T\hat{\beta})}$$

and $\hat{\beta}$ is the m.l.e. of β based on the training data. The value τ is a threshold and should be set according to how bad predicting a class label of 1 when it is in fact 0 is, compared to predicting a class label of 0 when it is in fact 1.

If in addition to our training data, we have another set of labelled data, we can plot the proportion of class 1 observations correctly classified against the proportion of class 0 observations incorrectly classified using $\hat{C}_\tau$, for different values of τ. This set of data is known as a *test set.* As τ varies between 0 and 1, the points plotted trace out what is known as a *Receiver Operating Characteristic* (ROC) curve. This gives a visual representation of how good a classifier our model is, and can serve as a way of comparing different classifiers. A classifier with ROC curve always above that of another classifier is certainly to be preferred. However, when ROC curves of classifiers cross, no classifier uniformly dominates the other. In these cases, a common measure of performance is the area under the ROC curve (AUC). If in a particular application, there is a certain probability of incorrectly classifying a class 0 observation that can be tolerated (say 5%), and the chance of incorrectly classifying a class 1 observation is to be minimised subject to this error tolerance, then ROC curves should be compared at the relevant point.

Of course all these comparisons are contingent on the particular test set used. Given a collection of data, it is advisable to randomly split it into training and test sets several times and average the ROC curves produced by each of the splits. Suppose the training sets are all of size n_{tr}, say. The average ROC curve is then a measure of the average performance of the classification procedure when it is fed n_{tr} observations where, thinking of the covariates now as (realisations of) random variables, this average is over the joint distribution of response and covariates.

3.1.3 *Logistic regression and linear discriminant analysis*

Additional support for the logistic link function can be gained by considering a classification problem where we have i.i.d. data $(Y_1, X_1), \ldots, (Y_n, X_n) \in \{0, 1\} \times \mathbb{R}^p$. Let $(Y, X) = (Y_1, X_1)$ (note X is not the design matrix here, it is a p-dimensional random vector). Assume that

$$X|Y = 0 \sim N_p(\mu_0, \Sigma), \qquad X|Y = 1 \sim N_p(\mu_1, \Sigma). \tag{3.1.1}$$

Suppose further that $\mathbb{P}(Y_i = 1) = \pi_1 = 1 - \pi_0$. Now

$$\begin{aligned}\log\left\{\frac{\mathbb{P}(Y = 1|X)}{\mathbb{P}(Y = 0|X)}\right\} &= \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}(\mu_1 + \mu_0)^T\Sigma^{-1}(\mu_1 - \mu_0) + X^T\Sigma^{-1}(\mu_1 - \mu_0) && (3.1.2)\\ &= \alpha + X^T\beta, && (3.1.3)\end{aligned}$$

with

$$\alpha := \log\left(\frac{\pi_1}{\pi_0}\right) - \frac{1}{2}(\mu_1 + \mu_0)^T\Sigma^{-1}(\mu_1 - \mu_0)$$
$$\beta := \Sigma^{-1}(\mu_1 - \mu_0).$$

Thus the log odds of the posterior class probabilities is precisely of the form needed for the logistic regression model to be correct.

Typically if it is known that the data generating process is (3.1.1), then a classifier is formed by replacing the population parameters π_1, μ_0, μ_1 and Σ in (3.1.2) with estimates, and then classifying to the class with the largest posterior probability. This gives Fisher's *linear discriminant analysis* (LDA), which you will have already met if you took Principles of Statistics.

The logistic regression model is more general in that it makes fewer assumptions. It does not specify the distribution of the covariates and instead treats them as fixed (i.e. it conditions on them). When the mixture of Gaussians model in (3.1.1) is correct, one can expect LDA to perform better. However, when (3.1.1) is not satisfied, logistic regression may be preferred.

3.1.4 Model checking

We have not discussed model checking for GLMs but it proceeds in much the same way as for the normal linear model, and residuals are the chief means for assessing the validity of model assumptions. With GLMs there are several different types of residuals one can consider. One form of residual builds on the analogy that the deviance is like the residual sum of squares from the normal linear model. The *deviance residuals* in a GLM are defined as

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

where $D(y_i; \hat{\mu}_i)$ is the i^{th} summand in the definition of $D(y; \hat{\mu})$, so

$$D(y_i; \hat{\mu}_i) = \frac{2}{a_i}[y_i\{\theta(y_i) - \theta(\hat{\mu}_i)\} - \{K(\theta(y_i)) - K(\theta(\hat{\mu}_i))\}].$$

In binomial regression (and also Poisson regression), one can sometimes test a particular model against a *saturated* model:

$$\begin{aligned} H_0 :&\mu_i = g^{-1}(x_i^T\beta) \;\; i = 1, \dots, n \qquad \text{against} \\ H_1 :&\mu_1, \dots, \mu_n \quad \text{unrestricted.} \end{aligned}$$

In this case,

$$w_{\text{LR}}(H_0) = \frac{D(y; \hat{\mu}) - D(y; y)}{\sigma^2} = \frac{D(y; \hat{\mu})}{\sigma^2},$$

but standard asmpytotic theory no longer ensures that this converges in distribution to a χ^2_{n-p}. Nevertheless, other asymptotic arguments can sometimes be used to justify referring the likelihood ratio statistic to χ^2_{n-p}, for instance when

- $Y_i \sim \frac{1}{n_i}\text{Bin}(n_i, \mu_i)$, with n_i large, and
- $Y_i \sim \text{Pois}(\mu_i)$ with μ_i large.

This is because in these cases, the individual Y_i get close to normally distributed random variables. Such asymptotics are known as *small dispersion asymptotics.*

3.2 Poisson regression

We have seen how binomial regression can be appropriate when the responses are proportions (including the important case when the proportions are in $\{0, 1\}$ i.e. the classification scenario). Now we consider count data e.g. the number of texts you receive each day, or the number of terrorists attacks that occur in a country each week. Another example where count data arises is the following: imagine conducting an (online) survey where perhaps you ask people to enter their college and their voting intentions. The survey may be live for a fixed amount of time and then you can collect to together the data into a 2-way *contingency table*:

College	Labour	Conservative	Liberal Democrats	Other
Trinity				
⋮				

When the responses are counts, it may be sensible to model them as realisations of Poisson random variables. A word of caution though. A Poisson regression model entails a particular relationship between the mean and variance of the responses: if $Y_i \sim \text{Pois}(\mu_i)$, then $\text{Var}(Y_i) = \mu_i$. In many situations we may find this assumption is violated. Nevertheless, the Poisson regression model can often be a reasonable approximation.

If the probability of occurence of an event in a given time interval is proportional to the length of that time interval and independent of the occurence of other events, then the number of events in any specified time interval will be Poisson distributed. Wikipedia lists a number of situations where Poisson data arise naturally:

- Telephone calls arriving in a system,
- Photons arriving at a telescope
- The number of mutations on a strand of DNA per unit length
- Cars arriving at a traffic light
- ...

The Poisson regression model assumes that our data $(Y_1, x_1), \ldots, (Y_n, x_n) \in \{0, 1, \ldots\} \times \mathbb{R}^p$ have $Y_1, \ldots, Y_n$ independent with $Y_i \sim \text{Pois}(\mu_i)$, $\mu_i > 0$. An example sheet question asks you to verify that the $\{\text{Pois}(\mu) : \mu \in (0, \infty)\}$ is an exponential dispersion family with dispersion parameter $\sigma^2 = 1$. In line with the GLM framework, we assume the μ_i are related to the covariates through $g(\mu_i) = x_i^T \beta$ for a link function g.

By far the most commonly used link function is the log link—this also happens to be the canonical link. In fact the Poisson regression model is often called the *log-linear model.* We only consider the log link here. Two reasons for the popularity of the log link are:

- $\{\log(\mu) : \mu \in (0, \infty)\} = \mathbb{R}$. The parameter space for β is then simply $\mathbb{R}^p$ and no restrictions are needed.
- Interpretability: if
$$\mu_i = \exp\left(\sum_{j=1}^{p} X_{ij}\beta_j\right) = \prod_{j=1}^{p} (e^{\beta_j})^{X_{ij}},$$
then we see that e^{β_j} is the multiplicative change in the expected value of the response for a unit increase in the j^{th} variable.

In the next practical class we'll look at data from the English Premier League and attempt to model the home and away scores Y_{ij}^{h} and Y_{ij}^{a} when team i is home to team j as independent Poisson random variables with respective means

$$\mu_{ij}^{\mathrm{h}} = \exp(\Delta + \alpha_i - \beta_j), \qquad \mu_{ij}^{\mathrm{a}} = \exp(\alpha_j - \beta_i).$$

Here Δ represents the home advantage (we expect it to be greater than 0) and α_i and β_i the offensive and defensive strengths of team i.

3.2.1 Likelihood equations

With $\log(\mu_i) = x_i^T\beta$, we have

$$\begin{aligned} L(\beta) &= \prod_{i=1}^n e^{-\mu_i(\beta)} \frac{\mu_i(\beta)^{y_i}}{y_i!} \\ &\propto \prod_{i=1}^n \exp(-e^{x_i^T\beta}) \exp(y_i x_i^T\beta), \end{aligned}$$

so

$$\ell(\beta) = -\sum_{i=1}^n \exp(x_i^T\beta) + \sum_{i=1}^n y_i x_i^T\beta.$$

Let us consider the case where we have an intercept term. We can either say that the first column of the design matrix X is a column of 1's, or we can include it explicitly in the model. In the latter case we take

$$\log(\mu_i) = \alpha + x_i^T\beta,$$

so the log-likelihood is

$$\ell(\alpha, \beta) = -\sum_{i=1}^n \exp(\alpha + x_i^T\beta) + \sum_{i=1}^n y_i(\alpha + x_i^T\beta).$$

Now differentiating w.r.t. α, we have

$$0 = \frac{\partial \ell(\hat{\alpha}, \hat{\beta})}{\partial \alpha} = \sum_{i=1}^n \{y_i - \exp(\hat{\alpha} + x_i^T\hat{\beta})\}.$$

Thus writing $\hat{\mu}_i := \exp(\alpha + x_i^T\beta)$, we have

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

The deviance and Pearson's χ^2-statistic

The fact above simplifies the deviance in a Poisson GLM. We have

$$\tilde{\ell}(\mu, \sigma^2) = -\sum_{i=1}^n \mu_i + \sum_{i=1}^n y_i \log(\mu_i),$$

so

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

when an intercept term is included.

Write $y_i = \hat{\mu}_i + \delta_i$, so we have that $\sum \delta_i = 0$. Then, by a Taylor expansion, assuming that $\delta_i/\hat{\mu}_i$ is small for each i,

$$\begin{aligned} D(y;\hat{\mu}) &= 2\sum_{i=1}^{n}(\hat{\mu}_i + \delta_i)\log\Big(1 + \frac{\delta_i}{\hat{\mu}_i}\Big) \\ &\approx 2\sum_{i=1}^{n}\Big(\delta_i + \frac{\delta_i^2}{\hat{\mu}_i} - \frac{\delta_i^2}{2\hat{\mu}_i}\Big) \\ &= \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}. \end{aligned}$$

The quantity in the final line is known as Pearson's χ^2 statistic.

Though we have described Pearson's χ^2 statistic as an approximation to the deviance, this does not mean that the deviance is superior in the sense that, for example, its distribution is closer to that of a χ^2_{n-p} (when the μ_i are large so small dispersion asymptotics are relevant). On the contrary, Pearson's χ^2 statistic, as its name suggest, is often to be preferred for this purpose.

3.2.2 Modelling rates with an offset

Often the expected value of a response count Y_i is proportional to a known value t_i. For instance, t_i might be an amount of time or a population size, such as in modelling crime counts for various cities. Or, it might be a spatial area, such as in modelling counts of a particular animal species. Then the sample rate is Y_i/t_i, with expected value μ_i/t_i.

In most such situations, it seems more natural to assume that it is the expected rate μ_i/t_i that is related to the covariates, rather that $\mathbb{E}(Y_i)$ itself. A log-linear model for the expected rate would then model $Y_i \sim \text{Pois}(\mu_i)$ with

$$\log(\mu_i/t_i) = \log(\mu_i) - \log(t_i) = x_i^T\beta$$

so

$$\mu_i = t_i \exp(x_i^T\beta).$$

This is the usual Poisson regression but with an *offset* of $\log(t_i)$. Since these are known constants, they can be readily incorporated into the estimation procedure.

3.2.3 Contingency tables

An r-way contingency table is a way of presenting responses which represent frequencies when the responses are classified according to r different factors.

We are primarily interested in $r = 2$ and $r = 3$. In these cases, we may write the data as

$$\begin{gathered} \{Y_{ij} : i = 1,\ldots,I,\ j = 1,\ldots,J\}, \qquad \text{or} \\ \{Y_{ijk} : i = 1,\ldots,I,\ j = 1,\ldots,J,\ k = 1,\ldots,K\} \end{gathered}$$

respectively.

Consider the example of the online survey that aimed to cross-classify individuals according to their college and voting intentions. This data could be presented as a two-way contingency table. If we also recorded people's gender, for example, we would have a three-way contingency table. A sensible model for this data is that the number of individuals falling into the ij^{th} cell, Y_{ij}, are independent $\text{Pois}(\mu_{ij})$.

Suppose we happened to end up with $n = 400$ forms filled. We could also imagine a situation where rather than accepting all the survey responses that happened to arrive in a given time, we fix the number of submissions to consider in advance, so we keep the survey live until we have 400 forms filled. In this case a multinomial model may be more appropriate.

Recall that a random vector $Z = (Z_1, \dots, Z_m)$ is said to have a multinomial distribution with parameters n and $p_1, \dots, p_m$, written $Z \sim \text{Multi}(n; p_1, \dots, p_m)$ if $\sum_{i=1}^m p_i =$ and

$$\mathbb{P}(Z_1 = z_1, \dots, Z_m = z_m) = \frac{n!}{z_1! \cdots z_m!} p_1^{z_1} \cdots p_m^{z_m},$$

for $z_i \in \{0, \dots, n\}$ with $z_1 + \cdots + z_m = n$.

In the second data collection scenario described above, only the overall total $n = 400$ was fixed, so we might model

$$(Y_{ij})_{i=1,\dots,I,\, j=1,\dots,J} \sim \text{Multi}(n; (p_{ij})_{i=1,\dots,I,\, j=1,\dots,J}),$$

where

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

At first sight, this second model might seem to fall outside the GLM framework as the responses Y_{ij} are not independent (adding up to n).

However, the following result suggests an alternative approach. Recall the **fact** that if Z_1, Z_2 are independent with $Z_i \sim \text{Pois}(\mu_i)$, then $Z_1 + Z_2 \sim \text{Pois}(\mu_1 + \mu_2)$. Obviously induction gives a similar result for any finite collection of independent Poisson random variables.

Proposition 10. *Let $Z = (Z_1, \dots, Z_m)$ be a random vector having independent components, with $Z_i \sim \text{Pois}(\mu_i)$ for $i = 1, \dots, m$. Conditional on $\sum Z_i = n$, we have that $Z \sim \text{Multi}(n; p_1, \dots, p_m)$, where $p_i = \mu_i / \sum \mu_j$ for $i = 1, \dots, m$.*

Proof. The joint distribution of $Z_1, \dots, Z_m$ is

$$\mathbb{P}_{\mu_1,\dots,\mu_m}(Z_1 = z_1, \dots, Z_m = z_m) = \exp\Big(-\sum_{i=1}^m \mu_i\Big) \prod_{i=1}^m \frac{\mu_i^{z_i}}{z_i!}, \quad z_i \in \{0, 1, \dots\},$$

and $S := \sum Z_i \sim \text{Pois}\Big(\sum \mu_j\Big)$. It follows that provided $\sum_i z_i = n$,

$$\begin{aligned}\mathbb{P}_{\mu_1,\dots,\mu_m}(Z_1 = z_1, \dots, Z_m = z_m | S = n) &= \frac{\exp\Big(-\sum \mu_j\Big) \prod(\mu_i^{z_i}/z_i!)}{\exp\Big(-\sum \mu_j\Big)\Big(\sum \mu_j\Big)^n / n!} \\ &= \frac{n!}{z_1! \dots z_m!} p_1^{z_1} \dots p_m^{z_m},\end{aligned}$$

where $p_i = \mu_i / \sum \mu_j$ for $i = 1, \dots, m$. □

Multinomial likelihood

First consider the multinomial likelihood obtained if we suppose that

$$(Y_{ij})_{i=1,\dots,I,\, j=1,\dots,J} \sim \text{Multi}(n; (p_{ij})_{i=1,\dots,I,\, j=1,\dots,J}),$$

where

$$p_{ij} = \frac{\mu_{ij}}{\sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}},$$

and

$$\log(\mu_{ij}) = \alpha + x_{ij}^T \beta.$$

Thus

$$p_{ij} = \frac{\exp(x_{ij}^T\beta)}{\sum_{i=1}^{I} \sum_{j=1}^{J} \exp(x_{ij}^T\beta)}.$$

Here the explanatory variables x_{ij} will depend on the particular model being fit.

[Consider the "colleges and voting intentions" example. Each of the 400 submitted survey forms can be thought of as realisations of i.i.d. random variables Z_l, $l = 1, \dots, 400$, taking values in the collection of categories $\{\text{Trinity}, \dots\} \times \{\text{Labour, Conservative}, \dots\}$. If we assume that the two components of the Z_l are independent, then we may write

$$p_{ij} = \mathbb{P}(Z_{l1} = \text{college}_i, Z_{l2} = \text{party}_j) = \mathbb{P}(Z_{l1} = \text{college}_i)\mathbb{P}(Z_{l2} = \text{party}_j) = q_i r_j, \qquad (3.2.1)$$

for some $q_i, r_j \geq 0$, $i = 1, \dots, I$, $j = 1, \dots, J$, with $\sum_{i=1}^{I} q_i = \sum_{j=1}^{J} r_j = 1$. To parametrise this in terms of β, we can take

$$x_{ij}^T = (\underbrace{0, \dots, 0, 1, 0, \dots, 0}_{I \text{ components}}, \underbrace{0, \dots, 0, 1, 0, \dots, 0}_{J \text{ components}}),$$

so $x_{ij}^T\beta = \beta_i + \beta_{I+j}$, and for identifiability we may take $\beta_1 = \beta_{I+1} = 0$.]

The log-likelihood for the multinomial model is

$$\begin{aligned} \ell_{\text{m}}(\beta|n) &= \sum_{i,j} y_{ij} \log\{p_{ij}(\beta)\} \\ &= \sum_{i,j} y_{ij} x_{ij}^T \beta - n \log\left(\sum_{i,j} \exp(x_{ij}^T\beta)\right), \end{aligned}$$

where we have emphasised the fact that the likelihood is based on the conditional distribution of the counts y_{ij} given the total n.

Poisson likelihood

Now consider the Poisson model, but where $\sum_{i,j} y_{ij} = n$. With $\log(\mu_{ij}) = \alpha + x_{ij}^T\beta$, we have log-likelihood

$$\begin{aligned} \ell_{\text{P}}(\alpha, \beta) &= -\sum_{i,j} \mu_{ij}(\alpha, \beta) + \sum_{i,j} y_{ij} \log\{\mu_{ij}(\alpha, \beta)\} \\ &= -\sum_{i,j} \exp(\alpha + x_{ij}^T\beta) + \sum_{i,j} y_{ij}(\alpha + x_{ij}^T\beta) \\ &= -\exp(\alpha) \sum_{i,j} \exp(x_{ij}^T\beta) + \sum_{i,j} y_{ij} x_{ij}^T\beta + n\alpha. \end{aligned}$$

Now let us reparametrise $(\alpha, \beta) \mapsto (\tau, \beta)$ where

$$\tau = \sum_{i,j} \mu_{ij} = \exp(\alpha) \sum_{i,j} \exp(x_{ij}^T \beta).$$

We have

$$\begin{aligned}\ell_{\mathrm{P}}(\tau, \beta) &= \sum_{i,j} y_{ij} x_{ij}^T \beta - n \log\left(\sum_{i,j} \exp(x_{ij}^T \beta)\right) + \{n \log(\tau) - \tau\} \\ &= \ell_{\mathrm{m}}(\beta | n) + \ell_{\mathrm{P}}(\tau).\end{aligned}$$

To maximise the log-likelihood above, we can maximise over β and τ separately. Thus if β^* is the m.l.e. from the multinomial model, and $\hat{\beta}$ is the m.l.e. from the Poisson model, we see that (assuming the m.l.e.'s are unique) $\beta^* = \hat{\beta}$. Several equivalences of the multinomial and Poisson models emerge from this fact.

- The deviances from the Poisson model and the multinomial model are the same.
- The fitted values from both models are the same. Indeed, in the multinomial model, the fitted values are

  $$n\hat{p}_{ij} := n \frac{\exp(x_{ij}^T \hat{\beta})}{\sum_{i=1}^I \sum_{j=1}^J \exp(x_{ij}^T \hat{\beta})}.$$

  For the Poisson model, the fitted values are

  $$\hat{\mu}_{ij} := \hat{\tau} \frac{\exp(x_{ij}^T \hat{\beta})}{\sum_{i=1}^I \sum_{j=1}^J \exp(x_{ij}^T \hat{\beta})}.$$

  But recall that since we have included an intercept term in the Poisson model,

  $$n = \sum_{i,j} y_{ij} = \sum_{i,j} \hat{\mu}_{ij} = \hat{\tau}.$$

Summary. Multinomial models can be fit using Poisson log-linear model provided that an intercept is included in the Poisson model. The Poisson models used to mimic multinomial models are known as *surrogate* Poisson models.

3.2.4 Test for independence of columns and rows

To test whether the rows and columns are independent (i.e. if (3.2.1) holds), we can consider a surrogate Poisson model that takes

$$\log(\mu_{ij}) = \mu + a_i + b_j,$$

where to ensure identifiability, we enforce the corner point constraints $a_1 = b_1 = 0$. Thus there are $1 + (I - 1) + (J - 1) = I + J - 1$ parameters. Provided the cell counts y_{ij} are large enough, small dispersion asymptotics can be used to justify comparing the deviance or Pearson's χ^2 statistic to $\chi^2_{IJ-I+J+1} = \chi^2_{(I-1)(J-1)}$.

3.2.5 Test for homogeneity of rows

Consider the following example. In a flu vaccine trial, patients were randomly allocated to one of two groups. The first received a placebo, the other the vaccine. The levels of antibody after six weeks were:

	Small	Moderate	Large	Total
Placebo	25	8	5	38
Vaccine	6	18	11	35

We are interested in the homogeneity of the different rows: is there a different response from the vaccine group? Here the row totals were fixed before the responses were observed. We can thus model the responses in each row as having a multinomial distribution.

If $n_i, i = 1, \dots, I$ denotes the sum of the i^{th} row, we model the response in the i^{th} row, Y_i as

$$Y_i \sim \text{Multi}(n_i; p_{i1}, \dots, p_{iJ}),$$

with $Y_1, \dots, Y_I$ independent. Note that $\sum_{j=1}^J p_{ij} = 1$ for all i.

The hypothesis of homogeneity of rows can be represented by requiring that $p_{ij} = q_j$ for all i, for some vector of probabilities $(q_1, \dots, q_J)^T$. Thus the mean in the ij^{h} cell is $\mu_{ij} := n_i q_j$.

You will discover for yourself on the example sheet that this form of multinomial model can be fitted using a surrogate Poisson model with

$$\log(\mu_{ij}) = \mu + a_i + b_j,$$

which is the same as for the independence example. For identifiability we may take $a_1 = b_1 = 0$. Here, the a_i are playing the role of intercepts for each row.

Three-way contingency tables

Now suppose we have a three-way contingency table with

$$Y \sim \text{Multi}(n; (p_{ijk}), i = 1, \dots, I,\ j = 1, \dots, J,\ k = 1, \dots, K).$$

Consider again that the table is constructed from i.i.d. random variables $Z_1, \dots, Z_n$ taking values in the categories

$$\{1, \dots, I\} \times \{1, \dots, J\} \times \{1, \dots, K\}.$$

Let us write $Z_1 = (A, B, C)$. Note that $p_{ijk} = \mathbb{P}(A = i, B = j, C = k)$. There are now eight hypotheses concerning independence which may be of interest. Broken into four classes, they are:

1. $H_1 : p_{ijk} = \alpha_i \beta_j \gamma_k$, for all i, j, k. Summing over j and k we see that $\alpha_i = \mathbb{P}(A = i)$. Thus this model corresponds to

   $$\mathbb{P}(A = i, B = j, C = k) = \mathbb{P}(A = i)\mathbb{P}(B = j)\mathbb{P}(C = k),$$

   i.e. A, B and C are independent.

2. $H_2 : p_{ijk} = \alpha_i \beta_{jk}$ for all i, j, k. As before we see that $\alpha_i = \mathbb{P}(A = i)$, and summing over i we get $\beta_{jk} = \mathbb{P}(B = j, C = k)$. This corresponds to saying A is independent of (B, C). Two other hypotheses are obtained by permutation of A, B, C.

3. $H_3 : p_{ijk} = \beta_{ij}\gamma_{ik}$ for all i, j, k. If we denote summing over an index with a '+', so for example

$$p_{i++} := \sum_{j,k} p_{ijk} = \sum_{j,k} \beta_{ij}\gamma_{ik} = \beta_{i+}\gamma_{j+},$$

   we see that

$$\mathbb{P}(B = j,\ C = k|A = i) = \frac{p_{ijk}}{p_{i++}} = \frac{\beta_{ij}}{\beta_{i+}}\frac{\gamma_{ik}}{\gamma_{i+}}$$

   Summing over k and j we see that

$$\mathbb{P}(B = j|A = i) = \frac{\beta_{ij}}{\beta_{i+}}, \qquad \mathbb{P}(C = k|A = i) = \frac{\gamma_{ik}}{\gamma_{i+}}.$$

   This means that

$$\mathbb{P}(B = j,\ C = k|A = i) = \mathbb{P}(B = j|A = i)\mathbb{P}(C = k|A = i),$$

   so B and C are conditionally independent given A. Two other hypotheses are obtained by permuting A, B, C.

4. $H_4 : p_{ijk} = \alpha_{jk}\beta_{ik}\gamma_{ij}$ for all i, j, k. This hypothesis cannot be expressed as a conditional independence statement, but means there are no three-way interactions.