Statistical modelling

Lecturer: Alberto J. Coca
alberto.j.coca@statslab.cam.ac.uk[1]

Course website: https://albertococacabrero.wordpress.com/teaching/lectures/statistical-modelling-michaelmas-2019/

Acknowledgements: the following notes (and the other materials of the course) are a lightly edited version of Dr. S. A. Bacallado's notes (and materials) which, in turn, were a lightly edited version of Dr. R. Shah's notes (and materials).

Introduction

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

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

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 $x_i = (x_{i1}, \ldots, x_{ip})^T$. 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 $x_i, i = 1, \ldots, n$ held fixed, and think of the dataset at hand as being one of the many elements of such a sequence.

In the Part II course Principles of Statistics, theory will be developed for data that are i.i.d. In our setting here, this assumption is not appropriate: the distributions of Y_i and Y_j may well be different if $x_i \neq x_j$. 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. Despite not being necessary, I encourage you to attend that course too.

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.

[1] Please let me know if you find any non-cosmetic typos. Sections between asterisks are not examinable.

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

(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 non-linear terms we may take, e.g. (quadratic),

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

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. $\text{rank}(X) = p$) so $X^T X$ is invertible (see example sheet). **We will often make this assumption below without stating it explicitly**. The fitted values, $\hat{Y} := X\hat{\beta}$ are then given by $X(X^T X)^{-1} X^T Y$. Let $P := X(X^T X)^{-1} X^T$. P is 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. **Facts:** Then $\mathbb{R}^n = V \oplus V^\perp$, so each $u \in \mathbb{R}^n$ may be written uniquely as $u = v + w$ 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 $v_{m+1}, \ldots, v_n$. Also, $(V^\perp)^\perp = V$.

Definition 1. A matrix $\Pi \in \mathbb{R}^{n \times n}$ is called an *orthogonal projection on to* $V \subseteq \mathbb{R}^n$ if $\Pi u = v$ when $u = v + w$ 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, span) of Π is V (by the first fact above and Definition 1), so $\text{rank}(\Pi) = \dim(V)$.

(ii) $I - \Pi$ is an orthogonal projection on to $V^\perp$, i.e., $I - \Pi$ 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) $\Pi^2 = \Pi = \Pi^T$, i.e., Π is idempotent and symmetric. The former is clear from the definition. To see that Π is symmetric observe that for all $u_1, u_2 \in \mathbb{R}^n$,

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

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 with ones and zeroes on its diagonal. As a result, $\text{rank}(\Pi) = \text{tr}(\Pi)$.

1.1.2 Analysis of OLS

Back to Statistics. Note that the matrix $P = X(X^TX)^{-1}X^T$ 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 $X^Tw = 0$, 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

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

Thus, the fitted values of OLS, $\hat{Y}$, are given by the projection of the vector of responses, Y, on to the column space of the matrix of predictors X.

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

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

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

the correlation matrix between Z_1 and Z_2 is the $n_1 \times n_2$ matrix with entries given by

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

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

We can now see that $\mathrm{Var}(\hat\beta_j) = \sigma^2 \|X_j^\perp\|^{-2}$. Indeed

$$\begin{aligned}\mathrm{Var}(\hat\beta_j) &= \frac{\mathrm{Var}((X_j^\perp)^T Y)}{\|X_j^\perp\|^4} \\ &= \frac{\mathrm{Var}((X_j^\perp)^T \varepsilon)}{\|X_j^\perp\|^4} \\ &= \frac{\sigma^2 (X_j^\perp)^T X_j^\perp}{\|X_j^\perp\|^4} \\ &= \sigma^2 \frac{1}{\|X_j^\perp\|^2}.\end{aligned}$$

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.

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

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

Further,

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

In fact it is the best linear unbiased estimator (BLUE), that is for any other unbiased estimator $\tilde\beta$ that is linear in Y, we have $\mathrm{Var}(\tilde\beta) - \mathrm{Var}(\hat\beta)$ is positive semi-definite.

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

Proof. Take an arbitrary linear unbiased estimator $\tilde\beta = CY$, and define D by $C = D + (X^TX)^{-1}X^T$. By unbiasedness, $\mathbb{E}\tilde\beta = DX\beta + \beta = \beta$ for all $\beta \in \mathbb{R}^p$, which implies that $DX = 0$. Then,

$$\mathrm{Var}(\tilde\beta) = \sigma^2(X^TX)^{-1} + \sigma^2 DD^T = \mathrm{Var}(\hat\beta) + \sigma^2 DD^T.$$

□

In particular, $\hat\beta$ being BLUE means that given a new observation $x^* \in \mathbb{R}^p$, we can estimate the regression function at x^* optimally in the sense that $\mathbb{E}\{(x^{*T}\hat\beta - x^{*T}\beta)^2\} \leq \mathbb{E}\{(x^{*T}\tilde\beta - x^{*T}\beta)^2\}$ for any linear and unbiased estimator $\tilde\beta$.

A closely related measure the quality of a regression procedure $\tilde\beta$ is its mean-squared prediction error (MSPE). This is defined here as

$$MSPE(\tilde\beta) := \frac{1}{n}\mathbb{E}(\|X\tilde\beta - X\beta\|^2).$$

Proposition 3.

$$MSPE(\hat\beta) = \sigma^2\frac{p}{n}.$$

Proof. Note that $X\hat\beta = PY = X\beta + P\varepsilon$, so the conclusion follows due to

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

□

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

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

1.2 Normal errors

1.2.1 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 affine transformations 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 $Z \sim N_d(\mu, \Sigma)$ when $\mathbb{E}(Z) = \mu$ and $\mathrm{Var}(Z) = \Sigma$. Note that $m + AZ \sim N_k(m + A\mu, A\Sigma A^T)$.

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 $(I - P)Y \sim N_n(0, \sigma^2(I - P))$ but it does not have a density of the form given above (unless $p = n$, not an interesting case in this course) as $I - P$ is not invertible.

Proposition 4. *Let Z_1 and Z_2 be jointly normal (i.e. (Z_1, Z_2) has a multivariate normal distribution). Then, $\mathrm{Cov}(Z_1, Z_2) := \mathbb{E}[\{Z_1 - \mathbb{E}(Z_1)\}\{Z_2 - \mathbb{E}(Z_2)\}^T] = 0$ if and only if Z_1 and Z_2 are independent.*

Proof. The *if* is immediate from the definition of covariance (and, of course, does not require the normality).

For the *only if* part, let Z_1' and 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 (Z_1', 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 $(Z_1', Z_2') \overset{d}{=} (Z_1, Z_2)$. □

χ^2 distribution

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

Proposition 5. *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

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

Student's t distribution

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

$$T \overset{d}{=} \frac{Z}{\sqrt{X/k}}$$

where Z and X are independent $N(0, 1)$ and χ_k^2 random variables respectively.

F distribution

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

$$F \overset{d}{=} \frac{X_1/k}{X_2/l}$$

where X_1 and X_2 are independent and follow χ_k^2 and χ_l^2 distributions respectively.

Notation. We will denote the upper α-points of the χ_k^2, t_k and $F_{k,l}$ distributions by $\chi_k^2(\alpha)$, $t_k(\alpha)$ and $F_{k,l}(\alpha)$ respectively. (So, for example, if $Z \sim \chi_k^2$ then $\mathbb{P}\{Z \geq \chi_k^2(\alpha)\} = \alpha$. 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

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

with appropriate independence between relevant random variables.

1.2.2 Maximum likelihood estimation

The method of least squares is just one way to construct an 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 $\hat{\theta}_{MLE}$ by choosing that $\theta \in \Theta$ which maximises the likelihood; this is called the maximum likelihood estimator (MLE). Often it is easier to work with the log-likelihood defined by

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

In the normal linear model we assume that the errors ε_i have $N(0, \sigma^2)$ distributions or, $\varepsilon \sim N_n(0, \sigma^2 I)$. Thus, we see that the likelihood for (β, σ^2) is

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2\right],$$

so the log-likelihood up to a constant 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 $(X^TX)^{-1}X^TY$, so $\hat{\beta}_{MLE} = \hat{\beta}$. The maximum likelihood estimate for σ^2 is

$$\hat{\sigma}^2_{MLE} := \frac{1}{n}\|Y - X\hat{\beta}\|^2 = \frac{1}{n}\|(I-P)Y\|^2 = \frac{1}{n}\|(I-P)\varepsilon\|^2;$$

see example sheet. Maximum likelihood does much more than simply give us another interpretation of OLS or a point estimate for σ^2. It allows us to perform *inference*: that is construct confidence interval for parameters and perform hypothesis tests.

1.2.3 Inference for the normal linear model

Distribution of $\hat{\beta}_{MLE}$ and $\hat{\sigma}^2_{MLE}$

We already know the mean and variance of $\hat{\beta}_{MLE}$ (equations (1.1.1) and (1.1.2)). As it is a linear transformation of Y, we know it must be normally distributed: $\hat{\beta}_{MLE} \sim N_p(\beta, \sigma^2(X^TX)^{-1})$. In addition, by Proposition 5,

$$\hat{\sigma}^2_{MLE} := \frac{1}{n}\|(I-P)\varepsilon\|^2 \sim \frac{\sigma^2}{n}\chi^2_{n-p}.$$

In particular, $\mathbb{E}(\hat{\sigma}^2) = (n-p)\sigma^2/n$, 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.

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 $\hat{\beta}_{MLE} = (X^TX)^{-1}X^TPY$ and $\tilde{\sigma}^2$ are independent.

Now that we know the joint distribution of $(\hat{\beta}, \tilde{\sigma}^2)$, we can easily construct confidence sets for β.

Inference for β

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

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

is a *pivot*, that is, its distribution does not depend on β or σ^2. 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

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

Note that $\prod_{j=1}^p C_j(\alpha)$ does not constitute a $1-\alpha$ confidence cuboid for the entire parameter vector β (too small generally), 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 5) and is independent of $\tilde{\sigma}^2$. Thus

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

so

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

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

$$H_0 : \beta_j = \beta_{0,j}$$
$$H_1 : \beta_j \neq \beta_{0,j}$$

and

$$H_0 : \beta = \beta_0$$
$$H_1 : \beta \neq \beta_0.$$

One can, for instance, propose the tests $\phi_j = \mathbf{1}\{\beta_{0,j} \notin C_j(\alpha)\}$ and $\phi = \mathbf{1}\{\beta_0 \notin C(\alpha)\}$; by the arguments above, these have significance level α under the respective null hypotheses H_0.

Prediction intervals

Let $x^* \in \mathbb{R}^p$ be a new observation. 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

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

We can also form a $(1-\alpha)$-level *prediction interval* for x^*, i.e., 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

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

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 $n \times p_0$ and X_1 is $n \times p - p_0$, and correspondingly $\beta_0 \in \mathbb{R}^{p_0}$ and $\beta_1 \in \mathbb{R}^{p-p_0}$. We are interested 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

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

To determine the right cut-off 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

$$\mathrm{Cov}((I-P)\varepsilon, (P-P_0)\varepsilon) = \mathbb{E}\{(I-P)\varepsilon\varepsilon^T(P-P_0)^T\} = \sigma^2(I-P)(P-P_0) = 0.$$

As the random vector

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

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.

In addition, we know that $\|(I-P)\varepsilon\|^2/\sigma^2 \sim \chi^2_{n-p}$. We now show that $\|(P-P_0)\varepsilon\|^2/\sigma^2 \sim \chi^2_{p-p_0}$. This follows from Proposition 5 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

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

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 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 (non-constant 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 and solution to practical).

(A3) $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ 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 $\hat{\varepsilon}_i$ 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, e.g., after discarding a subgroup of variables). If (A1) holds, we will have $\mathbb{E}(\hat{\varepsilon}_i) = 0$, so there should not be an obvious trend in the mean of the residuals.

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

$$\hat{\eta}_i := \frac{\hat{\varepsilon}_i}{\tilde{\sigma}\sqrt{1-p_i}}, \qquad \text{where } p_i := P_{ii} \quad \text{is the } \textit{leverage} \text{ of the } i^{th} \text{ observation}, \quad i = 1, \ldots, n.$$

The name *studentised* is due to the fact that if we replace $\tilde{\sigma}$ with an estimate $\tilde{\sigma}_{(-i)}$ derived from every sample except (x_i, Y_i), the resulting (*externally studentised*) residual has a t_{n-p-1} distribution (see example sheet). Note that $p_i < 1$ when $\hat{\varepsilon}_i \neq 0$, which will (almost surely) be the case (see example sheet). Assume $n >> p$. Then, $\tilde{\sigma}$ is a good estimate of σ and has low deviation, so the variance of $\hat{\eta}_i$ should be approximately 1. Hence, a standard check of the validity of (A2) involves plotting $\sqrt{|\hat{\eta}_i|}$ against the fitted values. One should expect a cloud of points around one, and a common evidence of violation of (A2) is that the variance of the errors increases with the fitted values. If, furthermore, (A1–A4) hold, then the studentised residuals $\hat{\eta}_i$ look roughly like an i.i.d. sample from a $N(0,1)$ distribution, and a good way of checking that the $\hat{\eta}_i$ look roughly standard normal is to look at a *Quantile–Quantile* (Q–Q) plot (see Practical 2 for details).

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, which is not necessarily desirable. The adjusted R^2, $\tilde{R}^2$ defined by

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

takes account of the number of parameters. It is generally used for model selection rather than model checking; the reason to introduced it now is for the theory to go in line with the practicals, so this paragraph will be rearranged once we cover model selection techniques.

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 and go back to the data to decide whether they really should be excluded when fitting the model. 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

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

where $p_i := P_{ii}$ is the leverage of the i^{th} observation. It measures the contribution that Y_i makes to the fitted value $\hat{Y}_i$. Since $\mathrm{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - p_i)$, 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 $\sum_{i=1}^n p_i = \mathrm{tr}(P) = p$ motivates a rule of thumb that says the influence of the i^{th} observation may be of concern if $p_i > 3p/n$.

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). 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.

Recall that a confidence ellipsoid for β is given by

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

A rule of thumb is that the influence of (Y_i, x_i) may be a cause for concern if $D_i > F_{p,n-p}(0.5)$, so removal of the i^{th} observation pushes the m.l.e. to the edge of or beyond a 50% confidence region centred on $\hat{\beta}$.

1.2.5 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

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

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 **va**riance (ANOVA). If all the n_j were equal, it would be called a balanced one-way ANOVA.

An alternative parametrisation is

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

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, α) 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, as the column space of the design matrix is unchanged.

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 **cova**riance (ANCOVA).

A word on causal inference and randomized experiments

Suppose that in the first example above, a particular exercise regime was deemed to have a significant positive effect on weight loss by ANOVA. Is it valid to infer that it was the exercise regimen that caused the weight loss? This is a delicate question that is at the heart of statistical science but we will only mention briefly in this course.

Whether we can make *causal* inferences from a regression analysis depends on how the experiment was conducted. Suppose that the subjects in the experiment were allowed to choose one of the j regimes. Then it is possible that fitter participants would choose the more difficult regimes. Suppose further that fitter participants are more likely to loose weight, regardless of the exercise regimen. Then we would expect that, even if the exercise regime had no effect, the data would suggest that more difficult regimes are associated to higher weight loss, simply because the level of fitness in these groups is higher. This is the origin of the refrain *correlation is not causation* — in this experiment, we can only prove correlation or association.

The issue here is that there is a latent variable, fitness, which is not included in the regression analysis. This variable is known as a *confounder* as it is correlated to both the input and output variables. The only way to avoid confounding is to conduct a *randomized experiment.* In such an experiment, the exercise regimes would be assigned to participants at random (by flipping a coin), in a manner independent to any potential latent variables.

Randomization, invented by C. Pierce and popularized by R. A. Fisher from the 1920s, is arguably the most important idea in Statistics. There are myriad ways in which an experiment can fail to produce causal inferences, and the design of experiments is a fascinating area of research. In fact, a lot of the terminology in regression methods, such as *design matrix*, has the flavor of randomized experiments. It is important to bear in mind that when regression methods are applied to observational data, we can only establish associations between inputs and response.

1.2.6 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, $\check{\beta}_0$, would have reduced MSPE $\sigma^2 p_0/n$, rather than $\sigma^2 p/n$. Moreover

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

for $j = 1, \dots, p_0$ (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.

We have already met two popular model selection techniques: the F-tests at the end of Section 1.2.3, and the (adjusted) coefficient of determination. We now see a few more.

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 for 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 models $\mathcal{F}_k := \{(f_{x_i,k}(\cdot;\theta_k))_{i=1}^n, \theta_k \in \Theta_k \subseteq \mathbb{R}^{p_k}\}, k = 1, \dots, K$, it is therefore desirable to select that which minimises $\bar{K}_k := \frac{1}{n}\sum_{i=1}^n K(g_{x_i}, \hat{f}_{x_i,k})$. This is equivalent to minimising

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

Of course, for a given model $\mathcal{F}$ 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) for a model $\mathcal{F}$ with log-likelihood function ℓ, 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 $\mathbb{E}(\text{AIC})/n \approx 2\mathbb{E}(\tilde{K})$ for large n, provided the true densities g_{x_i}, $i = 1, \ldots, n$ are contained in $\mathcal{F}$.

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

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

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

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

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, AIC$_c \approx$ 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

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

We say that sets of parameters are *mutually orthogonal* if the corresponding blocks of the design matrix have orthogonal column spaces. 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 $\|\hat{\beta}_j X_j\|^2$ (excluding the intercept term) in decreasing order, and pick variables corresponding to the first $p_0 - 1$ terms plus the intercept. 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,

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

where in the last equality we used that $\hat{\beta}_j = X_j^T Y / \|X_j\|^2$ by Proposition 1. 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 that 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 have 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 $\tilde{R}^2$ based criteria (or some other criterion).

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 increases 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 and that the inferential procedures we have seen fix the model(s) before performing inference.* 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.

What can we do to combat this problem? A simple option, e.g., 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.

Chapter 2

Exponential families and generalised linear models

2.1 Non-normal responses

Responses not always naturally live in $\mathbb{R}$ (even if predictors do): e.g., market prices, number of machinery failures in a factory, binary data (e.g, yes and no), etc. We elaborate on the last case.

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 $X\hat{\beta} \in [0,1]^n$. A better model may be, e.g.,

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

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

Variable transformations

A first approach to deal with non-normal responses was to transform the data so that (A1-A4) are approximately satisfied. If the responses are positive (including, e.g., natural numbers), the Box–Cox family of transformations is classical:

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

We can then use the transformed data to fit the normal linear model and a value for λ can be found by maximum likelihood estimation regarding $(\lambda, \beta, \sigma^2)$ as the parameter.

Variable transformations have the drawback that they attempt to achieve approximate normality, variance stabilisation and linearity in β of the expectation, all at once, which is generally too much to ask for from a single operation. A subsequent and superior alternative is to model the desired properties separately. Generalised linear models (GLMs) are an instance of this.

Generalised linear models (preliminaries)

Generalised linear models extend linear models to deal with situations such as those mentioned 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: $\mu_i = g(\eta_i)$, with $g = id$.

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 function g (strictly increasing and twice differentiable function as we will see).

2.2 Exponential families

We want to consider a class of distributions large enough to include the normal, Poisson, 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 at once for a number of families of distributions we may or may not have considered.

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 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 $p^y(1-p)^{1-y}$ on $\mathcal{Y} = \{0, 1\}$ for some $p \in [0, 1]$. Then, 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 (hint: take two values where it is finite and show that it is also finite at any intermediate value using convexity of e^x). If this contains an open interval about 0, then we have the series expansions

$$M(t) = \sum_{r=0}^{\infty} \mathbb{E}(Y^r) \frac{t^r}{r!},$$
$$K(t) = \sum_{r=1}^{\infty} \kappa_r \frac{t^r}{r!},$$

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

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

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

Definition 2. Let f_0 be a non-degenerate model function with c.g.f. K, and define $\Theta := \{\theta : K(\theta) < \infty\}$. Then the (exponentially tilted) family of model functions

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

is called the *natural exponential family* generated by f_0. The parameter θ is called the *natural parameter* and Θ is called the *natural parameter space.*

Remark 1. Note that we may write

$$f(y;\theta) = e^{\theta y - K(\theta)} f_0(y).$$

There are neat expressions for the mean and variance of any member of $\{f(y;\theta) : \theta \in \mathrm{int}(\Theta)\}$. We obtain these expressions by computing the first and second cumulants of $f(y;\theta)$. If $\theta, t+\theta \in \Theta$, the m.g.f. of $f(\cdot;\theta)$, $M(t;\theta)$, is

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

and its c.g.f. is $K(t,\theta) := \log M(t,\theta) = K(\theta+t) - K(\theta)$. If $\theta \in \mathrm{int}(\Theta)$, they both must be finite for t in a neighbourhood of 0. Thus, if Y has model function $f(y;\theta)$ for some $\theta \in \mathrm{int}(\Theta)$,

$$\mathbb{E}_\theta(Y) = \frac{d}{dt} K(t;\theta)\Big|_{t=0} = K'(\theta), \qquad \mathrm{Var}_\theta(Y) = \frac{d^2}{dt^2} K(t;\theta)\Big|_{t=0} = K''(\theta). \tag{2.2.1}$$

It is often useful to reparametrise the family of model functions in terms of their means, and this is what we discuss now. Fact: since f_0 was assumed to be non-degenerate, so must be every $f(y;\theta)$ for any $\theta \in \Theta$. Then, for $\theta \in \mathrm{int}(\Theta)$, $\mu(\theta) := \mathbb{E}_\theta(Y) = K'(\theta)$ is the *mean function* and satisfies $\mu'(\theta) = K''(\theta) > 0$, so μ is a smooth, strictly increasing function from $\mathrm{int}(\Theta)$ to $\mathcal{M} := \{\mu(\theta) : \theta \in \mathrm{int}(\Theta)\}$ ($\mathcal{M}$ for 'mean space'), with inverse function $\theta := \theta(\mu)$. This leads to the *mean value parametrisation*:

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

(If $\mathrm{int}(\Theta) \neq \Theta$, this can be extended by continuity to the edge point(s) of Θ, possibly setting $\mu = \pm\infty$; this will not be relevant to us though.) 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 be the standard normal density. Then $M(\theta) = e^{\theta^2/2}$ for any $\theta \in \Theta = \mathbb{R}$, so $K(\theta) = \frac{1}{2}\theta^2$ and 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 $\mu(\theta) = \theta$, $\mathcal{M} = \mathbb{R}$, $\theta(\mu) = \mu$, 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$(e^\theta), \theta \in \mathbb{R}$, family of distributions. The mean function is $\mu = e^\theta$ with mean space $\mathcal{M} = (0, \infty)$ and inverse $\theta = \log(\mu)$, and the variance function is $V(\mu) = \mu$.

2.3 Exponential dispersion families

Note that in the examples above, exponential tilting only generated new distributions by generalising the mean (in the second example the mean and variance are equal). Indeed, the natural exponential families are not broad enough for our purposes and we should like more control over the variance. In order to generalize the exponential family generated by a non-degenerate f_0 with c.g.f. K, we define the set Φ of numbers $\sigma^2 > 0$, such that $K(\cdot)/\sigma^2$ is the c.g.f. of some distribution f_{σ^2}, and in consequence, there exists an exponentially tilted model function
$$\exp\left[x\theta - K(\theta)/\sigma^2\right] f_{\sigma^2}(x).$$
Making the change of variables $y = x\sigma^2$ yields the model function
$$\frac{1}{\sigma^2}f_{\sigma^2}(y/\sigma^2)\exp\left[\frac{1}{\sigma^2}\{\theta y - K(\theta)\}\right].$$

Definition 3. An *exponential dispersion family* is a family of non-degenerate model functions,
$$f(y;\theta,\sigma^2) = a(\sigma^2, y)\exp\left[\frac{1}{\sigma^2}\{\theta y - K(\theta)\}\right], \qquad y \in \mathcal{Y}, \tag{2.3.1}$$
where

- $a(\sigma^2, y)$ is a positive function, K is the c.g.f. of a non-degenerate model function [1], and they are both known functions,

[1] *in fact, more generally, $K(\theta) = \log\int_{\mathcal{Y}} e^{\theta y}m(dy), \theta \in \Theta := \{\theta' \in \mathbb{R} : \int_{\mathcal{Y}} e^{\theta y}m(dy) < \infty\}$, where m is a σ-finite measure, such as in examples 3 and 4 below; we do not worry about this when checking if a family of model functions is an EDF and should only check that K is twice differentiable on Θ with $K'' > 0$ therein*

- the *dispersion parameter* σ^2 ranges over the set (necessarily containing 1) $\Phi := \{\sigma^2 \in (0,\infty) : K(\cdot)/\sigma^2$ is the c.g.f. of a model function[1]$\}$,
- and the *natural parameter* θ ranges over $\Theta := \{\theta' \in \mathbb{R} : K(\theta') < \infty\}$, which is assumed to be open (and necessarily contains 0).

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 model function in (2.3.1) is

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

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

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

Comparing this to (2.2.1), we see that exponential dispersion families generalise a given model function f_0 as exponential tilting plus an extra multiplicative parameter in the variance.

As before, we may define $\mu(\theta) := K'(\theta)$ and $\mathcal{M} := \{\mu(\theta) : \theta \in \Theta\}$. 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, $\theta(\mu)$ and the 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
$$f(y;\nu,\tau^2) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{y^2}{2\tau^2}\right) \exp\left\{\frac{1}{\tau^2}\left(\nu y - \frac{1}{2}\nu^2\right)\right\}, \qquad y \in \mathbb{R},$$
showing that this family is an exponential dispersion family with $\theta = \nu$, $\Theta = \mathbb{R}$, $\sigma^2 = \tau^2$, $\Phi = (0,\infty)$, $K(\theta) = \theta^2/2$, $\mu(\theta) = \theta$, $\mathcal{M} = \mathbb{R}$, $\theta(\mu) = \mu$, and $V(\mu) = 1 > 0$ ($\mu(\theta)$ and $V(\mu)$ can be obtained either by directly looking at the mean and variance or by differentiating $K(\theta)$). One can check that this is the exponential dispersion family generated by the standard normal distribution.

2. Consider the family Pois(λ), where $\lambda \in (0,\infty)$. We write their p.m.f.'s as
$$f(y;\lambda) = e^{-\lambda}\frac{\lambda^y}{y!} = \frac{1}{y!}\exp\left(y \log\lambda - \lambda\right), \qquad y \in \{0,1,2,\ldots\},$$
so this is an exponential dispersion family with $\theta = \log\lambda$, $\Theta = \mathbb{R}$, $\sigma^2 = 1$, $\Phi = \{1\}$, $K(\theta) = e^\theta$, $\mu(\theta) = e^\theta$, $\mathcal{M} = (0,\infty)$, and $V(\mu) = K''(\theta(\mu)) = \mu > 0$. One can check that this is the exponential dispersion family generated by the Pois(1) distribution, and so we see no difference with its natural exponential family (as somehow expected).

3. Let $Z \sim \mathrm{Bin}(n,p)$, where $n \in \mathbb{N}, p \in (0,1)$. Then $Y := Z/n \sim \frac{1}{n}\mathrm{Bin}(n,p)$ has p.m.f.
$$f(y;n,p) = \binom{n}{ny} p^{ny}(1-p)^{n(1-y)}, \qquad y \in \{0, 1/n, 2/n, \ldots, 1\}.$$

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

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

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

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

so $\mathcal{M} = (0,1)$ and its inverse is $\theta(\mu) = \log\{\mu/(1-\mu)\}$. Differentiating once more we see that

$$\begin{aligned} V(\mu) &= \frac{(1+e^{\theta(\mu)})e^{\theta(\mu)} - (e^{\theta(\mu)})^2}{(1+e^{\theta(\mu)})^2} \\ &= \frac{e^{\theta(\mu)}}{1+e^{\theta(\mu)}}\left(1 - \frac{e^{\theta(\mu)}}{1+e^{\theta(\mu)}}\right) \\ &= \mu(1-\mu) > 0. \end{aligned}$$

4. Consider the family $\Gamma(\alpha,\beta)$, where $\alpha,\beta \in (0,\infty)$, with densities

$$f(y;\alpha,\beta) = \frac{\beta^\alpha y^{\alpha-1}e^{-\beta y}}{\Gamma(\alpha)}, \qquad y \in (0,\infty). \tag{2.3.2}$$

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 density (2.3.2) then

$$\mathbb{E}_{\alpha,\beta}(Y) = \alpha/\beta \quad \text{and} \quad \mathrm{Var}_{\alpha,\beta}(Y) = \alpha/\beta^2.$$

If this family were an exponential dispersion family then

$$\mu = \alpha/\beta \quad \text{and} \quad \sigma^2 V(\mu) = \alpha/\beta^2.$$

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 $\alpha = \beta$, so $\sigma^2 V(\mu) = \sigma^2 \times \text{constant} \propto 1/\beta = 1/\alpha$. Thus we must have $\sigma^2 = \alpha^{-1}$ (or some constant multiple of it). In the new parametrisation where $\alpha = \sigma^{-2}$ and $\beta = (\mu\sigma^2)^{-1}$,

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

so this is an exponential dispersion family with $\theta = -\beta/\alpha$, $\Theta = (-\infty,0)$, $\sigma^2 = 1/\alpha$, $\Phi = (0,\infty)$, $K(\theta) = \log(-\theta^{-1})$, $\mu(\theta) = -\theta^{-1}$, $\mathcal{M} = (0,\infty)$, $\theta(\mu) = -\mu^{-1}$ and $V(\mu) = \theta(\mu)^{-2} = \mu^2$.

2.4 Generalised linear models

Definition 4. 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 monotone (generally increasing), twice differentiable function called the *link function.*

Note we must take g monotone for identifiability and twice differentiability will allow us to form quadratic approximations to the log-likelihood which will be useful for both computation and inference.

2.4.1 Choice of link function

Typically, g is chosen so that both computation and interpretability of the estimates of β are relatively simple. 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 $\mathcal{M} = (0, 1)$. Two popular choices for g in this situation are the *probit* function, i.e. the inverse of the standard normal probability distribution function, and the *logit* function, i.e. $g(\mu) = \log\{\mu/(1 - \mu)\}$ $(= \theta(\mu)$, where is the inverse of the mean function).

In a generalised linear model, 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 is the canonical link when the Y_i have scaled binomial distributions as above. As we see next, the canonical link function renders the log-likelihood concave and the likelihood equations simple, allowing for the construction of efficient algorithms to perform inference.

2.4.2 Likelihood equations

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

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

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

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

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). Indeed,

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

and $K'' > 0$ (note that if the design matrix $X = (x_1, \ldots, x_n)^T$ is full-rank it follows that the Hessian is negative definite and the log-likelihood is strictly concave). Thus, if

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

$\hat\beta$ must be a maximiser of the log-likelihood, and if it belongs to

$$\{\beta' \in \mathbb{R}^p : g^{-1}(x_i^T\beta') \in \mathcal{M} \quad \forall i = 1, \ldots, n\} = \{\beta' \in \mathbb{R}^p : x_i^T\beta' \in \Theta \quad \forall i = 1, \ldots, n\}$$

it is a maximum likelihood estimator and is characterised through the *likelihood equations*

$$\sum_{i=1}^n \frac{x_i}{\sigma^2 a_i}\{y_i - K'(x_i^T\beta)\} = 0.$$

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.

Consider data $(Y_1, x_1), \ldots, (Y_n, x_n)$ 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 more general setting than generalised linear models. Here we simply aim to sketch out the main results; for a rigorous treatment of the simpler case of i.i.d. data see your Principles of Statistics notes (the theory generalises without much changes). 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.

2.5.1 The score function

Let $\hat{\theta}$ be the maximum likelihood estimator of θ (assuming it exists and is unique, as guaranteed by the regularity conditions). 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 under regularity conditions, which ensure 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 easy 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: $Y = X\beta + \varepsilon$, $\varepsilon \sim N_n(0, \sigma^2 I_n)$. Then (see example sheet)

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

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

In fact we have the following result.

Theorem 6 (Cramér–Rao lower bound)**.** *Let $\tilde{\theta}$ be an unbiased estimator of $\theta \in int(\Theta)$. Then under regularity conditions,*

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

is positive semi-definite.

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

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

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, $\hat{\beta} := (X^T X)^{-1} X^T Y$ is unbiased and this model satisfies the regularity assumptions, 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 is 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.

We say a sequence of random vectors $Z_m \in \mathbb{R}^k$ *converges in distribution* to a random vector $Z \in \mathbb{R}^k$ if

$$\mathbb{P}(Z_m \in B) \to \mathbb{P}(Z \in B) \quad \text{as } m \to \infty$$

for all (Borel) sets B for which $\mathbb{P}(Z \in \partial B) = 0$, where $\partial B := \bar{B} \setminus \mathrm{int}(B)$.

For example, the multidimensional central limit theorem (CLT) states that if $Z_1, Z_2, \ldots$ are i.i.d. random vectors in $\mathbb{R}^k$ with 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

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

Asymptotic normality of maximum likelihood estimators

Theorem 7. *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 $I(\theta)$. Then denoting the maximum likelihood estimator of $\theta \in int(\Theta)$ when there are n observations by $\hat{\theta}^{(n)}$, under regularity conditions we have*

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

Equivalently,

$$i^{1/2}(\theta)(\hat{\theta}^{(n)} - \theta) \xrightarrow{d} N_d(0, I_d).$$

A short-hand and informal version of writing this (which is fine for this course, but should not be taken as common practice outside of it) is that

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

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

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 it turns out that

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

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

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

where z_α is the upper α-point of $N(0,1)$. The coverage of this confidence interval tends to $1 - \alpha$ as $n \to \infty$, i.e. $\mathbb{P}_{\theta_j}(\theta_j \in C_j(\alpha)) \to 1 - \alpha$ and $\mathbb{P}_\theta(\theta \in C(\alpha)) \to 1 - \alpha$ as $n \to \infty$. Similarly, an asymptotic $1 - \alpha$ level confidence set for θ is given by

$$C(\alpha) := \{\theta' : (\hat{\theta} - \theta')^T i(\hat{\theta})(\hat{\theta} - \theta') \leq \chi_d^2(\alpha)\}.$$

With these, we can perform hypothesis tests of the form $H_0 : \theta_j = \theta_{0,j}$ versus $H_1 : \theta_j \neq \theta_{0,j}$, and $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ for some $\theta_{0,j} \in \mathbb{R}$ and $\theta_0 \in \mathbb{R}^d$. One can, for instance, use the tests $\phi_j = \mathbf{1}\{\theta_{0,j} \notin C_j(\alpha)\}$ and $\phi = \mathbf{1}\{\theta_0 \notin C(\alpha)\}$; by the arguments above, these have asymptotic significance level α under the respective null hypotheses H_0, i.e. $\mathbb{E}_{\theta_{0,j}}\phi_j \to \alpha$ and $\mathbb{E}_{\theta_0}\phi \to \alpha$ as $n \to \infty$.

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 nothing. You will have had it drilled into you long ago in Analysis I that even the first trillion terms of a sequence need not say anything about 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. 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

$$\begin{aligned} H_0 &: \theta_1 = 0 \qquad \text{against} \\ H_1 &: \theta_1 \neq 0. \end{aligned}$$

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

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

Theorem 8 (Wilks' theorem)**.** *Suppose that H_0 is true. Then, under regularity conditions*

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

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

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 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 $i_\beta(\beta, \sigma^2)$ for the $p \times p$ top left submatrix of $i(\beta, \sigma^2)$ and $i_{\sigma^2}(\beta, \sigma^2)$ for the bottom right entry, we have

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

Whether σ^2 is known or unknown in a generalised linear model depends on which model we have chosen. When σ^2 is unknown we may estimate it as

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

which can be motivated by recalling that

$$\mathbb{E}\{(Y_i - \mu_i)^2\} = \sigma^2 a_i V(\mu_i).$$

For notational simplicity let us set $\tilde{\sigma}^2 = \sigma^2$ when σ^2 is known. Then, a study of the asymptotic behaviour of $\tilde{\sigma}^2$ together with the asymptotic results we have studied show that

$$\hat{\beta} \sim AN_p(\beta, i_\beta^{-1}(\hat{\beta}, \tilde{\sigma}^2)),$$

justifying the following asymptotic $(1-\alpha)$-level confidence interval for β_j:

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

where z_α is the upper α-point of $N(0,1)$. Since in practice we have finite data, it can be argued that when n is moderately large and σ^2 is unknown, replacing $z_{\alpha/2}$ by $t_{n-p}(\alpha/2)$ gives a generally superior confidence interval for β_j. Tests for β_j can be constructed immediately from these intervals.

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 $\hat{\beta}, \check{\beta} \in \mathbb{R}^p$ for the m.l.e. of β under the alternative and the null models, where in the latter $\theta = (\beta_0^T, \mathbf{0}^T)^T$ with $\mathbf{0} = (0, \dots, 0)^T \in \mathbb{R}^{p-p_0}$. Define $\tilde{\ell}(\mu, \sigma^2)$ by

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

Note that

$$\max_{\mu \in \mathbb{R}^n} \tilde{\ell}(\mu, \sigma^2) = \tilde{\ell}(y, \sigma^2),$$

provided the maximum on the LHS exists. Thus $\tilde{\ell}(y, \sigma^2)$ is the maximised log-likelihood of the so-called saturated model which imposes no restrictions on the μ_i. Define $\hat{\mu}, \check{\mu} \in \mathbb{R}^n$ by

$$\hat{\mu}_i = g^{-1}(x_i^T \hat{\beta}), \qquad \check{\mu}_i = g^{-1}(x_i^T \check{\beta}).$$

Note that, for example,

$$\tilde{\ell}(\hat{\mu}, \sigma^2) = \max_{\mu \in \mathbb{R}^n : \mu_i = g^{-1}(x_i^T \beta), \text{ some } \beta \in \mathbb{R}^p} \tilde{\ell}(\mu, \sigma^2) = \ell(\hat{\beta}, \sigma^2).$$

The *deviances* of the models corresponding to H_1 and H_0 are

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

The deviance may be thought of as the appropriate generalisation to GLMs of the residual sum of squares from the linear model and we remark that it does not depend on σ^2. Note that the deviance is reduced in the larger model.

Notice that

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

so by Wilks' theorem we may test if $\beta_1 = 0$ at asymptotic level α by rejecting the null hypothesis when the value of this test statistic is larger than $\chi^2_{p-p_0}(\alpha)$. When n is moderately large and σ^2 is unknown, we use[2] $(p - p_0) F_{p-p_0, n-p}(\alpha)$ as the critical value.

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

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

[2] Typo: I added the term $(p - p_0)$

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 would 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, $m \in \mathbb{N}$, we update

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

We terminate the algorithm when successive iterations produce negligible difference in the log-likelihood. 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. Thus Fisher scoring updates

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + i^{-1}(\hat{\beta}^{(m-1)})U(\hat{\beta}^{(m-1)}).$$

Fisher scoring may not necessarily converge to $\hat{\beta}$ but almost always does.

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

$$U_j(\beta) = \sum_{i=1}^{n} \frac{(y_i - \mu_i)X_{ij}}{a_i\sigma^2 V(\mu_i)g'(\mu_i)} \qquad j = 1, \dots, p,$$
$$i_{jk}(\beta) = \sum_{i=1}^{n} \frac{X_{ij}X_{ik}}{a_i\sigma^2 V(\mu_i)\{g'(\mu_i)\}^2} \qquad k = 1, \dots, p.$$

Choosing the canonical link $g(\mu) = \theta(\mu)$ simplifies $U_j(\beta)$ and $i_{jk}(\beta)$ since $\theta'(\mu) = 1/V(\mu)$.

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

$$R_i(\mu) = g'(\mu_i)(y_i - \mu_i).$$

Then we may write

$$
\begin{aligned}
U(\beta) &= \sigma^{-2} X^T W(\mu) R(\mu) \\
i(\beta) &= \sigma^{-2} X^T W(\mu) X,
\end{aligned}
$$

where μ has $\mu_i = g^{-1}(x_i^T \beta)$. Let us set

$$
\begin{aligned}
W^{(m)} &:= W(\hat{\mu}^{(m)}) \\
R^{(m)} &:= R(\hat{\mu}^{(m)}),
\end{aligned}
$$

where $\hat{\mu}_i^{(m)} = g^{-1}(x_i^T \hat{\beta}^{(m)})$. Then we see that

$$
\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + (X^T W^{(m-1)} X)^{-1} X^T W^{(m-1)} R^{(m-1)}.
$$

If we define the *adjusted dependent variable* $Z^{(m)}$ by

$$
Z^{(m)} := \hat{\eta}^{(m)} + R^{(m)},
$$

where $\hat{\eta}^{(m)} = X\hat{\beta}^{(m)}$, then

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

See example sheet 1 question 3 for the final equality. Note that $g(y_i) \approx g(\hat{\mu}_i^{(m)}) + g'(\hat{\mu}_i^{(m)})(y_i - \hat{\mu}_i^{(m)}) =: Z^{(m)}$ if $\hat{\mu}_i^{(m)} \approx y_i$. Thus the sequence of approximations to $\hat{\beta}$ are given by the *iterative reweighted least squares* (IRLS) of the adjusted dependent variable where the (expected) local curvature of the log-likelihood is incorporated through the weights.

We can take the initial guess $\hat{\beta} = 0$ or, with this alternative formulation, 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 y, in which case $Z^{(0)} = g(y) := (g(y_1), \dots, g(y_n)))^T$ and $W^{(0)} = W(\hat{\mu}^{(0)})$.

2.7 Model checking

Model checking for GLMs 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. The *raw residuals*, $Y_i - \hat{\mu}_i$ tend not to be the most useful for model checking as their variances are not constant.

The *Pearson residuals* are one attempt to correct for this and are defined by

$$
e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\tilde{\sigma}^2 a_i V(\hat{\mu}_i)}},
$$

where $\tilde{\sigma}^2$ is defined as in Section 2.5.4. If $\hat{\mu}_i \approx \mu_i$ then we should have $\mathbb{E}(e_i) \approx 0$ and $\mathrm{Var}(e_i) \approx 1$. In fact, for some generalised linear models we have that the Pearson residuals have an approximate centred normal distribution (with variance ≈ 1 but generally less than 1; the standarised Pearson residuals defined below will have an approximate standard normal distribution), so it is common to examine normal Q–Q plots to check these models and to spot out unusual observations. Two important examples where this occurs are

- $Y_i \sim \frac{1}{n_i}\mathrm{Bin}(n_i, p_i)$, with n_i large for all i, and

- $Y_i \sim \text{Pois}(\mu_i)$ with μ_i large for all i.

This is because in these cases, the (appropriately centred and rescaled) individual Y_i get close to normally distributed random variables as $\min_i\{n_i\}, \min_i\{\mu_i\} \to \infty$, and one can show that the Pearson residuals will also be close to normally distributed and $\sum_{i=1}^n e_i^2 \to^d \chi^2_{n-p}$ (n fixed). The latter is known as the *Pearson chi-squared statistic*, and can be used for goodness-of-fit tests. These asymptotics are called *small dispersion asymptotics* and, informally, they arise due to growing information in the fixed number of observations rather than information growing from increasingly many observations.

Without applying asymptotics,

$$\text{var}(Y_i - \hat{\mu}_i) \approx \sigma^2 a_i V(\mu_i)(1 - h_i(\mu)),$$

where $h_i(\mu) := H_{ii}(\mu)$ with $H(\mu)$ an $n \times n$ (projection) matrix given by

$$H(\mu) := W^{1/2}(\mu) X \left(X^T W(\mu) X\right)^{-1} X^T W^{1/2}(\mu).$$

Let the *leverage* of the i^{th} observation be $\hat{h}_i := h_i(\hat{\mu})$ and the *standarised Pearson residuals* be

$$r_i = \frac{e_i}{\sqrt{1 - \hat{h}_i}}.$$

In comparison to the normal linear model, it still holds that $h_i \in [0, 1]$ and $\sum_{i=1}^n h_i = p$ (since H is a projection matrix) but, in a generalised linear model, observations with extreme values of its covariates may not have high leverage. As a measure of influence we can take *Cook's distance* of the i^{th} observation,

$$D_i = \frac{1}{p} \frac{\hat{h}_i}{1 - \hat{h}_i} r_i^2.$$

Another form of residuals 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

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

where $D_i(y_i; \hat{\mu}_i)$ (not to be confused with Cook's distance) is the i^{th} summand in the definition of $D(y; \hat{\mu})$, so

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

and $\sum_{i=1}^n d_i^2 = D(y; \hat{\mu})$. Under small dispersion asymptotics, the d_i converge to normal distributions, and, if σ^2 is known, the fitted model can be tested against the *saturated* model, i.e.,

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

because, in this case,

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

converges to a χ^2_{n-p} distribution (n fixed).

2.8 Model selection

AIC applies to generalised linear models without modification, and so do the (non-examinable) warnings about inference after model selection. Procedures based on residual sums of squares, e.g., the coefficient of determination or forward and backward selection, can be generalised to generalised linear models by replacing the residual sums of squares by the appropriate deviances. E.g., if $\bar{\mu}$ are the fitted means for the intercept-only model, the coefficient of determination can be generalised by

$$R^2 = \frac{D(y; \bar{\mu}) - D(y; \hat{\mu})}{D(y; \bar{\mu})}.$$

Orthogonality does not generalise to generalised linear models.

Chapter 3

Specific regression problems

3.1 Binomial regression

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

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

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 genuine email (called 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 IRLS 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.

There is an interesting latent variable interpretation of the model with any of these three links (and for more general links) whenever $n_i = 1$ for all $i = 1, \ldots, n$: $Y_i = \mathbb{1}_{\{Y_i^* > 0\}}$, where the unobserved/latent variable is given by $Y_i^* = x_i^T\beta + \varepsilon_i$ with $\varepsilon_i \sim^{i.i.d.} F$ for different c.d.f.s F for each of the links; indeed,

$$\mu_i = \mathbb{E}Y_i = \mathbb{P}(Y_i^* > 0) = \mathbb{P}(\varepsilon_i > -x_i^T\beta) = 1 - F(-x_i^T\beta),$$

and, defining $F^{-1}(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\}$, it follows that

$$\eta_i := x_i^T\beta = -F^{-1}(1-\mu_i).$$

Note, furthermore, that if F is symmetric about the origin, i.e. $F(x) = 1 - F(-x)$ for all $x \in \mathbb{R}$,

$$\eta_i = F^{-1}(\mu_i).$$

Thus, we can interpret the coefficients of each of these models as the effects of a unit increase in the corresponding variable on a latent variable whose sign gives us the observed response and which follows a linear model with errors following log-Weibull, standard normal and standard logistic distributions, respectively (you do not need to know anything about the first and last distributions).

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 coefficients from a model with logit link (a *logistic regression* model) have an additional and easier interpretation. 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

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

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. When the responses are counts, it may be sensible to model them as realisations of Poisson random variables. A word of caution though. As with all GLMs, 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. Indeed, if the probability of occurrence of an event in a given time interval is proportional to the length of that time interval and independent of the occurrence of other events, then the number of events in any specified time interval will be Poisson distributed.

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 $Y_i \sim \text{Pois}(\mu_i)$, $\mu_i > 0$. It follows from the second example in Section 2.3 that $\{\text{Pois}(\mu) : \mu \in (0, \infty)\}$ is an exponential dispersion family with dispersion parameter $\sigma^2 = 1 = a_i$ for all $i = 1, \ldots, n$.

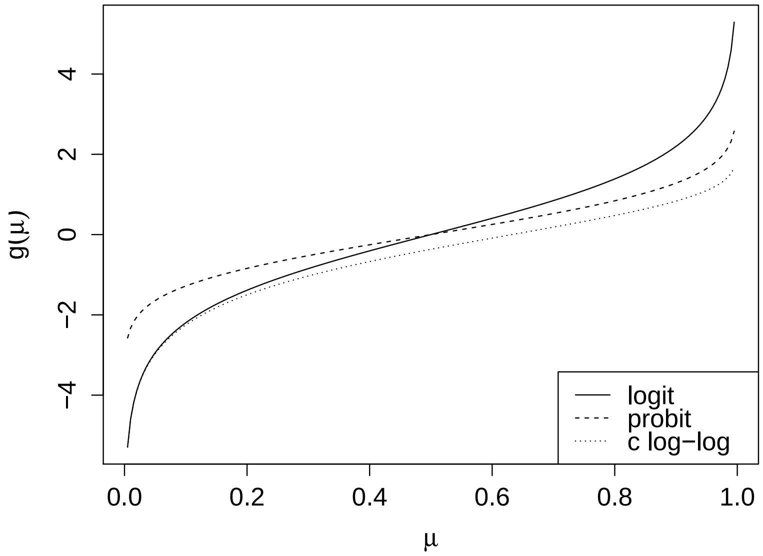

Figure 3.1: The graphs of three commonly used link functions for binomial regression.

3.2.1 Link functions

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. 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.

3.2.2 The deviance and Pearson's χ^2-statistic

In a Poisson GLM we have

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

In an example sheet you are asked to show that, if an intercept term is included in a GLM with $\sigma^2 = a_i = 1$ and with a canonical link, then, writing $\hat{\mu}_i := \exp(x_i^T\hat{\beta})$,

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

Therefore,

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

when an intercept term is included. Furthermore, 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 Pearson's χ^2 statistic for this model. Recall from Section 2.7 that both Pearson's χ^2 statistic and the deviance converge to a χ^2_{n-p} distribution under small dispersion asymptotics (so, herein, as the μ_i diverge) if the dispersion parameter is known (as here). For finite data, it is often better to use Pearson's χ^2 statistic for inference in the sense that its distribution is closer to the χ^2_{n-p} distribution.

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

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

respectively.

Consider count data arising from conducting an (online) survey where 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 together the data into a 2-way *contingency table*:

	Labour	Conservative	Liberal Democrats	Other
Darwin	×	×	×	×
Clare	×	×	×	×
⋮				

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 Pois(μ_{ij}), where $\log(\mu_{ij}) = \alpha + x_{ij}^T\beta$ (note that herein we model the intercept α explicitly). Then, we have log-likelihood (up to additive constants)

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

Multinomial model

Rather than accepting all the survey responses that happened to arrive in a given time, an alternative experimental design would be to fix the number of submissions to consider in advance, so we keep the survey live until we have n forms filled. In this case a multinomial model may be more appropriate.

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

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

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

Thus, we now might model

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

where

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

and

$$\log(\mu_{ij}) = x_{ij}^T \beta,$$

so that

$$p_{ij} = \frac{\exp(x_{ij}^T \beta)}{\sum_{i'=1}^I \sum_{j'=1}^J \exp(x_{i'j'}^T \beta)}$$

(note that, without loss of generality, we may take no intercept in this model).

Here the explanatory variables x_{ij} will depend on the particular model being fitted. Consider the "colleges and voting intentions" example. Each of the n submitted survey forms can be thought of as realisations of i.i.d. random variables Z_l, $l = 1, \ldots, n$, taking values in the collection of categories {Trinity, Selwyn, . . .} × {Labour, Conservative, . . .}. 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 $q_i, r_j \geq 0$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, with $\sum_{i=1}^I q_i = \sum_{j=1}^J r_j = 1$. To parametrise this in terms of $\beta = (\log(q_1), \ldots, \log(q_I), \log(r_1), \ldots, \log(r_J))^T$, we can take

$$x_{ij}^T = (\underbrace{0, \ldots, 0, \overbrace{1}^{i)}, 0, \ldots, 0}_{I \text{ components}}, \underbrace{0, \ldots, 0, \overbrace{1}^{I+j)}, 0, \ldots, 0}_{J \text{ components}}),$$

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 (up to additive constants)

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

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

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).

Connection between Poisson and multinomial models

Now let us reparametrise $(\alpha, \beta) \mapsto (\tau, \beta)$ in the Poisson log-linear model, where

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

Assuming $\sum_{i,j} y_{ij} = n$ we have (up to additive constants)

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

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 expected information matrices are the same.
- The fitted values from both models are the same. Indeed, in the multinomial model, the fitted values are

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

  whilst 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 fitted 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.*

Underlying this relationship is the following result.

Proposition 9. *Let $Z = (Z_1, \ldots, Z_m)$ be a random vector having independent components, with $Z_i \sim \mathrm{Pois}(\mu_i)$ for $i = 1, \ldots, m$. Conditional on $\sum Z_i = n$, we have that $Z \sim \mathrm{Multi}(n; p_1, \ldots, p_m)$, where $p_i = \mu_i / \sum \mu_j$ for $i = 1, \ldots, m$.*

Proof. Recall the **fact** that if $Z_i \overset{ind.}{\sim} \mathrm{Pois}(\mu_i)$, then $S := \sum Z_i \sim \mathrm{Pois}\Big(\sum \mu_j\Big)$. It follows that provided $\sum_i z_i = n$,

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

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

Tests for independence

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

$$\log(\mu_{ij}) = \alpha_i + \beta_j,$$

where to ensure identifiability, we enforce the corner point constraint $\beta_1 = 0$ (or, equivalently, $\log(\mu_{ij}) = \mu + \alpha_i + \beta_j$ with $\alpha_1 = \beta_1 = 0$). Thus there are $I + J - 1$ parameters. Provided the cell counts y_{ij} are large enough and the above model is true, small dispersion asymptotics can be used to justify comparing the deviance or Pearson's χ^2 statistic to $\chi^2_{IJ-I-J+1} = \chi^2_{(I-1)(J-1)}$.

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 as follows:

1. $H_1 : p_{ijk} = q_i r_j s_k$, for all i, j, k. Summing over j and k we see that $q_i = \mathbb{P}(A = i)$. By the analogous argument, $r_j = \mathbb{P}(B = j)$ and $s_k = \mathbb{P}(C = k)$. 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 mutually independent.

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

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

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

   we see that

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

   Summing over k and j we see that

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

   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. $H_4 : p_{ijk} = q_{jk}r_{ik}s_{ij}$ for all i, j, k. This hypothesis cannot be expressed as any independence statement, but means there are no three-way interactions.

These models can be tested using the appropriate surrogate Poisson models and small dispersion asymptotic results if the cell counts are large enough. For instance, for H_3 we may equivalently test the model $Y_{ijk} \sim^{ind.} \text{Pois}(\mu_{ijk}), \sum_{ijk} Y_{ijk} = n, \log \mu_{ijk} = \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \epsilon_{ik}$ with corner point constraints $\beta_1 = \gamma_1 = \delta_{i1} = \delta_{1j} = \epsilon_{i1} = \epsilon_{1k} = 0$, where $i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, K$; its deviance and Pearson's χ^2 statistic should be approximately distributed as a χ^2_{IJK-L} distribution, with $L = I + J - 1 + K - 1 + (I - 1)(J - 1) + (I - 1)(K - 1)$.

"By-row" multinomial model

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 as follows:

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

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.

Then, if n_i is the sum of the i^{th} row, $i = 1, \dots, I$, we may model the response in the i^{th} row, Y_i, as

$$Y_i \sim^{ind.} \text{Multi}(n_i; p_{i1}, \dots, p_{iJ}), \quad p_{ij} = \frac{\mu_{ij}}{\sum_{j'=1}^{J} \mu_{ij'}}, \quad \log \mu_{ij} = x_{ij}^T \beta, \quad i = 1, \dots, I, j = 1, \dots, J.$$

In the example sheet, you will show that one can instead fit a surrogate Possion model for each row, i.e.,

$$Y_{ij} \sim^{ind.} \text{Pois}(\mu_{ij}), \qquad \sum_{j=1}^{J} Y_{ij} = n_i, \qquad \log \mu_{ij} = \alpha_i + x_{ij}^T \beta.$$

Here, the α_i are playing the role of intercepts for each row.

In these settings we are usually interested in the homogeneity of the different rows: is there a different response from the vaccine group? 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^{th} cell is $\mu_{ij} := n_i q_j$ or, equivalently, $\log \mu_{ij} = \alpha_i + \beta_j$, where for identifiability we may take $\alpha_1 = \beta_1 = 0$, so it is the same as for the two-way independence example

3.3 The delta method

The delta method states, roughly, that smooth functions of asymptotically normal estimators are also asymptotically normal. This can be used to perform inference of functions of the coefficients β in a GLM which may not be linear or monotonic.

Theorem 10. *Adopt the general setting of Section 2.5. If $f : \mathbb{R}^p \to \mathbb{R}^d$ has a derivative ∇f which is continuous at $\beta \in \mathbb{R}^p$ and*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Sigma)$$

for some $\Sigma \in \mathbb{R}^{p \times p}$ (may depend on β), then

$$\sqrt{n}(f(\hat{\beta}) - f(\beta)) \xrightarrow{d} N\left(0, [\nabla f]^T \Sigma \nabla f\right).$$

The intuition behind this fact is that the distribution of the estimator $\hat{\beta}$ concentrates quickly around the true parameter β, and in a small neighbourhood of β we can approximate f with a linear function. Therefore, $f(\hat{\beta})$ is approximately a linear function of an approximately normal vector. We give a non-examinable proof of the one dimensional version of this theorem.

**Proof (case $p = d = 1$)*.* By a Taylor series expansion around β with mean-value remainder,

$$f(\hat{\beta}) = f(\beta) + \nabla f(\lambda\hat{\beta} + (1-\lambda)\beta)(\hat{\beta} - \beta)$$

for some $\lambda \in [0,1]$. Therefore,

$$\sqrt{n}(f(\hat{\beta}) - f(\beta)) = \nabla f(\lambda\hat{\beta} + (1-\lambda)\beta)\sqrt{n}(\hat{\beta} - \beta).$$

On the right hand side, $\sqrt{n}(\hat{\beta} - \beta)$ converges to $N(0, \Sigma)$ in distribution by assumption. Furthermore, $\hat{\beta}$ converges to β in probability. By the continuous mapping theorem, $\nabla f(\lambda\hat{\beta} + (1-\lambda)\beta)$ converges to $\nabla f(\beta)$ in probability. Applying Slutsky's theorem then yields the desired result. □

We can apply this to an example in Poisson regression with data (y_i, x_i) for $i = 1, \ldots, n$, with the canonical link function. Suppose the response y_i is the number of immune cells which can be seen on a slide under a microscope, and x_i is a vector of characteristics of the slide which might predict the response. For a new slide with predictors x^* we want to perform inference on the probability that the response $Y^* \sim \text{Pois}(\mu^*)$, where Y^* is independent from the previous responses and $\log \mu^* = (x^*)^T\beta$, is smaller or equal to some $k \in \mathbb{N}$. This can be written as

$$f_k(\beta) = \mathbb{P}(Y^* \leq k) = e^{-\mu^*} \sum_{j=0}^{k} \frac{(\mu^*)^j}{j!}.$$

The delta method and the asymptotic normality of MLE tell us that our fitted value $f_k(\hat{\beta})$ is asymptotically normal; more specifically,

$$\sqrt{n}(f_k(\hat{\beta}) - f_k(\beta)) \xrightarrow{d} N\left(0, [\nabla f_k(\beta)]^T I^{-1}(\beta) \nabla f_k(\beta)\right),$$

where $I(\beta) = \lim_{n\to\infty} i(\beta)/n$ and $i = i^{(n)}$ is the Fisher information matrix with n observations. As we know the Fisher information matrix for Poisson regression, all that remains to find the asymptotic variance is to compute the derivative of f_k. Applying the chain rule

$$\begin{aligned}
\nabla_\beta f_k(\beta) &= -\nabla_\beta \mu^* e^{-\mu^*} \sum_{j=0}^{k} \frac{(\mu^*)^j}{j!} + e^{-\mu^*} \sum_{j=1}^{k} \nabla_\beta \mu^* \frac{(\mu^*)^{j-1}}{(j-1)!} \\
&= \left[P(Y^* \leq k-1) - P(Y^* \leq k)\right] \nabla_\beta \mu^* \\
&= -P(Y^* = k)\mu^* x^* \\
&= -e^{(x^*)^T\beta - e^{-(x^*)^T\beta}} \frac{((x^*)^T\beta)^k}{k!} x^*.
\end{aligned}$$

Then, a $(1-\alpha)$-level asymptotic confidence interval for $P(Y^* \leq k)$ is given by, e.g.,

$$\left[f_k(\hat{\beta}) - z_{\alpha/2}\sqrt{\nabla_\beta f_k(\hat{\beta})^T i^{-1}(\hat{\beta}) \nabla_\beta f_k(\hat{\beta})},\, f_k(\hat{\beta}) + z_{\alpha/2}\sqrt{\nabla_\beta f_k(\hat{\beta})^T i^{-1}(\hat{\beta}) \nabla_\beta f_k(\hat{\beta})} \right].$$