Part II — Statistical Modelling

Based on lectures by A. J. Coca

Notes taken by Joseph Tedds using Dexter Chua's header and Gilles Castel's snippets.

Michaelmas 2019

These notes are not endorsed by the lecturers, and I have modified them (often significantly) after lectures. They are nowhere near accurate representations of what was actually lectured, and in particular, all errors are almost surely mine.

Introduction to the statistical programming language R
Graphical summaries of data, e.g. histograms. Matrix computations. Writing simple functions. Simulation. [2]

Linear Models
Review of least squares and linear models. Characterisation of estimated coefficients, hypothesis tests and confidence regions. Prediction intervals. Model selection. BoxCox transformation. Leverages, residuals, qq-plots, multiple $\mathbb{R}^2$ and Cooks distances. [5]

Overview of basic inferential techniques
Asymptotic distribution of the maximum likelihood estimator. Approximate confidence regions. Wilks theorem. The delta method. Posterior distributions and credible intervals. [3]

Exponential dispersion families and generalised linear models (glm)
Exponential families and meanvariance relationship. Dispersion parameter and generalised linear models. Canonical link function. Iterative solution of likelihood equations. Regression for binomial data; use of logit and other link functions. Poisson regression models, and their surrogate use for multinomial data. Application to 2- and 3-way contingency tables. Hypothesis tests and model selection, including deviance and Akaikes Information Criterion. Residuals and model checking. [8]

Examples in R
Linear and generalised linear models. Interpretation of models, inference and model selection. [6]

Contents

0 Introduction **3**

1 Linear Models **4**
1.1 Ordinary least squares (OLS) 4
1.2 Orthogonal projection 5
1.3 Analysis of OLS 5
1.4 Normal Errors 8
1.4.1 Multivariate normal and related distributions 8
1.4.2 Maximum likelihood estimation 10
1.4.3 Inference for the normal linear model 11
1.4.4 Testing significance of groups of variables 12
1.4.5 Model checking 14
1.5 ANOVA & ANCOVA 15
1.6 Model selection 16
1.7 Inference after model selection 18

2 Exponential families and generalised linear models **19**
2.1 Non-normal responses 19
2.2 Exponential Families 19
2.3 EDFs 21
2.4 GLMs 22
2.4.1 Choice of the link function 23
2.4.2 Likelihood equations 23
2.5 Inference 24
2.5.1 The score function 24
2.5.2 Fisher information 25
2.5.3 Asymptotic normality of MLE and Wilks' theorem 25
2.5.4 Inference in GLMs 26
2.6 Computation 27
2.7 Model checking 28
2.8 Model selection 29

3 Specific regression problems **30**
3.1 Binomial regression 30
3.1.1 Link functions 30
3.2 Poisson regression 30
3.2.1 Link functions 31
3.2.2 Deviance and Pearson chi-squared statistic 31
3.2.3 Contingency tables 31

0 Introduction

This course is unusual in that 8 of the lectures are taken as practicals, with the following guidance.

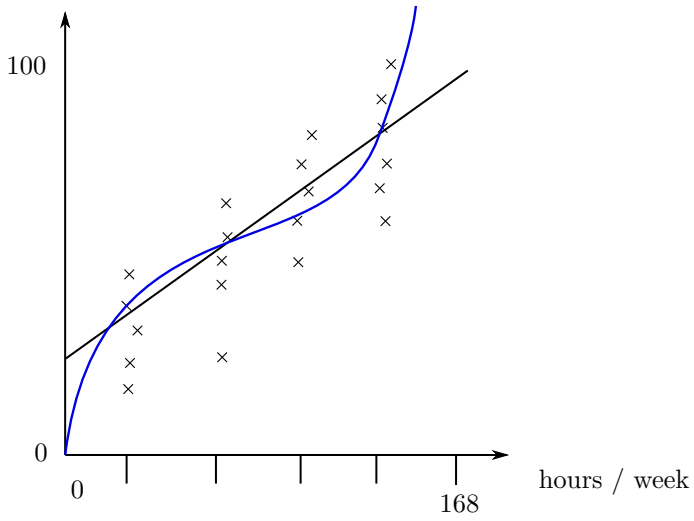
- Ideally use Linux, some things may not work on other operating systems
- Use R and R Studio

We study Data:

- $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n), i = 1, \ldots, n$, $n =$ sample size.
- $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ - predictors, covariates , independent or explanatory variables.
- y_i - targets, responses, dependent variables.

Objective: understand the functional relationship relating the y_i's to the $\mathbf{x}_i$'s to develop a regression function.

Example. $x_i =$ number of hours / week student i invests in on statistical modelling, $y_i =$ final grade of student i.



In the next section we model the Y's (note they are now upper-case) as random variables, as $Y_i = f(x_i, \theta) + \varepsilon_i$ independent.

- f is linear in θ
- $\varepsilon_i \approx$ errors / noise with potential causes as measurement errors or our limited understanding of the world.
- $\mathbb{E}[Y_i \mid X_i] = f(x_i, \theta) + \mathbb{E}[\varepsilon_i \mid x_i]$

In the sections thereafter, $\mathbb{E}[Y_i \mid x_i] = f_i(x_i, \theta)$, f_i is not necessarily linear in θ.

Warning. A word of caution, statistical models are not a perfect representation of the world, but they are useful approximations to make decisions.

1 Linear Models

1.1 Ordinary least squares (OLS)

Consider the linear regression model $Y = X\beta + \epsilon$,

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathbb{R}^{n \times p}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_n \end{pmatrix} \in \mathbb{R}^p, \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \in \mathbb{R}^n.$$

where

(i) $\mathbb{E}[\epsilon_i] = 0$ - does not mean unbiased, but centred

(ii) $\operatorname{var} \epsilon_i = 0$ - homoskedastic

(iii) $\operatorname{cov}(\epsilon_i, \epsilon_j) = 0$ - uncorrelated = linear independence $\neq$ independence

Definition (Design matrix)**.** The design matrix X, unless otherwise stated : $p \leq n$, and $\mathbf{X}$ is full rank i.e. rank $X = p$.

Note, $\theta = \beta$ in the introduction. If we want intercept,

$$X = \begin{pmatrix} 1 & x_1^T \\ \vdots & \vdots \\ 1 & x_n^T \end{pmatrix}, \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \in \mathbb{R}^{p+1}$$

.
If we want higher order terms e.g. quadratic

$$X = \begin{pmatrix} 1 & x_1^T & x_{11}^2 & \cdots & x_{1p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^T & x_{n1}^2 & \cdots & x_{np}^2 \end{pmatrix}, \theta = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \in \mathbb{R}^{2p+1}.$$

Remember, linear means linear in θ

Definition (Least squares)**.** The *least squares estimator*, $\hat{\beta}$ is defined as

$$\hat{\beta} = \underset{\mathbf{b} \in \mathbb{R}^n}{\operatorname{argmin}} \|Y - X\mathbf{b}\|^2.$$

On the example sheet, we will show that $\hat{\beta} = (X^T X^{-1}) X^T Y$

The fitted values are given by

$$\hat{Y} = X\hat{\beta} = \overbrace{X(X^T X^{-1}) X^T}^{P} Y = PY.$$

We call P the 'hat' matrix and it is an orthogonal projection onto the column space of X.

1.2 Orthogonal projection

Let $V \subseteq \mathbb{R}^n$ be linear. Its orthogonal complement is

$$V^\perp = \{\omega \in \mathbb{R}^n : \omega^T \cdot \mathbf{v} = 0 \ \forall \ \mathbf{v} \in V\}.$$

Fact. (i) $\mathbb{R} \cong V \oplus V^\perp$, so $\forall\ \mathbf{u} \in \mathbb{R}^n\ \exists\ \mathbf{v} \in V, \omega \in V^\perp$ such that $\mathbf{u} = \mathbf{v} + \mathbf{w}$

(ii) $(V^\perp)^\perp = V$

Definition (Orthogonal projection)**.** $\pi \in \mathbb{R}^{n\times n}$ is an *orthogonal projection* onto V if $\pi\mathbf{u} = \mathbf{v}$ whenever $\mathbf{u} = \mathbf{v} + \mathbf{w}, \mathbf{v} \in V, \mathbf{w} \in V^\perp$. π is an orthogonal projection if it is an orthogonal projection onto its column space.

Let π be an orthogonal projection onto V, properties

(i) The column space /range / image/ span of π is V (immediate from the fact above and the definition) so rank $\pi = \dim V$.

(ii) $I - \pi$ is an orthogonal projection onto $V^\perp$. Let $\mathbf{u} = \mathbf{v} + \mathbf{w}, \mathbf{v} \in V, \mathbf{w} \in v^\perp$,

$$(I - \pi)\mathbf{u} = \mathbf{0} + \mathbf{w}.$$

(iii) π is idempotent ($\pi^2 = \pi$) and π is symmetric ($\pi^T = \pi$). The former is by definition, the latter

$$\forall\ \mathbf{u}_1, \mathbf{u}_2 \in \mathbb{R}^n, (\pi\mathbf{u}_1)^T(I - \pi)\mathbf{u}_2 = \begin{cases} 0 \\ \mathbf{u}_1^T((\pi^T - \pi^T\pi)\mathbf{u}_2 \end{cases}.$$

$$\pi^T = \pi^T\pi \iff \pi i = \pi^T\pi = \pi^T.$$

In fact $\pi^2 = \pi = \pi^T$ is an alternative definition for an orthogonal projection. Let $\mathbf{v} \in$ span π.

$$\exists\ \mathbf{u} \in \mathbb{R}^n \pi\mathbf{u} = \mathbf{v} \implies \pi\mathbf{v} = \pi^2\mathbf{u} = \pi\mathbf{u} = \mathbf{v}.$$

Let $\mathbf{v} \in (\text{span } \pi)^\perp$,

$$\pi\mathbf{v} = \pi^T\mathbf{v} = 0.$$

(iv) Orthogonal bases of V and $V^\perp$ are eigenvectors π with eigenvalues 1 or 0. Thus, $\pi = \mathbf{u}D\mathbf{u}^T, \mathbf{u}$ orthonormal ($\mathbf{u}^T\mathbf{u} = \mathbf{u}\mathbf{u}^T = I$), D is diagonal matrix with 1's and 0's

1.3 Analysis of OLS

Recall

$$\hat{\beta} = \arg\min_{\mathbf{b}\in\mathbb{R}^p} \|Y - X\mathbf{b}\|^2 = \left(X^TX\right)^{-1} X^TY.$$

and $\hat{Y} = X\hat{\beta} = PY$

$$PX\mathbf{b} = X(X^TX)^{-1}X^TX\mathbf{b} = X\mathbf{b}.$$

If $w \in (\text{span } \pi)^\perp$

$$Pw = X\left(X^TX\right)^{-1}\underbrace{X^T\mathbf{w}}_{\mathbf{0}} = \mathbf{0}.$$

P is an orthogonal projection onto span X, $\hat{Y}$ is a projection of Y onto span X.

The reverse is true: if π is an orthogonal projection onto V, then

$$\pi\mathbf{u} = \arg\min_{\mathbf{v}\in V}\|\mathbf{u}-\mathbf{v}\|^2 \ \forall\ \mathbf{u}\in\mathbb{R}^n.$$

$\hat{\beta}$ is the vector of coefficients of the closest vector in span X to Y as a linear combination of the columns of X Alternative representation of OLS:
Let $X_j = (X_{\cdot j}), X_{-j}$ is X without X_j, $P_{-j} = X_{-j}\left(X_{-j}^T X_{-j}\right)^{-1} X_{-j}^T$

Proposition. Let $X_j^\perp = (I - P_{-j})X_j$. Then

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

Proof.

$$\begin{aligned}
\left(X_j^\perp\right)^T Y &= X_j^T(I-P_{-j})Y \\
&= X_j^T(I-P_{-j})(PY+(I-P)Y) \\
&= X_j^T(I-P_{-j})PY + X_j^T \underbrace{X_j^T(I-P_{-j})(I-P)Y}_{X_j^T(I-P)Y=0} \\
&= X_j^T(I-P_{-j})PY
\end{aligned}$$

With the reduction of the second term coming from $V_{-k}^\perp \subseteqq V^\perp$

$$\begin{aligned}
\left(X_j^\perp\right)^T &= X_j^T(I-P_{-j})X \\
&= X_j^T \begin{pmatrix} 0 & \cdots & 0 & \underset{j^{\text{th}}\text{ element}}{(I-P_{-j})} & 0\cdots & 0 \end{pmatrix} \\
&= \begin{pmatrix} 0 & \cdots & 0 & X_j^T(I-P_{-j})^2X_j & 0 & \cdots & 0 \end{pmatrix} \\
&= \begin{pmatrix} 0 & \cdots & 0 & \|X_j^\perp\|^2 & 0 & \cdots & 0 \end{pmatrix}
\end{aligned}$$

Hence

$$\left(X_j^\perp\right)^T Y = \left(X_j^\perp\right)^T X\hat{\beta} = \begin{pmatrix} 0 & \cdots & 0 & \|X_j^\perp\|^2\beta_j & 0 & \cdots & 0 \end{pmatrix}.$$

□

Recall if $\mathbf{z}_i \in \mathbb{R}^{n_i}, i = 1, 2$

$$\operatorname{cov}(\mathbf{z}_1,\mathbf{z}_2) = \mathbb{E}\left[(\mathbf{z}_1 - \mathbb{E}[\mathbf{z}_1])(\mathbf{z}_2 - \mathbb{E}[\mathbf{z}_2]\right].$$

$$(\operatorname{cov}(\mathbf{z}_1,\mathbf{z}_2))_{ij} = \frac{\operatorname{cov}(\mathbf{z}_1,\mathbf{z}_2)_{ij}}{\sqrt{(\operatorname{var}\mathbf{z}_1)_{ii}(\operatorname{var}\mathbf{z}_2)_{jj}}}.$$

$$\forall\ \mathbf{a}_i\in\mathbb{R}^{n_i},\quad \operatorname{cov}(\mathbf{z}_1+\mathbf{a}_1,\mathbf{z}_2+\mathbf{a}_2) = \operatorname{cov}(\mathbf{z}_1,\mathbf{z}_2).$$

and if $A\in\mathbb{R}^{d\times n}, b\in\mathbb{R}^d$

$$\mathbb{E}[\mathbf{b}+A\mathbf{z}_1] = \mathbf{b} + A\mathbb{E}[\mathbf{z}_1].$$

Then,

$$\begin{aligned}\operatorname{var}\hat{\beta}_j &= \frac{1}{\|X_j^{\perp}\|^4}\operatorname{var}\left(X_j^{\perp}\right)^T Y \\ &= \frac{1}{\|X_j^{\perp}\|^4}\operatorname{var}(X_j^{\perp})^T\varepsilon \\ &= \frac{1}{\|X_J^{\perp}\|^4}\left(X_j^{\perp}\right)^T \operatorname{var}\varepsilon\ X_j^{\perp} \\ &= \frac{\sigma^2}{\|X_j^{\perp}\|^2}\end{aligned}$$

.

Now, $\hat{\beta} \in \mathbb{R}^p$, $\hat{\beta}$ is unbiased. Indeed

$$\mathbb{E}_\beta\left[\hat{\beta}\right] = \mathbb{E}\left[(X^TX)^{-1}X^T(X\beta+\epsilon)\right] = \beta.$$

.

$$\begin{aligned}\operatorname{var}(\hat{\beta}) &= \operatorname{var}((X^TX)^{-1}XY) \\ &= \operatorname{var}((X^TX)^{-1}X^T\epsilon) \\ &= (X^TX)^{-1}X^T\underbrace{\operatorname{var}\epsilon}_{\sigma^2 I}X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1}\end{aligned}$$

.

This is optimal in the following sense.

Theorem (Gauss-Markov)**.** $\hat{\beta}$ is BLUE (Best Linear Unbiased Estimator) i.e. $\forall\ \tilde{\beta}$ linear (in Y) and unbiased estimator

$$\operatorname{var}\tilde{\beta} - \operatorname{var}\hat{\beta} \text{ is positive semidefinite.}$$

Proof. Let $\tilde{\beta} = CY = \hat{\beta} + \underbrace{(C-(X^TX)^{-1}X^T)}_{D}Y$. Note $\mathbb{E}_\beta\left[\tilde{\beta}\right] = \beta\ \forall\ \beta \in \mathbb{R}^p$, so

$$0 = DX\beta \implies DX = 0$$

as this is true $\forall\ \beta$. Then,

$$\operatorname{var}\tilde{\beta} = \operatorname{var}\hat{\beta} + \operatorname{var}DY + 2\operatorname{cov}(\hat{\beta}, DY).$$

$$\operatorname{var}DY = \operatorname{var}D\epsilon = \sigma^2 DD^T,$$

which is positive semi-definite by definition.

.

$$\begin{aligned}\operatorname{cov}(\hat{\beta}, DY) &= \operatorname{cov}((X^TX)^{-1}X^T\epsilon, D\epsilon) \\ &= (X^TX)^{-1}\underbrace{X^TD^t}_{=0}\sigma^2\end{aligned}$$

.

□

Consequently, if x^* is a new observation

Exercise.

$$\mathbb{E}\left[((x^*)^T\hat{\beta} - (x^*)^T\beta)^2\right] \leq \mathbb{E}\left[\left((x^*)^T\tilde{\beta} - (x^*)^T\beta\right)^2\right] \ \forall\ \tilde{\beta} \text{ LUE}.$$

We can also measure the quality of a regression procedure $\tilde{\beta}$ by its mean-squared prediction error:

$$\text{MSPE}(\tilde{\beta}) = \frac{1}{n}\mathbb{E}\left[\|X\tilde{\beta} - X\beta\|^2\right].$$

Proposition.

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

Proof. First note that

$$X\hat{\beta} = PY = X\beta + P\epsilon.$$

$$\|X\hat{\beta} - X\beta\|^2 = \|P\epsilon\|^2 = \epsilon^T P\epsilon^T = \text{Tr}(\epsilon^T P\epsilon) = \text{Tr}(P\epsilon^T\epsilon).$$

$$\begin{aligned}\mathbb{E}\left[\text{LHS}\right] &= \text{Tr}(P\mathbb{E}\left[\epsilon^T\epsilon\right] \\ &= \sigma^2\text{Tr}P \\ &= \sigma^2 p\end{aligned}$$

.

□

Lastly, $\hat{\epsilon} = Y - \hat{Y} = (I-P)Y$ is the vector of residuals. This satisfies

$$\begin{aligned}\text{cov}(\hat{\epsilon}, \hat{Y}) &= \text{cov}((I-P)\epsilon, P\epsilon) \\ &= \sigma^2 X(X^TX)^{-1}\underbrace{X^T(I-P)^T}_{=0} = 0\end{aligned}$$

.

So $\hat{\epsilon}$ and $\hat{Y}$ are uncorrelated.

1.4 Normal Errors

1.4.1 Multivariate normal and related distributions

Definition (Multivariate normal)**.** $Z \in \mathbb{R}^d$ is *multivariate normal* if $\forall\ t \in \mathbb{R}^d, t^TZ$ is univariate normal. Thus $\forall\ m \in \mathbb{R}^{k, A\in\mathbb{R}^{k\times d}}, m + AZ$ is (multivariate) normal.

Fact. Normal distributions are uniquely characterised by their mean and variance. So write

$$Z \sim N_d(\mu, \Sigma).$$

if $\mathbb{E}\left[Z\right] = \mu, \text{var}\, Z = \Sigma$

$$\implies m + Az \sim N_k(m + A\mu, A\epsilon A^T).$$

If Σ is invertible , Z has density

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

Proposition. Let Z_1, Z_2 be jointly normal (i.e. (Z_1, Z_2) is multivariate normal).

$$\operatorname{cov}(Z_1, Z_2) = 0 \iff Z_1, Z_2 \text{ are independent } (Z_1 \perp Z_2).$$

Proof. The backward direction is immediate.
In the other direction, let $Z_1' \overset{d}{Z_1}, Z_2' \overset{d}{=} Z_2$. Note

$$\mathbb{E}\left[(Z_1', Z_2')\right] = \left(\mathbb{E}\left[Z_1'\right], \mathbb{E}\left[Z_2'\right]\right) = \mathbb{E}\left[(Z_1, Z_2\right].$$

$$\operatorname{var}(Z_1', Z_2') = \begin{pmatrix} \operatorname{var} Z_1' & \operatorname{cov}(Z_1', Z_2') \\ \operatorname{cov}(Z_1', Z_2') & \operatorname{var} Z_2' \end{pmatrix} = \begin{pmatrix} \operatorname{var} Z_1 & 0 \\ 0 & \operatorname{var} Z_2 \end{pmatrix} = \operatorname{var}(Z_1, Z_2).$$

Also, (Z_1', Z_2') is normal because sums of independent normals is normal. Thus the conclusion follows by the fact above. □

Definition (χ^2 distribution)**.** $X \sim \chi^2_k$ (on k degrees of freedom), if

$$X \overset{d}{=} \sum_{j=1}^{k} Z_j^2, \quad Z_j \overset{\text{iid}}{\sim} N(0,1).$$

Proposition. Let $\pi \in \mathbb{R}^{n\times n}$ be an orthogonal projection with rank k and $\epsilon \sim N_n(0, \sigma^2 I)$. Then

$$\|\pi\epsilon\|^2 \sim \sigma^2 \chi^2_k.$$

Proof. Recall that $\pi = UDU^T$ and noting that $U^T\epsilon \sim N_n(0, \sigma^2 I)$. Then ,

$$\begin{aligned} \|\pi\epsilon\|^2 &= \epsilon^T UDU^T UDU^T \epsilon \\ &= \|Du^T\epsilon\|^2 \\ &\overset{d}{=} \|D\epsilon\|^2 \\ &= \sum_{j:D_{jj}=1} \epsilon_j^2 \\ &\overset{d}{=} \sigma^2 \sum_{j:D_{jj}=1} Z_j^2 \end{aligned}$$

.

□

Definition (t-student distribution)**.** $T \sim t_k$ (on k degrees of freedom) if

$$T \overset{d}{=} \frac{Z}{\sqrt{X/k}}, Z \sim N(0,1), X \sim X_k^2 \text{ independent.}$$

Definition (F distribution)**.** $F \sim F_{k,l}$ (on k, l degrees of freedom) if

$$F \overset{d}{=} \frac{X_1/k}{X_2/l}, X_1 \sim \chi^2_k, X_2 \sim \chi^2_l \text{ independent.}$$

1.4.2 Maximum likelihood estimation

Let $Y \in \mathbb{R}^n$ has density $f(\cdot, \theta), \theta \in \Theta \subseteq \mathbb{R}^p$ unknown (Θ paraneter space unknown). If data y is a realisation of Y, the likelihood function is

$$L(\theta) = f(y : \theta), \theta \in \Theta.$$

Then

$$\hat{\theta}_{\text{MLE}} = \arg\min_{\theta \in \Theta} L(\theta).$$

It is usually easier to work with the log-likelihood $\ell(\theta) = \log L(\theta)$. Many times we define them up to constants

Example. The t-statistic is given by

$$\frac{\hat{\beta} - \beta}{\sqrt{\text{var}(\hat{\beta})}}.$$

In the second practical we look at

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

Where

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

and $\hat{\sigma}^2$ is the MLE for σ^2.

Assume that $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ or $\varepsilon \sim N_n(0, \sigma^2)$. Then

$$Y = X\beta + \varepsilon \sim N_n(X\beta, \sigma^2 I)$$

and the likelihood is

$$L(\beta, \sigma^2) = \frac{1}{\sqrt{(2\pi)^n \sigma^{2n}}} \exp\left(\frac{1}{2\sigma^2}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta)\right), \beta \in \mathbb{R}^p, \sigma^2 > 0.$$

Thus the log-likelihood (up to constants) is

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

It follows that

$$\hat{\beta}_{\text{MLE}} = \hat{\beta} = (X^T X)^{-1} X^T Y$$

and

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

Both estimators are intuitive under our model assumptions without $\varepsilon \sim N_n(0, \sigma^2 I)$ necessarily. However, the distributional assumptions on ε induce distributions on the estimators, which will allow us to perform inference.

1.4.3 Inference for the normal linear model

Distributions of $\hat{\beta}_{\text{MLE}}$ and $\hat{\sigma^2}_{\text{MLE}}$:
Note

$$\hat{\beta}_{\text{MLE}} = \beta + (X^T X)^{-1} X^T \varepsilon \sim N_p(\beta, \sigma^2 (X^T X)^{-1})$$

also

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

In particular, $\mathbb{E}\left[\hat{\sigma^2}_{\text{MLE}}\right] = \frac{\sigma^2}{n}(n-p) \implies \hat{\sigma^2}_{\text{MLE}}$ is biased (but asymptotically unbiased). So let

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

What can we say about their joint distribution? Recall that PY and $(I-P)Y$ are uncorrelated;

$$\begin{pmatrix} PY \\ (I-P)Y \end{pmatrix} = \begin{pmatrix} P \\ I - P \end{pmatrix} Y$$

is a linear transformation of Y (normal). Then we know by our earlier work that $PY \perp (I-P)Y$. Note $\hat{\beta} = (X^T X)^{-1} X^T PY \implies \hat{\beta} \perp \tilde{\sigma^2}$.
Inference for β :
Note that

$$\frac{\hat{\beta} - \beta}{\sqrt{\tilde{\sigma^2}}} = \frac{N_p(0, (X^T X)^{-1})}{\sqrt{\chi^2_{n-p}/(n-p)}},$$

with the numerator and denominator independent. This is a *pivot* i.e. id does not depend on the unknown parameters (β, σ^2) and hence can be used for inference.

Example.

$$C_j(\alpha) := \{b \in \mathbb{R} : \left| \frac{\hat{\beta}_j - b}{\sqrt{\tilde{\sigma^2}(X^T X)^{-1}_{jj}}} \right| \leq t_{n-p}(\frac{\alpha}{2})\} \quad j = 1. \ldots, p,$$

where if $T \sim t_{n-p}$, then

$$\mathbb{P}\left(-t_{n-p}(\frac{\alpha}{2}) \leq T \leq T_{n-p}(\frac{\alpha}{2})\right) = 1 - \alpha.$$

Since

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

then $\mathbb{P}_{\beta,\sigma^2}(\beta_j \in C_j(\alpha)) = 1 - \alpha$. So $C_j(\alpha)$ is a $(1-\alpha)$-confidence interval for β_j. Note,

$$\prod_{j=1}^{p} C_j(\alpha)$$

is not a $(1-\alpha)$-confidence interval cuboid for β (too small). Though

$$\prod_{j=1}^{p} C_j(\frac{\alpha}{p})$$

has a confidence / coverage of $\geq 1-\alpha$. However, the latter is too large / conservative generally.

A smaller (and exact) alternative. Consider

$$\|X(\hat{\beta}-\beta\|^2 = \|P\varepsilon\|^2 \sim \sigma^2\chi^2_p$$

independent of $\tilde{\sigma^2}$. Let

$$C(\alpha) = \{\mathbf{b} \in R^p : \frac{\|X(\hat{\beta}-\beta\|^2}{\tilde{\sigma^2}} \leq F_{p,n-p}(\alpha)\}.$$

If $F \sim F_{p,n-p}$, then $\mathbb{P}(F \leq F_{p,n-p}(\alpha)) = 1-\alpha$. Then $\mathbb{P}_{\beta,\sigma^2}(\beta \in C(\alpha)) = 1-\alpha$. Note, the same arguments allows us to construct hypotheses tests.

Example. We can test $H_0 : \beta_j = \beta_{j,0}$ v.s. $H_1 : \beta_j \neq \beta_{j,0}$ with

$$\phi_j = 1\{\beta_{j,0} \notin C_j(\alpha)\}.$$

We can also test $H_0 : \beta = \beta_0$ v.s. $\beta \neq \beta_0$ with

$$\phi = 1\{\beta_0 \notin C(\alpha)\}.$$

Prediction Intervals:
Let $\mathbf{x}^*$ be a new observation. Note

$$\mathbf{x}^{*T}(\hat{\beta}-\beta) = \mathbf{x}^{*T}(X^TX)^{-1}X^T \sim N\left(0, \mathbf{x}^{*T}(X^TX)^{-1}\mathbf{x}^*\right),$$

and

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

can be used to perform inference for the regression function at $\mathbf{x}^*$ i.e. $x^{*T}\beta$. Let $Y^* = x^*\beta + \varepsilon^*, \varepsilon \sim N(0,\sigma^2)$ is independent of ε. We can also construct a $(1-\alpha)$-prediction interval for Y^* i.e. a random interval I depending only on Y such that $\mathbb{P}(Y^* \in I) = 1-\alpha$. Indeed,

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

So we can use the pivot

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

Note that the confidence intervals for Y^* will be larger than those for $x^{*T}\beta$ because of the additional variability / uncertainty coming from ϵ^*.

1.4.4 Testing significance of groups of variables

Let

$$X = (X_0.X_1), \beta = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, X_0 \in \mathbb{R}^{n\times p_0}, X_1 \in \mathbb{R}^{n\times(p-p_0)}, \beta_0 \in \mathbb{R}^{p_0}, \beta_1 \in \mathbb{R}^{p-p_0}.$$

Without loss of generality, we wish to test $H_0 : \beta_1 = 0$ v.s. $H_1 : \beta_1 \neq 0$. We can construct a generalised likelihood ratio test : recall if Y has density $f(y, \theta), \theta \in \Theta$ unknown, the likelihood ratio test for $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \notin \Theta_0$, where $\Theta_0 \subseteq \Theta$ reject H_0 for large values of

$$\omega_{LR} := 2 \log \frac{\sup_{\theta' \in \Theta} L(\theta')}{\sup_{\theta' \in \Theta_0} L(\theta')} = 2(\sup_{\theta' \in \Theta} \ell(\theta') - \sup_{\theta' \in \Theta_0} \ell(\theta').$$

Notation. Write $\check{\beta}_0$ and $\check{\sigma}^2$ for the MLEs under the null i.e. $Y = X_0\beta_0 + \varepsilon$

$$\varepsilon \sim N_n(0, \sigma^2 I), \text{ so } \check{\beta}_0 = (X_0^T X_0)^{-1} X_0^T Y, \check{\sigma}^2 = \frac{1}{n}\|Y - X_0\check{\beta}_0\|^2.$$

Then

$$\omega_{LR} = 2(-\frac{n}{2}\log\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}\|Y - X\hat{\beta}^2\| + \frac{n}{2}\log\check{\sigma}^2 + \frac{1}{2\check{\sigma}^2}\|Y - X_0\check{\beta}_0\|^2 = n\log(\|(I-P_0)Y\|^2/\|(I-P)Y\|^2).$$

Note $I - P_0 = I - P + P - P_0$ and $PP_0 = P_0$

$$(I - P)(P - P_0) = 0 \implies \|(I - P_0)Y\|^2 = \|(I - P)Y\|^2 + \|(P - P_0)Y\|^2$$

and

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

We claim that $P - P_0$ is an orthogonal projection. Indeed, $P - P_0$ is symmetric and by $P_0P = P_0^T P = (PP_0)^T = P_0^T = P_0$, so

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

Also note that if π is an orthogonal projection, then $\text{rank}\pi = \text{tr}\pi$ so,

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

and we conclude that $\|(P - P_0)\varepsilon\|^2 \sim \sigma^2\chi^2_{p-p_0}$. Note

$$\text{cov}((I - P)\varepsilon, (P - P_0)\varepsilon) = \sigma^2(I - P)(P - P_0) = 0$$

and

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

is a linear transformation of ε (normal) so $(I - P)\varepsilon \perp (P - P_0)\varepsilon$. So we take the test

$$\phi = 1\{\frac{\|(P - P_0)Y\|^2/(p - p_0)}{\|(I - P)Y\|^2}/(n - p) \geq F_{p-p_0,n-p}(\alpha)\}.$$

This test has a significance level of α.

1.4.5 Model checking

The validity of our inferential conclusions depends on our assumptions about the distribution of ε being correct. If any of them fail we have the following remarks:

Remark. (i) $\mathbb{E}[\varepsilon_i] = 0$. If not, the coefficients in the linear model should be interpreted with care ($Y = X\beta + \mu + (\varepsilon - \mu)$), $\tilde{\sigma}^2$ is inflated, F-tests will have the correct size (example sheet) but they may lose power.

(ii) $\operatorname{var}\varepsilon_i = \sigma^2$. If not, $\hat{\beta}$ is not as efficient as it could be, and confidence sets and hypothesis test levels may not be correct. If the variance are known up to a multiplicative constant, then do weighted least squares (ex sheet).

(iii) $\operatorname{cov}(\varepsilon_i, \varepsilon_j) = 0 \ \forall \ i, j$. This usually occurs with temporal / spatial data, confidence sets and levels may be wrong.

(iv) ε is normal. If the first assumptions hold, then the inferential procedures hold asymptotically thanks to the central limit theorem.

Any violation of the above may be checked by looking at

$$\hat{\epsilon} = Y - X\hat{\beta} = (I - P)Y.$$

To check $\mathbb{E}[\varepsilon_i] = 0$ plot $\hat{\varepsilon}$ against $\hat{Y}$ (and sometimes v.s. the covariance), if the assumption holds we should expect to see no trends.
To check $\operatorname{var}\varepsilon_i$ (and $\operatorname{cov}(\varepsilon_i, \varepsilon_j)$) note that $\operatorname{var}\hat{\varepsilon} = \sigma^2(I - P)$ so define the studentised residuals

$$\hat{\eta}_i = \frac{\hat{\varepsilon}_i}{\tilde{\sigma}\sqrt{1 - P_i}},$$

where $P_i = P_{ii}$ is the leverage of the ith observation. Note if $\tilde{\sigma} \leftrightarrow \tilde{\sigma}_{-i}$ (thet corrected MLE when excluding (x_i, Y_i). Then $\hat{\eta}_i \sim t_{n-1-p}$. Note: if $\hat{\varepsilon}_i \neq 0$ a.s., $p_i \neq 1$. Also, assume $n \gg p \implies \tilde{\sigma} \approx \sigma$ with low deviation and $\operatorname{var}(\hat{\eta}_i) \approx 1$. Thus, plot $\sqrt{(\hat{\eta}_i)}$ vs $\hat{Y}$ ($|\cdot|$ avoids the 2-sidedness ; $\sqrt{\cdot}$ brings the studentised residuals closer to 1 if $\operatorname{var}(\hat{\eta}_i) \approx 1$). We expect a flat cloud of points around 1. Furthermore, under our model assumptions, $\hat{\eta}_i$ are approximately $\overset{\text{iid}}{\sim} N(0, 1)$. We check this with a Q-Q (Quantile-Quantile) plot.
Coefficients of determination Popular measure of goodness of fit of a model. It compares the residual sum of squares (RSS) of the model vs that of an intercept only model

$$R^2 = \frac{\|Y - \hat{Y}1_n\|^2 - \|(I - P)Y\|^2}{\|Y - \hat{Y}1_n\|^2}, 1_n = (1, 1, \ldots, 1)^T \in \mathbb{R}^n.$$

Note

$$R^2 = \frac{\operatorname{var}_n Y - \hat{\sigma}^2}{\operatorname{var}_n Y} \ \in [0, 1].$$

So R^2 is the proportion of the total variation of the data explained by the model. If we have a larger R^2 value, then we have a better fit, but it always grows whenever we add a new variable. The adjusted

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

Observation A few observations may not agree with our model assumptions i.e. outliers. So, we go back to the data, they could be very informative or we may have to exclude them.
Leverage $\hat{Y}_i = (PY)_i = \sum_{j=1}^n P_{ij}Y_j$ and $\text{var}(\hat{\varepsilon}_i) = \sigma^2(1-p_i), p_i = P_{ii}$ - the leverage of the ith observable . Thus p_i measures the contribution of Y_i to $\hat{Y}_i$. If $\pi \approx 1$, the model is forced to go through Y_i. It is possible that excluding such an observation does not change $\hat{\beta}$ much, but R^2 and F-tests with H_0 : intercept-only model may be highly affected therefore p_i is a measure of influence. Note $\sum_{i=1}^n p_i = \text{tr}(P) = p$ so our rule of thumb is that if $p_i > 3\frac{p}{n}$ then there is a concern that our ith observation is too influential.
Cook's Distance Defined as

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

where $\hat{\beta}_{-i}$ is the MLE when excluding (X_i, Y_i). Note

$$D_i = \frac{1}{p}\frac{P_i}{1-p_i}\hat{\eta}_i^2.$$

Recall that

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

So we get another rule of thumb, the ith observation's influence may be worrying if $D_i > F_{p,n-p}(0.5)$ i.e. removing (X_i, Y_i) pushes the MLE beyond a 50% confidence interval around $\hat{\beta}$.

1.5 ANOVA & ANCOVA

So far, our predictor is $\in \mathbb{R}$ (continuous variables) can deal with categories (or factors) similarly: e.g. let the responses be the weight loss (WL) under J exercise regimes, with the first being no exercise (control group). Our model is Y_{jk} is the weight loss of the kth participant in regime j (n_j of them)

$$Y_{jk} \overset{\text{iid}}{\sim} N(\mu_j, \sigma^2), j = 1, \ldots, J, k = 1, \ldots, n_j.$$

Equivalently, we can say $Y = X\beta + \varepsilon$ with

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

and we have that $\varepsilon \sim N_{\sum_{j=1}^{J} n_j}(0\sigma^2 I)$. This type of model is called one-way analysis of variace (ANOVA). If $n_i = n_j \ \forall \ i \neq j$ then it is called a balanced one-way ANOVA.

Question. Why do we use ANOVA? Statistics describe the variability in a data set. Fisher proposed the model above to reflect that different groups have different variability, and developed associated F-tests for the means. These are in terms of the variances within and between groups, so the model and / or the tests are called ANOVA.

An alternative parametrisation $Y_{jk} = \mu + \alpha_j + \varepsilon_{jk}$ where the $\varepsilon_{jk} \overset{\text{iid}}{\sim} N(0\sigma^2)$, μ is called the baseline effect, α_j is the jth regime's effect in relation to μ. Note: this model is not identifiable, indeed $\mu + c, \alpha - c$ gives the same model $\forall\ c \in \mathbb{R}$. So we need a constraint, generally we take the corner point constraint (default in R) e.g. $\alpha_1 = 0$ so it is simpler to test with respect to baseline effects / control group. We can further subdivide the dataset e.g. I different food diets so Y_{ijk} is WL of kth participant in diet I and regime j (n_{ij} of them). Our model is now

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},\ \varepsilon_{ijk} \overset{\text{iid}}{\sim} N(0, \sigma^2),\ i = 1, \dots, I, j = 1, \dots, J, k = 1, \dots, n_{ij}$$

and the γ_{ij} are the interaction effects. This is called a two-way ANOVA and if we set all γ_{ij} to 0, additive two-way ANOVA. Typically, our (corner-point) constraint is $\alpha_1 = \beta_1 = \gamma_{11} = 0$. we can also include continuous variables e.g. blood pressure, and the resulting normal linear model is called analysis of covariance (ANCOVA).
A word on causal inference and randomised experiments. (non-examinable)
Causal conclusions depend on experiment design. Suppose that given our ANOVA analysis we find that one of our regimes has a positive effect on weight loss. If we allowed the participants to choose their regime it is likely that fitter participants will pick harder routines.

1.6 Model selection

Intuitively, we wish to select "the right model" to focus on the variables of interest. Mathematically, if only β_0 is non-zero in $\beta = \begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix}$, then

$$\text{var}(\check{\beta}_{0,j}) \leq \text{var}(\hat{\beta}_j) \text{ (example sheet) and MSPE}(\check{\beta}_0) = \sigma^2 \frac{P_0}{n} \leq \sigma^2 \frac{P}{n}.$$

We have already met two popular model selection techniques F-tests from section 1.2.3 and $\tilde{R}^2$ (or R^2).
Akaike's information criterion (AIC)
Let $\mathcal{M} = \{f_k(\cdot; \theta_k), \theta_k \in \Theta_k\} k = 1, \dots, K$ be a collection of models. Assume $Y_i \overset{\text{iid}}{\sim} g_i, i = 1, \dots, n$ (g may not be in the collection). For simplicity let $g_i i = g \ \forall\ g$ let $\hat{\theta}_k$ be the MLE for $\mathcal{M}_k$ and θ_k and $\hat{f}_k(\cdot) = f_k(\cdot; \hat{\theta}_k)$ AIC minimises $K(g, \hat{f}_k)$ over k where

$$K(g, f) = \int \log \frac{g}{f} g$$

or equivalently minimise $\mathbb{E}_g \left[\log \hat{f}_k\right]$.

Fact.

$$\frac{1}{n}\sum_{i=1}^{n}\log \hat{f}_k(Y_i)$$

is an estimator of $\mathbb{E}_g\left[\log \hat{f}_k\right]$ with bias $\approx \dim\frac{\theta_k}{n}$. So AIC minimises

$$\mathrm{AIC}(\mathcal{M}_k) = 2n(\frac{1}{n}\sum_{i=1}^{n}\log \hat{f}_k(Y_i) = \frac{\dim\theta_k}{n} = -2(\ell_k\hat{\theta}_k - \dim\theta_k = -2\,(\text{ maximised log-likelihood - no of parame}$$

Note: for the normal linear model $\mathcal{M}_k$

$$\mathrm{AIC}(\mathcal{M}_k) = -2\left(-\log((2\pi\hat{\sigma}_k^2)^{\frac{n}{2}}) - \frac{n}{2} - 2(p_k+1)\right).$$

Note: the models above are general so AIC doesn't require nestedness.

Orthogonality Given p covariates (including intercept), we can form 2^{p-1} models with intercept, and can use e.g. AIC or $\tilde{R}^2$ to select one. This is unfeasible if p is large, unless X has "sufficient orthogonality". Let $X = (X_0, X_1)$ and $\beta = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}$. Then β_0, β_1 are orthogonal sets of parameters if $X_0^T X_1 = 0$. If so, $\hat{\beta} = \begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}$. Indeed,

$$\begin{aligned}
\hat{\beta} &= (X^TX)^{-1}X^TY\\
&= \left(\begin{pmatrix}X_0^T\\ X_1^T\end{pmatrix}(X_0X_1)\right)^{-1}\begin{pmatrix}X_0^T\\ X_1^T\end{pmatrix}Y\\
&= \begin{pmatrix}X_0^T & 0\\ 0 & X_1^TX_1\end{pmatrix}^{-1}\begin{pmatrix}X_0^TY\\ X_1^TY\end{pmatrix}\\
&= \begin{pmatrix}(X_0^TX_0)^{-1}X_0^TY\\ (X_1^TX_1)^{-1}X_1^TY\end{pmatrix} = \begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix}.
\end{aligned}$$

2 or more sets of parameters are mutually orthogonal (MO) if the column spaces of corresponding blocks of X are orthogonal. If so, $\hat{\beta}$ decomposes into the estimators of the blocks, and if there are sufficiently many of these, the strategy above may indeed be feasible. In particular, if all columns of X are orthogonal, and we wish to maximise $\tilde{R}^2$ or R^2 (or minimise the residual sum of squares - RSS) for a fixed p_0; then order $(\hat{\beta}_j\|X_j\|)^2$ (excluding the intercept) increasingly and pick the highest $p_0 - $'1 terms. Indeed, if $S \subseteq \{1, ..., p\}$ and P_s be the orthogonal projection onto the corresponding columns of X,

$$\begin{aligned}
\|(I-P_s)Y\|^2 &= \|Y - \sum_{j\in S}\hat{\beta}_jXj\|^2\\
&= Y^TY - 2\sum\hat{\beta}_jX_j^TY + \sum_{i,j}\hat{\beta}_i\hat{\beta}_jX_i^TX_j\\
&= \|Y\|^2 - \sum\hat{\beta}_j\|X_j\|^2
\end{aligned}$$

.

Exact orthogonality is uncommon, unless we designed X or we transformed it (at the risk of losing interpretability). Orthogonality between intercept and the rest is common, by the common transformation of mean-centring the columns of X.
Forward (fwd) and backward (bwd) selection
If no or little orthogonality, popular strategies are as follows
Forward selection

(i) Fit the intercept only model S_0

(ii) Compute all models with one more parameter and keep the model with the lowest RSS

(iii) Repeat 2 until all (or sufficiently many) covariates are included

(iv) Choose one model from the resulting sequence $S_0 \subset S_1 \subset \cdots$ using e.g. AIC or $\tilde{R}^2$.

Backward selection
The same as above, bur start with all covariates until reaching the intercept-only model (removing one covariate at a time)

1.7 Inference after model selection

Inference from 1.2.3 assumes that the model was fixed prior to data-collection. Therefore, we cannot use the same data for selection and inference. The easy option is to split the data and use each bit for each purpose

2 Exponential families and generalised linear models

2.1 Non-normal responses

Responses not always naturally live in $\mathbb{R}$: e.g. prices (> 0), counting data ($\mathbb{N}$), binary options (yes & no). Thus, the normal linear model may not be appropriate.

Variable transformations

A first approach is to transform the data so that our linear model assumptions are met approximately. If responses > 0 (including $\mathbb{N}$), the Box-Cox transformation is classical:

$$y \mapsto y^{\lambda} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases}.$$

One finds λ by MLE regarding $(\lambda, \beta, \sigma^2)$ as the parameter. Drawback, we should at once, achieve approximate normality, variance stabilisation and linearity in β. A more recent and superior strategy is to model these separately by e.g. GLM.

GLMs (preliminaries)

We can think of normal linear models (NLM) as consisting of 3 components

(i) Random component: $Y_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2),\ i = 1, \ldots, n.$

(ii) Systematic component: a linear predictor $\eta = (\eta_1, \ldots, \eta_2), \eta_i = x_i^T \beta_i.$

(iii) Link between (i) & (ii) $\eta_i = g(\mu_i), g = id.$

GLMs generalise NLM in (i) & (iii), allowing more general distributions (including the prototypical Poisson & binomial) and links, whilst keeping its conceptual and computational simplicity. To generalise (i), we need to introduce exponential dispersion families (EDFs), which in turn are built from exponential families (EFs)

2.2 Exponential Families

Let $f_0 : \mathcal{Y} \to [0, \infty), \mathcal{Y} \subseteq \mathbb{R}$ be a density or probability mass function which we refer to as model function. For any non-degenerate f_0, i.e. if $Y \sim f_0$, $\operatorname{var} Y > 0$ we can form a family of model functions by exponential tilting.

$$f(y; \theta) = \frac{e^{y\theta} f_0(y)}{\int_{\mathcal{Y}} e^{y'\theta} f_0(y') \mathrm{d}y'}, y \in \mathcal{Y}, \theta \in \{\theta \in \mathbb{R} : \int_{\mathcal{Y}} e^{y'\theta} f_0(y') \mathrm{d}y' < \infty\}.$$

Note:

$$\int_{\mathcal{Y}} e^{y'\theta} f_0(y') \mathrm{d}y' = \mathbb{E}_{f_0}\left[e^{\theta \mathcal{Y}}\right],$$

i.e. the moment generating function.

Moment & cumulant generating functions (m.g.f. & c.d.f.)

The m.g.f. & c.d.f. of a random variable Y or of its model function is $M(t) = \mathbb{E}\left[e^{tY}\right]$ & $K(t) = \log M(t)$ whenever they are finite. Claim: $\{t \in \mathbb{R} : M(t) < \infty\}$ is an interval containing the origin. Indeed, if $M(t), M(t') < \infty$. Then $\forall\ \rho \in [0, 1], y \in \mathcal{Y}$

$$e^{\left((1-\rho)t + \rho t'\right)y} \leq (1-\rho)e^{ty} + \rho e^{t'y} < \infty.$$

If $M(t)$ is defined in a neighbourhood about the origin e.g. if $\exists\ t < 0 < t'$ with $M(t), M(t') < \infty$, they have series expansions

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

κ_r is the rth cumulant of Y.

Fact.

$$\mathbb{E}\left[Y^r\right] = \frac{\mathrm{d}^r}{\mathrm{d}t} M(t)\Big|_{t=0} = M^r(0) \& \kappa_r = K^r(0).$$

Check $\kappa_1 = \mathbb{E}\left[Y\right], \kappa_2 = \operatorname{var} Y$

Definition. Let f_0 be non-degenerate with c.g.f K and define $\Theta = \{\theta \in \mathbb{R} : K(\theta) < \infty\}$. Then we say that the exponentially tilted family of model functions $\{f(\cdot;\theta), \theta \in \Theta\}$ is the natural exponential family of f_0, θ is the natural parameter and Θ is the natural parameter space.

Note: $f(y,\theta) = e^{y\theta - K(\theta)} f_0(y)\ \forall\ y \in \mathcal{Y}, \theta \in \Theta$ and if $\theta, t+\theta \in \Theta$ the m.g.f of $f(\cdot,\theta)$ is

$$M(t,\theta) = \int_{\mathcal{Y}} e^{ty} e^{y\theta - K(\theta)} f_0(y) \mathrm{d}y = e^{K(t+\theta) - K(\theta)} \int_{\mathcal{Y}} f(y, t+\theta) \mathrm{d}y$$

and its c.g.f is $K(t,\theta) = \log M(t,\theta)$. Then if $Y \sim f(\cdot;\theta)$ for $\theta \in \int(\theta)$, we can characterise its mean and variance as

$$\mathbb{E}_\theta\left[Y\right] = \frac{\mathrm{d}}{\mathrm{d}t} K(t,\theta)\Big|_{t=0}, \operatorname{var}_\theta Y = \frac{\mathrm{d}^2}{\mathrm{d}t^2} K(t,\theta)\Big|_{t=0}.$$

Fact. If f_0 is non-degenerate, then $f(\cdot,\theta)$ is so $\ \forall\ \theta \in \Theta$. Then,

$$\forall\ \theta \in \mathrm{int}\Theta, \mu(\theta) = \mathbb{E}_\theta\left[Y\right] = K'(0),$$

is the mean function and satisfies $\mu'(\theta) = K''(\theta) = \operatorname{var}_\theta Y > 0$, and can reparametrise $\{f(\cdot,\theta\}$ through their means: Let $\mathcal{M} = \{\mu(\theta) : \theta \in \mathrm{int}\Theta\}$ be the mean space, and $\theta(\mu)$ be the inverse of $\mu(\theta)$, then the mean value parametrisation is

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

If $\mathrm{int}\theta \neq \theta$ we can extend the definition by continuity, possibly setting $y = \pm\infty$. The function $V : \mathcal{M} \to (0,\infty)$, $\mu \mapsto \operatorname{var}_\theta(Y) = K''(\theta(\mu))$ is called the variance function.

Example. (i) Let f_0 be the standard normal density. Check: $M(t) = e^{\frac{1}{2}\theta^2}, \theta \in \mathbb{R}$, so $K(\theta) = \frac{1}{2}\theta^2$ and

$$f(y;\theta) = e^{y\theta - \frac{1}{2}\theta^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-\theta)^2}.$$

. This is the $N(\theta,1), \theta \in \mathbb{R}$ family, so $\mu(\theta) = \theta, \mathcal{M} = \mathbb{R}, \theta(\mu) = \mu$ and $V(\mu) = 1$ which can be checked by differentiating K.

(ii) Let f_0 be the pmf of Poi(1) i.e. $f_0(y) = e^{-1}\frac{1}{y!}, y \in \mathbb{N} \cup \{0\}$. Check that $M(\theta) = \exp(e^\theta - 1), \Theta = \mathbb{R}$, so $K(\theta) = e^\theta - 1$ and

$$f(y,\theta) = e^{y\theta-(e^\theta-1)}e^{-1}\frac{1}{y!} = e^{-e^\theta}\frac{\left(e^\theta\right)^y}{y!}, y \in \mathbb{N} \cup \{0\}, \theta \in \mathbb{R}.$$

This is the $\{\text{Poi}(e^\theta) : \theta \in \mathbb{R}\}, \mu(\theta) = e^\theta \mathcal{M} = \mathbb{R}^+, \theta(\mu) = \log\mu$ and $V(\mu) = \mu$ all of which can be checked differentiating K.

2.3 EDFs

In the first example, exponentially tilting f_0 generalised it through the mean only. In the second example, it also generalised the variance but only because the mean variance coincide therein. We wish to generalise the variance too. Let K be the c.g.f. of some non-degenerate model function, and

$$\Phi = \{\sigma^2 > 0 : \frac{K(\cdot)}{\sigma^2} \text{ is the c.g.f. of some } f_{\sigma^2}\}.$$

By exponentially tilting any f_{σ^2} we obtain the following family of model functions

$$e^{x\theta-K(\theta)/\sigma^2}f_{\sigma^2}(x) \overset{y=x\sigma^2}{=} e^{\frac{1}{\sigma^2}(y\theta-K(\theta)}f_{\sigma^2}\left(\frac{y}{\sigma^2}\right), y \in \mathcal{Y}, \theta \in \Theta = \{\theta \in \mathbb{R}, K(\theta) < \infty\}, \sigma^2 \in \Phi.$$

Definition (Exponential dispersion family)**.** An *exponential dispersion family* is a family of non-degenerate model functions

$$f(y;\theta,\sigma^2) = a(\sigma^2,y)\exp\left(\frac{1}{\sigma^2}\{\theta y - K(\theta)\}\right), y \in \mathcal{Y},$$

where a (positive) and K (a c.g.f of a non-degenerate σ-finite measure) are known, $\sigma^2 \in \Phi \subseteq (0,\infty)$ is the dispersion parameter, θ is the natural parameter, and Θ is an open interval.

On the example sheet, we will show that if $\theta, \theta + \sigma^2 t \in \Theta$, then the c.g.f of $f(\cdot,\theta,\sigma^2)$ is

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

Since Θ is open $\forall\, \theta \in \Theta, \sigma^2 \in \Phi, \theta + \sigma^2 t \in \Theta$ for any t in some neighbourhood of the origin and $\mathbb{E}[Y] = \mathbb{E}_{\theta,\sigma^2}[Y] = K'(0,\theta,\sigma^2) = K'(\theta)$ and $\text{var}\, Y = \text{var}_{\theta,\sigma^2} Y = K''(o,\theta,\sigma^2) = \sigma^2 K''(\theta)$. Let $\mu(\theta) = K'(\theta)$ be the mean function, $\mathcal{M} = \{\mu(\theta) : \theta \in \Theta\}$ the mean space, $\theta(\mu)$ the inverse function of $\mu(\theta)$ (note $K''(\theta) > 0$ by non-degeneracy and the variance function is

$$V : \mathcal{M} \to \mathbb{R}^+, \mu \mapsto K''(\theta(\mu))$$

(despite $\text{var}\, Y = \sigma^2 K''(\theta)$.

Example. (i) Let f_0 be the standard normal density, so $K(\theta)/\sigma^2 = \frac{1}{2\sigma^2}\theta^2$ is the cgf of $N(0,\frac{1}{\sigma^2})$ for any $\theta \in \Theta = \mathbb{R}, \sigma^2 \in \Phi = (0,\infty)$.

$$f(x,\theta,\sigma^2) = e^{x\theta-\theta^2/(2\sigma^2)}\sqrt{\frac{\sigma^2}{2\pi}}\exp\left(-\frac{\sigma^2}{2}x^2\right) = a(\sigma^2,y)\exp\left(\frac{1}{\sigma^2}(y\theta - \frac{\theta^2}{2})\right)..$$

So $\mu(\theta) = \theta, \mathcal{M} = \mathbb{R}, \theta(\mu) = \mu, V(\mu) = 1$. Note: we generated $\{N(\theta,\sigma^2) : \theta \in \mathbb{R}, \sigma^2 > 0\}$. Traditionally, one checks that a given family is an EDF (rather than constructing them)

(ii) $\mathrm{Poi}(\lambda), \lambda > 0$ is EDF (simple to check), and it cannot be generalised further with the use of dispersion.

(iii) Let $Z \sim \mathrm{Bin}(n,p), n \in \mathbb{N}, p \in (0,1)$. Then $Y = \frac{Z}{n} \sim \frac{1}{n}\mathrm{Bin}(n,p)$ and

$$f(y,n,p) = \mathbb{P}_{n,p}(Y=y) = \mathbb{P}_{n,p}(Z=ny) = \binom{n}{ny} p^{ny}(1-p)^{n-ny} = \binom{n}{ny} \exp\left(ny \log \frac{p}{1-p} + n\log(1 -$$

So $\theta = \log \frac{p}{1-p}, \Theta = \mathbb{R}, \sigma^2 = \frac{1}{n}, \Phi = 1/\mathbb{N}$,

$$K(\theta) = -\log(1-p(\theta)) = -\log(1 - \frac{e^\theta}{1+e^\theta}) = \log(1+e^\theta).$$

$\mu(\theta) = K'(\theta) = \frac{e^\theta}{1+e^\theta}, \mathcal{M} = (0,1), \theta(\mu) = \log\frac{\mu}{1-\mu}, V(\mu) = K''(\theta(\mu)) = \mu(1-\mu)$.

(iv) Let $Y \sim \Gamma(\alpha,\beta)$ then

$$f(y,\alpha,\beta) = \frac{\beta^\alpha y^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)}, y > 0.$$

It is not immediately obvious to see that this is an EDF. Note that $y^{\alpha-1} = \exp((\alpha-1)\log y)$, so it will absolve it into a, so we could first guess that $\sigma^2 = \sigma^2(\alpha)$. Note $\mathbb{E}[Y] = \alpha/\beta, \operatorname{var} Y = \alpha/\beta^2$ so $\mu = \alpha/\beta, \sigma^2 V(\mu) = \alpha/\beta^2$. In particular, when $\alpha = \beta, \mu = 1, \sigma^2 V(1) = \frac{1}{\beta}$ so $\sigma^2 = C/\alpha$ for some constant $C > 0$

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

.

So can take $c = 1$ and this is an EDF with $\theta = -\frac{\beta}{\alpha}, \Theta = \mathbb{R}^-, \sigma^2 = \frac{1}{\alpha}, \Phi = \mathbb{R}^+, K(\theta) = \log\left(-\theta^{-1}\right), \mu(\theta) = -\theta^{-1}, \mathcal{M} = \mathbb{R}^+, \theta(\mu) = -\mu^{-1}$ and $V(\mu) = \theta(\mu)^{-2} = \mu^2$.

2.4 GLMs

Definition. A GLM for data $(Y_1, x_1^T), \ldots, (Y_n, x_n^T)$ is defined by the following properties

(i) $Y_1, \ldots, Y_n$ are independent (given the x_i) with model functions in the same EDF

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

(i.e. $a\&K$ are fixed in i) where $\sigma_i^2 = \sigma^2 a_i$ with $a_1, \ldots, a_n > 0$ known and $\sigma^2 > 0$ possibly unknown.

(ii) The mean of Y_i (μ_i) and the ith component of the linear predictor $\eta_i := x_i^T\beta$ are linked by the equation

$$g(\mu_i) = \eta_i \quad i = 1, \ldots, n,$$

where the *link function* $g : \mathcal{M} \to \mathbb{R}$ is strictly monotone (for identifiability), generally increasing, and twice differentiable (for computations and inference).

2.4.1 Choice of the link function

The link function g is generally chosen so that it is relatively easy to compute and interpret the estimated coefficients of β. Given g, the β's for which the GLM is defined are those satisfying

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

If $g(\mathcal{M}) \subsetneq \mathbb{R}$ it may be complicated to compete the estimator of β. So, typically, one chooses g surjective i.e. $g(\mathcal{M}) = \mathbb{R}$.

Example. In our binomial example $\mathcal{M} = (0, 1)$ and one can take the Probit function, i.e. the inverse of the standard normal probability distribution function, or the logit function, $g(\mu) = \log\{\mu/(1-\mu)\} (= \theta(\mu))$.

The general choice $g(\mu) = \theta(\mu)$ is called the canonical link function e.g. $g = id, \log$, and logit respectively in the examples above. As seen next, it renders the log-likelihood concave and the likelihood equations simple, so it allows for the construction of efficient algorithms to perform inference.

2.4.2 Likelihood equations

A typically optimal estimator of β is the MLE. The log-likelihood of a GLM is

$$\begin{aligned}\ell(\beta, \sigma^2) &= \log\left(\prod_{i=1}^n a(\sigma^2 a_i, Y_i) \exp\left(\frac{1}{\sigma^2 a_i}\left(\theta\left(g^{-1}\left(X_i^T\beta\right)\right)Y_i - K\left(\theta\left(g^{-1}\left(X_i^T\beta\right)\right)\right)\right)\right)\right) \\ &= \sum_{n=1}^i \frac{1}{\sigma^2 a_i}\left(\theta\left(g^{-1}\left(X_i^T\beta\right)\right)Y_i - K\left(\theta\left(g^{-1}\left(x_i^T\beta\right)\right)\right)\right) + \sum_{i=1}^n \log a\left(\sigma^2 a_i, Y_i\right).\end{aligned}$$

$$\beta \in \{\beta' \in \mathbb{R}^p : g^{-1}\left(X_i^T\beta'\right) \in \mathcal{M} \ \forall \ i = 1, \ldots, n\}, \sigma^2 a_i \in \Phi.$$

when $g(\mu) = \theta(\mu)$, it simplifies due to $\theta\left(g^{-1}\left(X_i^T\beta\right)\right) = X_i^T\beta$ and

$$\beta \in \{\beta' \in \mathbb{R}^p : x_i^T\beta \in \Theta \ \forall \ i = 1, \ldots, n\}.$$

Furthermore, it renders the log-likelihood concave in β for σ^2 fixed. Indeed,

$$\frac{\partial}{\partial\beta_j}\ell(\beta, \sigma^2) = \frac{\partial}{\partial\beta_j}\left(\sum_{i=1}^n \frac{1}{\sigma^2 a_i}\left(X_i^T\beta Y_i - K(x_i^T\beta)\right)\right) = \sum_{i=1}^n \frac{x_i}{\sigma^2 a_i}\left(y_i - K'(x_i^T\beta)\right),$$

and

$$\frac{\partial^2}{\partial\beta_j\partial\beta_k}\ell(\beta, \sigma^2) = -\sum_{i=1}^n \frac{x_{ij}x_{ik}}{\sigma^2 a_i}K''(x_i^T\beta) \quad j, k = 1, \ldots, p.$$

So the Hessian satisfies $H\ell(\beta,\sigma^2) = -\sum_{i=1}^n \frac{x_i^T x_i^T}{\sigma^2 a_i} K''(x_i\beta)$. Recall that $\sigma^2, a_i, K''(x_i^T\beta) > 0$ then $H\ell$ is negative semi-definite (and if $X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix}$ is full-rank it is negative definite), so ℓ is concave (strictly concave) in β for σ^2 fixed. In particular, if $\overline{\nabla}_\beta \ell(\beta,\sigma^2)\Big|_{\hat\beta} = 0$ with $\hat\beta \in \{\beta \in \mathbb{R}^p : x_i^T\beta \in \Theta \ \forall\ i = 1,\dots,n\}$. Then $\hat\beta$ is a MLE and it is easy to characterise through the likelihood equations $\equiv \sum_{i=1}^n \frac{x_i}{\sigma^2 a_i}(Y_i - K'(x_i^T\beta) = 0$.

2.5 Inference

Consider data $(Y_1, x_1^T), \dots, (Y_n, x_n^T)$ with Y_i independent (given the x_i) and (joint) model function for $Y = (Y_1, \dots, Y_n)^T$ in

$$\{\prod_{i=1}^n f_{x_i}(y_i;\theta), y_i \in \mathcal{Y}, \theta \in \Theta \subseteq \mathbb{R}\}$$

(more general than GLM). We review some theory for MLEs.

2.5.1 The score function

Note: $\ell(\theta;Y) = \ell(\theta) = \sum_{i=1}^n \log f_{X)=i}(Y_i,\theta)$. Then the score function is $U(\theta;Y) = U(\theta) = \nabla_\theta \ell(\theta)$. Let $\hat\theta = \hat\theta^{(n)}$ be the MLE of $\theta \in \text{int}\Theta$, throughout we will assume the so-called regularity assumptions / conditions ("RCs", see principles of statistics for iid version) which, in particular guarantee that $\hat\theta$ exists (in intΘ with high probability) and is unique. Then $U(\theta) = 0$ at $\theta = \hat\theta$ and only there. More is true under "RCs" (which ensure differentiation and integration can be interchanged.

(i) $\mathbb{E}_\theta[U(\theta;Y)] = 0$.

(ii) $\text{var}_\theta U(\theta;Y) = -\mathbb{E}_\theta[H_\theta \ell(\theta;Y)]$

Indeed,

$$\begin{aligned}
\mathbb{E}_\theta[U(\theta;Y)] &= \int_{\mathcal{Y}^n} \nabla_\theta \log f(y,\theta) f(y,\theta) \mathrm{d}y, \quad f(y,\theta) = \prod_{i=1}^m f_{X_i}(y_i;\theta) \\
&= \int_{\mathcal{Y}^n} \nabla_\theta f(y,\theta) \mathrm{d}y \\
&= \nabla_\theta \int_{\mathcal{Y}^n} f(y,\theta)\mathrm{d}y = \nabla_\theta 1 = 0,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}_\theta[H_\theta \ell(\theta,Y)] &= \mathbb{E}_\theta\left[\frac{H_\theta f(\theta;Y)}{f(\theta:Y)} - \left(\frac{\nabla_\theta f(\theta;Y)}{f(\theta;Y)}\right)\left(\frac{\nabla_\theta f(\theta;Y)}{f(\theta;Y)}\right)^T\right] \\
&= \int_{\mathcal{Y}^n} H_\theta f(\theta;y)\mathrm{d}y - \text{var}_\theta(U(\theta;Y)) \\
&= -\text{var}_\theta(U(\theta;Y)).
\end{aligned}$$

2.5.2 Fisher information

The Fisher information (matrix) is $i^{(n)}(\theta) = i(\theta) = \mathrm{var}_\theta(U(\theta, Y))$. We interpret this quantity as a measure of how easy it is to estimate θ (as the "true parameter") with "the higher, the easier". The observed information matrix is $j(\theta) = -H_\theta \ell(\theta; Y)$ (so $i(\theta) = \mathbb{E}_\theta[j(\theta)]$.

Example. The normal linear model (NLM) $Y = X\beta + \varepsilon, \varepsilon \sim N_p(0, \sigma^2 I_p)$. On the example sheet you will show that

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

Note that $\mathrm{var}_\theta \hat{\beta} = i^{-1}(\beta)$.

Theorem (Cramer-Rao lower bound)**.** Let $\tilde{\theta}$ be an unbiased estimator of $\theta \in \mathrm{int}\Theta$. Then under "RCs" $\mathrm{var}_\theta \tilde{\theta} - i^{-1}(\theta)$ is positive semi-definite.

Thus, since $\hat{\beta}$ is unbiased and the NLM satisfies the "RCs" $\hat{\beta}$ has minimal variance among all unbiased estimators (and not only linear as in the Gauss-Markov theorem) "True" for MLEs.

2.5.3 Asymptotic normality of MLE and Wilks' theorem

Recall: a sequence of random vectors $Z_m \in \mathbb{R}^k$ converges in distribution to a random vector $Z \in \mathbb{R}^k$ if $\mathbb{P}(Z_m \in B) \underset{m\to\infty}{\to} \mathbb{P}(Z \in B)$ for all B (Borel) set in $\mathbb{R}^k$ such that $\mathbb{P}(Z \in \partial B) = 0$ where $\partial B = \overline{B} \setminus \mathrm{int}B$. We write $Z_m \xrightarrow{d} Z$

Theorem. Assume that $i^{(n)}(\theta)/n \underset{n\to\infty}{\to} I(\theta)$, where $I(\theta)$ is a positive definite matrix. Then, under "RCs",

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

Equivalently,

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

Informally, write $\hat{\theta} \sim AN_d(\theta, i^{-1}(\theta))$ (informal because $i(\theta) = i^n(\theta)$ so it cannot be a limit)

Note: θ is unknown, so cannot use the theorem above immediately. If $i^{-1}(\theta)$ is continuous then $i^{-1}(\hat{\theta})$ is a good estimate and $\hat{\theta} \sim AN_d(\theta, i^{-1}(\hat{\theta})$. Inference

$$C_j = \{\theta_j' \in \mathbb{R} : |\hat{\theta}_j - \theta_j'| \le z_{\alpha/2}\sqrt{(i^{-1}(\hat{\theta})_{jj}}\},$$

where $\mathbb{P}(Z \ge z_\alpha) = \alpha$ if $Z \sim N(0,1)$, is a $(1-\alpha)$-level asymptotic confidence interval for θ_j. Also,

$$C = \{\theta' \in \mathbb{R}^d : (\hat{\theta} - \theta')^T i(\theta')(\hat{\theta} - \theta') \le \chi^2_d(\alpha)\}$$

where $\mathbb{P}\left(X \ge \chi^2_d(\alpha)\right) = \alpha$ if $X \sim \chi^2_d$, is a $(1-\alpha)$-level asymptotic confidence set for θ. We can test $H_0 : \theta_j = \theta_{j,0}$ v.s. $H_1 : \theta_j \ne \theta_{j,0}$ with $\phi = 1\{\theta_{j,0} \notin C_j\}$ and $H_0 : \theta = \theta_0$ v.s. $H_1 \theta \ne \theta_0$ with $\phi = 1\{\theta_0 \in C\}$ at level $1-\alpha$.

How to test sub-models? i.e. $H_0 : \theta \in \Theta_0$ v.s. $H_1 : \theta \notin \Theta_0, \Theta_0 \subseteq \Theta$ with $\dim\Theta_0 < \dim\Theta$. Let $\Theta = \mathbb{R}^d, \Theta_0 = \mathbb{R}^{d_0}, d_0 < d$ wlog $\theta = (\theta_0^T, \theta_1^T)^T, \theta_0 \in \Theta_0$. Can test $H_0\theta_1 = 0$ v.s. $H_1 : \theta_1 \neq 0$ with $\phi_{\text{LR}} = 1\{\omega_{\text{LR}} \geq \chi^2_{d-d_0}(\alpha)\}$,

$$\omega_{\text{LR}} = 2\log\frac{\sup_{\theta'\in\Theta} L(\theta')}{\sup_{\theta'\in\Theta_0} L(\theta')}.$$

Theorem (Wilks' Theorem)**.** Assume H_0 is true and "RCs". Then

$$\omega_{\text{LR}}(H_0) \xrightarrow{d} \chi^2_{d-d_0}.$$

Remark. (i) Thus, ϕ_{LR} is an asymptotic α-significance level test i.e.

$$\mathbb{E}_{H_0}[\phi_{\text{LR}}] = \mathbb{P}_{H_0}\left(\omega_{\text{LR}} \geq \chi^2_{d-d_0}(\alpha)\right) \to \alpha \text{ as } n \to \infty.$$

Recall: $C_j \& C$ are asymptotic $(1-\alpha)$-level confidence sets for θ_j, θ ie. $\mathbb{P}_{\theta_j}(\theta_j \in C_j), \mathbb{P}_\theta(\theta \in C) \to 1-\alpha$ as $n \to \infty$. Hence, these are asymptotic (as $n \to \infty$), but in practice, we only have finite data. For our purposes, "$n = O(100)$" will be okay.

(ii) The previous theorems are consistent with the NLM since $t_{n-p} \xrightarrow{d} N(0,1)$ and $F_{q,n-p} \xrightarrow{d} \chi^2_q \ \forall\ p, q \in \mathbb{N}$.

(iii) Wilks' Theorem can be used to test $H_0 : \theta_j = 0$ v.s. $H_1 : \theta_j \neq 0$. Unlike in NLM, this test is different (and preferable) to ϕ_j in general.

2.5.4 Inference in GLMs

Let $i(\beta, \sigma^2)$ be the Fisher information in a GLM, on the example sheet, we show

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

where $i_\beta \in \mathbb{R}^{p\times p}, i_{\sigma^2} \in (0,\infty)$. Recall that $\text{var}(Y_i) = \mathbb{E}\left[(Y_i - \mu_i)^2\right] = \sigma^2 a_i V(\mu_i)$. We take

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

when σ^2 unknown, and $\tilde{\sigma^2}$ when known. Then $\hat{\beta} \sim AN_p(\beta, i^{-1}(\hat{\beta}, \tilde{\sigma^2})$ i.e.

$$i_{\hat{\beta}}^{\frac{1}{2}}(\hat{\beta}, \tilde{\sigma^2})(\hat{\beta} - \beta) \xrightarrow{d} N_p(0, I),$$

and

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

is an asymptotic $(1-\alpha)$-level confidence interval for β_j. For n moderately large and σ^2 unknown , better to replace $z_{\frac{\alpha}{2}}$ by $t_{n-p}(\frac{\alpha}{2})$. Asymptotic tests for β_j follow immediately. How to test sub-models?
Wlog, $\beta = (\beta_0^T, \beta_1^T)^T, \beta_0 \in \mathbb{R}^{p_0}, p_0 < p$. We wish to test $H_0 : \beta_1 = 0$ v.s.

$H_1 : \beta_1 \neq 0$. Write $\hat{\beta}, \check{\beta} \in \mathbb{R}^p$ for the MLEs for β in the full and sub-models, and define $\hat{\mu}, \check{\mu}$ by $\hat{\mu}_i = g^{-1}(X_i^T \hat{\beta}), \check{\mu} = g^{-1}(X_i^T \check{\beta})$. Let

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

and note $\ell(\hat{\beta}, \sigma^2) = \tilde{\ell}(\hat{\mu}, \sigma^2), \ell(\check{\beta}, \sigma^2) = \tilde{\ell}(\check{\mu}, \sigma^2)$ and $\max_{\mu \in \mathbb{R}^n} \tilde{\ell}(\mu, \sigma^2) = \tilde{\ell}(y, \sigma^2)$ i.e. the RHS are the maximised log-likelihoods for the full and sub-model, and for the so-called saturated model, which imposes no restrictions on the μ_i. Let $D(y; \mu) 2\sigma^2 [\tilde{\ell}(y, \sigma^2) - \tilde{\ell}(\mu, \sigma^2)$ (it does not depend on σ^2). We define the deviances of the full and sub-models by $D(y; \hat{\mu}$ and $D(y; \check{\mu})$, respectively. On the example sheet, we see these generalise the RSS from the NLM. Note if σ^2 is known, they are the (rescaled) LR test statistics to test the full and sub-models v.s. saturated model, respectively. Furthermore,

$$\omega_{\mathrm{LR}}(H_0) = \frac{1}{\sigma^2} [D(y; \hat{\mu}) - D(y; \check{\mu})],$$

so by Wilks' Theorem, $\omega_{\mathrm{LR}} \xrightarrow{d} \chi^2_{p-p_0}$. For n moderately large and σ^2 unknown, it is better to replace critical values by those of $F_{p-p_0, n-p}$ distribution.

2.6 Computation

Let $U(\beta) = \nabla_\beta \ell(\beta, \sigma^2), j(\beta) = -H_\beta \ell(\beta, \sigma^2), i(\beta) = \mathbb{E}_\beta [j(\beta)]$. Under "RCs", we wish to find the $\hat{\beta}$ satisfying $U(\hat{\beta}) = 0$. By Taylor's Theorem, if $\beta \approx \beta_0$,

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

so if $\beta_0 \approx \hat{\beta}$, $U(\beta_0) \approx j(\beta_0)(\hat{\beta} - \beta_0) \implies \hat{\beta} \approx \beta_0 + j^{-1}(\beta_0) U(\beta_0)$, assuming that j is invertible.

Newton-Raphson

(i) Initial guess $\hat{\beta}^{(0)} \in \mathbb{R}^p$

(ii) In the mth iteration, $m \in \mathbb{N}$, set $\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + j^{-1}(\hat{\beta}^{(m-1)}) U(\hat{\beta}^{(m-1)})$

(iii) Repeat (ii) until "numerical convergence" e.g. until the log-likelihood increases less than a prescribed tolerance

Issue: $j(\hat{\beta}^{-1})$ may be singular or close to it. Recall that $i(\beta)$ is positive definite under "RCs".

Fisher scoring

The same as N-R, replacing $j^{-1}(\hat{\beta}^{(m-1)})$ by $i^{-1}(\hat{\beta}^{(m-1)})$. This is not guaranteed to converge to $\hat{\beta}$ but it generally does. We have an alternative formulation:

Iterative reweighted least squares On the example sheet, we show

$$U_j(\beta) = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{\sigma^2 a_i \sqrt{\mu_i} [g'(\mu_i)]^2}, \; i_{jk} = \sum_{i=1}^{n} \frac{X_{ij} X_{ik}}{\sigma^2 a_i \sqrt{\mu_i} g'(\mu_i)}.$$

Let $W(\mu) \in \mathbb{R}^{n \times n}$ diagonal with $W_{ii}(\mu) = (a_i \sqrt{\mu_i} [g'(\mu + i)]^2)^{-1}$ and $R(\mu) \in \mathbb{R}^n$ such that $R_i(\mu) = g'(\mu_i)(y_i - \mu_i)$. Then,

$$U(\beta) = \frac{1}{\sigma^2} X^T W(\mu) R(\mu), \; i(\beta) = \frac{1}{\sigma^2} X^T W(\mu) X.$$

So, in the mth iteration FS sets

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

Where $\hat{\mu}^{(m-1)} \in \mathbb{R}^n$ is defined by $\hat{\mu}_i^{(m-1)} = g^{-1}(X_i^T \hat{\beta}^{(m-1)}$. Let $\hat{\eta}^{(m)} := X\hat{\beta}^{(m)}$ and define the adjusted dependent variable $Z^{(m)}$ by

$$Z^{(m)} = \hat{\eta}^{(m)} + R(\hat{\mu}^{(m)}).$$

Then,

$$\hat{\beta}^{(m)} = (X^T W(\hat{\mu}^{(m-1)}X)^{-1} X^T W(\hat{\mu}^{(m-1)}) Z^{(m-1)} = \underset{b\in\mathbb{R}^p}{\operatorname{argmin}} \{\sum_{i=1}^{n} W_{ii}(\hat{\mu}^{(m-1)})(Z_i^{(m-1)} - X_i^T b)^2\}.$$

Note: $g(y_i) \approx g(\mu_i) + g'(\mu_i)(y_i - \mu_i)$ if $y_i \approx \mu_i$, so $Z_i^{(m)} \approx g(y_i)$ if $\hat{\mu}_i^{(m)} \approx y_i$. Hence, FS applies IRLS to the sequence of approximations $Z^{(0)}, Z^{(1)}, \ldots$ incorporating the (expected) local curvature of the log-likelihood of the GLM. We may take $\hat{\beta}^{(0)} = 0$ or $\hat{\mu}^{(0)} = y$ in which case $Z^{(0)} = g(y) = (g(y_1), \ldots, g(y_n))^T$.

2.7 Model checking

Residuals are key again, but several notions of residuals:

- Raw residuals - $Y_i - \hat{\mu}_i$ have different variances so they are undesirable.
- Pearson's residuals -

$$e_i := \frac{Y_i - \hat{\mu}_i}{\sqrt{\tilde{\sigma^2}} a_i \sqrt{\hat{\mu}_i}},$$

  if $\hat{\mu}_i \approx \mu_i$ then $\mathbb{E}[e_i] \approx 0, \operatorname{var} e_i \approx 1$. In fact, for some GLMs, $e_i \overset{d}{\approx}$ normal dist, so we can use $Q-Q$ plots to check the models. Two important examples :

  (i) $Y_i \overset{\text{ind}}{\sim} \frac{1}{n_i}\text{Bin}(n_i, p_i)$, n_i large $\forall\ i$.

  (ii) $Y_i \overset{\text{ind}}{\sim} \text{Poi}(\mu_i), \mu_i$ large $\forall\ i$.

  Indeed, for n fixed and $n_i, \mu_i \to \infty$ then $e_i \overset{d}{\to}$ normal distribution, and $\sum_{i=1}^n e_i^2 \overset{d}{\to} \chi^2_{n-p}$. The latter is Pearson's Chi-squared statistic, and can be used to test these models (goodness of fit tests). These are called small dispersion asymptotics (SDA) and arise from having growing information for each observation rather than information growing from increasingly many observations ($n \to \infty$). Without asymptotics, $\operatorname{var} Y_i - \hat{\mu}_i \approx \sigma^2 a_i V(\mu_i)(1 - h_i(\mu)), h_i(\mu) = H_{ii}(\mu)$ where $H(\mu) \in \mathbb{R}^{n\times n}$ diagonal with

$$H_{ii}(\mu) = W^{\frac{1}{2}}(\mu) X (X^T W(\mu) X)^{-1} X^T W(\mu).$$

  Let $\hat{h}_i = H_{ii}(\hat{\mu})$ (leverage of the ith observation) and

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

(standardised Pearson's residuals). It still holds that $\hat{h}_i \in [0,1]$ and $\sum_{i=1}^n \hat{h}_i = p_i$ but observations with extreme covariates need not have high leverage. Measure of influence : Cook's distance

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

– Deviance residuals -

$$d_i := \text{sgn}(y_i - \hat{\mu}_i)\sqrt{D_i(y_i;\hat{\mu}_i)}, D_i(y_i;\hat{\mu}_i) = \frac{2}{a_i}\left[y_i\{\theta(y_i) - \theta(\hat{\mu}_i)\} - \{K(\theta(y_i)) - K(\theta(\hat{\mu}_i))\}\right],$$

so $\sum_{i=1}^n d_i^2 = D(y;\hat{\mu})$. Under SDA and σ^2 known, the d_i converge to normal distributions and

$$\frac{D(y,\hat{\mu})}{\sigma^2} = \omega_{\text{LR}}(H_0) \xrightarrow{d} \chi^2_{n-p}$$

under the fitted model, so we can test it versus the saturated model.

2.8 Model selection

AIC applies to GLMs without modification and the inference after model selection warnings do too. R^2 and forward and backward selection must be adapted by replacing RSS by the appropriate deviances e.g.

$$R^2 = \frac{D(y,\overline{\mu}) - D(y,\hat{\mu})}{D(y,\overline{\mu})},$$

$\overline{\mu}$ is the fitted means for the intercept-only model. Orthogonality does not hold generally in GLMs.

3 Specific regression problems

Data: $(y_1, x_1), \ldots, (y_n, x_n) \in \mathbb{R} \times \mathbb{R}^p$.

3.1 Binomial regression

Model: $Y_i \overset{\text{ind}}{\sim} \frac{1}{n_i}\text{Bin}(n_i, \mu_i), n \in \mathbb{N}, \mu_i \in (0,1), g(\mu_i) = X_i^T\beta$. Check this is a GLM with $\sigma^2 = 1, a_i = \frac{1}{n_i}, \theta(\mu_i) = \log\frac{\mu_i}{1-\mu_i}$.

3.1.1 Link functions

Ideally, $g((0,1)) \to \mathbb{R}$. Indeed, in increasing order of popularity:

(i) Complementary log-log link $g(\mu) = \log(-\log(1-\mu))$

(ii) Probit link $g(\mu) = \Phi^{-1}(\mu)$, Φ is the cdf of the standard normal distribution.

(iii) Logit (canonical) link $g(\mu) = \log\frac{\mu}{1-\mu}$

Claim: If $n_i = 1 \ \forall \ i$, in the three cases (and more generally), we can write

$$Y_i = 1\{Y_i^* > 0\}, Y_i^* = X_i\beta + \varepsilon_i, \varepsilon \overset{\text{iid}}{\sim} F$$

for some cdf F. Indeed,

$$\mu_i = \mathbb{E}[Y_i] = \mathbb{P}(Y_i^* > 0) = \mathbb{P}\left(\varepsilon_i > -X_i^T\beta\right) = 1 - F(-X_i^T\beta),$$

and if $F^{-1}(p) = \inf\{x \in \mathbb{R}, F(x) \geq p\}$, then $X_i^T\beta = -F^{-1}(1-\mu_i)$. Note if F is symmetric about the origin, i.e. $F(t) = 1 - F(-t)$, then $-F^{-1}(1-\mu_i) = F^{-1}(\mu_i)$. Thus, the coefficients in theses binomial regression models ($n_i = 1 \ \forall \ i$) van be interpreted as the effects of a unit increase in the corresponding covariate on a latent variable that satisfies a linear model with log-Weibull, standard normal, standard logistic distributions respectively. The logit is the most popular because it is the canonical link and due to its additional interpretability. Note that

$$\frac{\mu_i}{1-\mu_i} = \prod_{j=1}^{p}(e^{\beta_j})^{X_{ij}},$$

so e^{β_j} is the multiplicative change in the odds $\frac{\mu_i}{1-\mu_i}$ for a unit increase in the jth variable.

3.2 Poisson regression

Model: $Y_i \overset{\text{iid}}{\sim} \text{Poi}(\mu_i), \mu_i > 0 g(\mu_i) = X_i^T\beta$. On the example sheet we see this is a GLM with $\sigma^2 = 1 = a_i \ \forall \ i$ and $\theta(\mu_i) = \log\mu_i$. Recall $\mathbb{E}[Y_i] = \text{var}\, Y_i = \mu_i$ so this can be a limiting in modelling counting data. If Y_i counts independent events occurring in a time interval and the probability of occurrence is proportional to this time interval, the model is exact.

3.2.1 Link functions

Again, wish $g((0,\infty)) = \mathbb{R}$. Most popular link is the log (or canonical) link (aka log-linear regression model). Note $\mu_i = \prod_{j=1}^{p}(e^{\beta_j})^{X_{ij}}$ so e^{β_j} as the multiplicative change in the expected response value for a unit increase of the jth variable (rest of covariates fixed).

3.2.2 Deviance and Pearson chi-squared statistic

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

$$D(y:\hat{\mu}) = 2[\overline{\ell}(y,\sigma^2) - \overline{\ell}(\hat{\mu},\sigma^2)] = 2\left[\sum_{i=1}^{n} y_i \log\frac{y_i}{\hat{\mu}_i} - \sum_{i=1}^{n}(y_i - \hat{\mu}_i)\right].$$

Example sheet: for any GLM with $\sigma^2 = a_i = 1\ \forall\ i, g$ canonical and including an intercept, $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\hat{\mu}_i$ hence

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

Assume $\frac{\delta_i}{\hat{\mu}_i}$ small. Then, by Taylor and $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\hat{\mu}_i$

.

$$\begin{aligned} D(y;\hat{\mu}) &= 2\sum(\hat{\mu}_i + \delta_i)\left(\frac{\delta_i}{\hat{\mu}_i} - \frac{1}{2}\left(\frac{\delta_i}{\hat{\mu}_i}\right)^2 + O\left(\frac{\delta_i}{\hat{\mu}_i}\right)^3\right) \\ &= 2\sum(\delta_i + \frac{1}{2}\frac{\delta_i^2}{\hat{\mu}_i}) + O\left(\frac{\delta_i}{\hat{\mu}_i}^2\right) \\ &\approx \sum_{i=1}^{n}\frac{\delta_i^2}{\hat{\mu}_i} \\ &= \sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}. \end{aligned}$$

The RHS is Pearson's χ^2-statistic and is better approximated by χ^2_{n-p} than the deviance (under SDA).

3.2.3 Contingency tables

Assume we have counting data classified according to r factors i.e. $Y_{i_1,\ldots,i_r},\ i_s = 1,\ldots,I_s, s = 1,\ldots,r$. An r-way contingency table is a concise way to present this type of data e.g. conduct an online survey where, for given period of time, participants reveal their college and voting intentions, if we add up participants with the same characteristics we can present the data by a 2-way contingency

ollege
Labour
Conservatives
other
Darwin

table. A possible model is $Y_{ij} \overset{\text{iid}}{\sim} \text{Poi}(\mu_{ij}), i = 1, \dots I, j =$

Clare

$1, \dots, J, \mu_{ij} > 0, \log \mu_{ij} = \alpha + X_{ij}^t \beta$ where α is an explicit intercept. The log-likelihood is

$$\begin{aligned}\ell_p(\alpha, \beta) &= \log\left(\prod_{i,j} e^{-\mu_{ij}} \frac{\mu_{ij}^{Y_{ij}}}{Y_{ij}!}\right) \\ &= \sum_{i,j} [Y_{ij} \log \mu_{ij} - \mu_{ij}] + C \\ &= \alpha \sum_{ij} Y_{ij} + \sum_{ij} X_{ij}^T \beta - \sum_{ij} \exp(X_{ij}^T \beta) + C.\end{aligned}$$

Multinomial model

Assume that instead of conducting the survey for a fixed amount of time, we run it until we get n responses which means the responses cannot be independent. The mmultinomial model is a better candidate. Recall that $(Z_1, \dots, Z_n) \sim \text{Multi}(n, p_1, \dots, p_m) n, m \in \mathbb{N}, p_k \in [0,1], k = 1, \dots, m, \sum_{k=1}^m p_k = 1$ if $\mathbb{P}(Z_1 = z_1, \dots, Z_m = z_m) = \frac{n!}{z_1! \cdots z_m!} p_1^{z_1} \cdots p_m^{z_m}$ i.e. it counts the outcomes of n independent rolls of a biased die with m faces; we can model $(Y_{ij}) \sim \text{Multi}(n, (p_{ij}))$ for $i = 1, \dots, I, j = 1, \dots, J$ where $p_{ij} = \frac{\mu_{ij}}{\sum_{i',j'} \mu_{i'j'}}$ and $\log \mu_{ij} = X_{ij}^T \beta$ (wlog $\alpha = 0$). Note: the x_{ij} depend on the fitted model e.g. if we assume that the voting intentions are independent of the college, then $p_{ij} = q_i r_j, q_i, r_j \in [0,1], \sum_{i=1}^I q_i = \sum_{j=1}^J r_j = 1$ and we can take $\beta \in \mathbb{R}^{I+J}, \beta_i = \log q_i, \beta_{I+j} = \log r_j, \beta_1 = \beta_I = 0$ and

$$X_{ij}^T = (\underbrace{0, \dots, 0, \overset{i}{1}, 0, \dots, 0}_{I} \underbrace{0, \dots, 0, \overset{I+j}{1}, 0 \dots, 0}_{J}).$$

The log-likelihood for the multinomial model is

$$\begin{aligned}\ell_\mu(\beta) &= \log\left(\frac{n!}{\prod_{i,j} Y_{ij}!}\prod_{i,j} p_{ij}^{Y_{ij}}\right) = \sum_{i,j} Y_{ij}\log p_{ij} + C \\ &= \sum_{i,j} Y_{ij}\left[\log\mu_{ij} - \log\sum_{i',j'}\mu_{i'j'}\right] + C \\ &= \sum_{i,j} Y_{ij}X_{ij}^T\beta - \left(\sum_{i,j} Y_{ij}\right)\log\left(\sum_{i,j}\mu_{ij}\right) + C \\ &= \sum_{i,j} Y_{ij}X_{ij}^T\beta - n\log\left(\sum_{i,j}\mu_{ij}\right) + C.\end{aligned}$$

Remark. Note that the multinomial model cannot be a GLM (dependent responses). However, it is intimately related to the (conditional) Poisson log-linear model, and we can use the theory for the latter.

Connection between Poisson and multinomial models
Let us reparametrise $(\alpha, \beta) \mapsto (\tau, \beta)$ in the Poisson model, where

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

Then if $\sum_{i,j} Y_{ij} = n$

$$\begin{aligned}\tilde{\ell}_p(\tau,\beta) &= \sum_{i,j} Y_{ij}X_{ij}^T\beta + n\log\frac{\tau}{\sum_{i,j}\exp(X_{ij}^T\beta)} - \tau + C \\ &= \sum_{ij} Y_{ij}X_{ij}^T\beta - n\log\left(\sum_{i,j}\exp\left(X_{ij}^T\beta\right)\right) + n\log\tau - \tau + C \\ &= \ell_m(\beta) + \ell_p(\tau).\end{aligned}$$

We can maximise over $\beta\&\tau$ separately and, under "RCs" (so unique MLEs)

$$\hat{\beta}_m = \text{argmax}_\beta\ell_m(\beta) = \text{argmax}_\beta\tilde{\ell_p}(\tau,\beta) = \hat{\beta}_p = \hat{\beta},$$

and $\hat{\tau} = n$. So,

$$\hat{\alpha} = \log\frac{n}{\sum_{i,j}\exp(X_{ij}^T\hat{\beta})}.$$

Consequently,

(i) The deviances are equal

(ii) The Fisher information matrix's coincide

(iii) The fitted values are the same i.e.

$$n\hat{p}_{ij} = n\frac{e^{X_{ij}^T\hat{\beta}}}{\sum_{i',j'} e^{X_{i'j'}^T\hat{\beta}}} = e^{\hat{\alpha}+X_{ij}^T\hat{\beta}} = \hat{\mu}_{ij}.$$

The multinomial models can be fitted by fitting Poisson log-linear models with intercept. Poisson models used for this purpose are called surrogate Poisson models. Underlying this is the following result.

Proposition. Let $Z_k \overset{\text{iid}}{\sim} \text{Poi}(\mu_k), \mu_k > 0, k = 1, \ldots, m$. Then

$$(Z_1, \ldots, Z_m)\Big|_{\sum_{k=1}^m Z_k = n} \sim \text{Multi}(n; p_1, \ldots, p_m),\ p_k = \frac{\mu_k}{\sum_{k'=1}^m \mu_{k'}}.$$