Statistical Modelling II

2023-2024

260923

Evaluation [80% exam (January - 10h) (online)
10% } In-term exam / coursework / quiz
10% } week 7 and week 12 (online)

Activities Tuesday 1300-1500 Bancroft 2.40 week 1-6, 8-12 } ALL students

one of { Monday 1100-1300 Bancroft 1.15A week 2-6, 8-12
Wednesday 0900-1100 Maths 302 week 2-6, 8-12

Syllabus Statistical Modelling II MTH6134
"Generalized linear models"

		(weeks)	
1	Introduction	1	} HMA
2	Normal linear model (regression) (normal)	2-4	
3	Generalized linear models	4-6	
4	Binary data (logistic regression) (binomial)	8-9	} Dr Silvia Livierani
5	Count data (Poisson regression) (Poisson)	9-11	
6	Survival data (exponential regression) (exp)	11-12	

Module resources
- Qreview (automatic)
- Typed notes } always download the latest version from qmplus
- Exercises
- Labs from qmplus
- Scanned notes
- (Forum) qmplus

Motivating examples: Seattle heart failure model
MAGGIC heart failure calculator

1 - Introduction

1.1 Topics to cover

Let us recall the standard linear regression (cp in SMI). The model explains a vector of response values Y using explanatory variables. The model is

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

- Y: response vector, $n \times 1$
- X: design matrix, $n \times p$
- β: parameter vector, $p \times 1$
- ε: vector of errors, $n \times 1$, has multivariate normal distribution with zero mean (vector) and covariance matrix $\Sigma = \sigma^2 I$.

The vector of estimates is $\hat{\beta} = (X^T X)^{-1} X^T Y$

with variance-covariance matrix

$$V(\hat{\beta}) = \sigma^2 (X^T X)^{-1}.$$

Refresher!

Generalized linear models are a family of models that include the above regression as a special case. We will use distributions from the exponential family to model count data (Poisson) (Negative binomial) zero/one data (Bernoulli, binomial), Survival data (exponential, gamma), thus extending what we can do with regression.

1.2 Examples of generalized linear models

All the models in this course consist of two components: a distribution and a link function (this function relates the expectation with the predictor).

- Binary data - In a clinical trial, we give r_i patients a dose x_i of a drug and Y_i is the number of those having positive response $Y_i \sim Bin(r_i, \pi_i)$

Suppose we have n trials $\begin{cases} r_1, r_2, \ldots r_n \text{ patients} \\ x_1, x_2, \ldots x_n \text{ doses} \\ Y_1, Y_2, \ldots Y_n \text{ response values} \end{cases}$

Some models...

- all the π_i are equal - only one parameter to estimate $\pi = \pi_1 = \pi_2$
- The trials are conducted in three hospitals - 3 parameters
- each trial with own probability - n parameters
- The probability depends on the dose...
  - linearly $\pi_i = \beta_0 + \beta_1 x_i$ (identity link) — dist. Binomial link identity
  - logistic model $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$ (logit link) — dist. Binomial link logit

- Normal data (same as above)
  Consider a single predictor x_i so response is $Y_i \sim N(\mu_i, \sigma^2)$ so that we relate expectation μ_i with link to the linear predictor and have

$$\mu_i = \beta_0 + \beta_1 x_i$$
$$g(\mu_i) = \beta_0 + \beta_1 x_i$$

dist. Normal link identity

031023

- Poisson data $Y_i \sim \text{Poisson}(\mu_i)$ we relate the poisson to the linear predictor using log-link

$$\log(\mu_i) = \beta_0 + \beta_1 x_i \quad \text{with inverse} \quad \mu_i = e^{\beta_0 + \beta_1 x_i}$$

Dist. Poiss link log

Exercise. Invert the logit link $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$

$$\pi_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}$$

2 - Normal model

2.1 Likelihood

In statistics, we have several methods of estimating parameters: least squares, maximum likelihood, minimum chi-squa

Bayesian methods.

Likelihood is a general estimation methodology that can be applied to models involving distributions.

Notation: $\underline{\theta} = (\theta_1, \theta_2, \ldots \theta_p)$ vector of parameters to be estima

Ex. $\underline{\theta} = (\mu, \sigma^2)$ in normal

$\underline{\theta} = (\pi)$ for binomial

$\underline{y} = (y_1, \ldots, y_n)^T$ vector of observations ← this is the data, which are realization

$\underline{Y} = (Y_1, Y_2, \ldots Y_n)^T$ vector of random variables

Definition (2.1) For a discrete random variable, the likelihood is the probability of observing the data, as function of the parameter

$$L(\underline{\theta}; \underline{y}) = L(\underbrace{\theta_1, \ldots, \theta_p}_{\text{parameter(s)}}; \underbrace{y_1, \ldots y_n}_{\text{data is fixed}}) = Pr(\underline{Y} = \underline{y}; \underline{\theta})$$

$$= Pr(Y_1 = y_1, Y_2 = y_2, \ldots Y_n = y_n; \theta_1, \ldots, \theta_p)$$ under independence of r.v.

$$= \prod_{i=1}^{n} Pr(Y_i = y_i; \underline{\theta})$$

For continuous random variable, we use the joint probability density of the data

~~$L(\theta_1, \ldots \theta_p)$~~

$$L(\theta_1, \ldots, \theta_p; y_1, \ldots y_n) = f_{Y_1, Y_2, \ldots Y_n}(y_1, y_2, \ldots, y_n; \underline{\theta})$$ under independe

$$= \prod_{i=1}^{n} F_{Y_i}(y_i; \underline{\theta})$$

Example. Simple linear regression $Y_i \sim N(\beta_0+\beta_1 x_i, \sigma^2)$ $i=1,\dots,n$ all independent so $\underline{\theta}=(\beta_0,\beta_1,\sigma^2)$ and the likelihood of $\underline{\theta}$ given data is

$$L(\beta_0,\beta_1,\sigma^2; y_1, y_2, \dots y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i-\beta_0-\beta_1 x_i)^2}{\sigma^2}}$$

$$= (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}$$

Recall normal density
$X \sim N(\mu,\sigma^2)$
$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$

2.2 Method of maximum likelihood

The maximum likelihood estimates $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_p$ are the values that maximize the likelihood $L(\underline{\theta}; \underline{y})$. We commonly use traditional calculus techniques to maximize $L(\underline{\theta}; \underline{y})$. As $L(\underline{\theta}; \underline{y})$ is a product of functions, we work with the log-likelihood

$$\ell(\underline{\theta}; \underline{y}) := \log L(\underline{\theta}; \underline{y}).$$

We find derivatives (partial derivatives) $\frac{\partial \ell}{\partial \theta_j}$ and set up a simultaneous system of equations

$$\frac{\partial \ell}{\partial \theta_j} \stackrel{!}{=} 0.$$

m.l.e.

We write the maximum likelihood estimate $\hat{\underline{\theta}} = \hat{\underline{\theta}}(\underline{Y})$ to emphasize its dependence on data.

Example (regression, continued) The log-likelihood is

$$\ell(\beta_0,\beta_1,\sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2.$$

$\ell(\underbrace{\beta_0,\beta_1,\sigma^2}_{\underline{\theta}\text{ vector of parameters}}; \underline{y}) =$

To compute MLE, we do partial derivatives with respect to β_0, β_1, σ

$$\frac{\partial l}{\partial \beta_0} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial l}{\partial \beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i)$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

} Simultaneous system of equations leading to m.l.e.s

From the first two derivatives, we have the system

$$\begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

and retrieve the old-time estimates

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

From the third derivative, we have

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$\hat{\beta}_0, \hat{\beta}_1$ unbiased (like SMI)
$\hat{\sigma}^2$ biased (not like SMI)

In fact $E(\hat{\sigma}^2) = \frac{n-2}{n}\sigma^2$ that shows the bias and an unbiased estimate of σ^2 is $\frac{n}{n-2}\hat{\sigma}^2$.

2.3 Large sample distribution of the maximum likelihood estimator

We have the following result for asymptotic cases (large sample

Theorem 2.1 Under general conditions, for large n,

$$\hat{\underset{\sim}{\theta}} \sim N_p\left(\underset{\sim}{\theta}, V^{-1}\right),$$

$\hat{\underset{\sim}{\theta}}$: m.l.e.
N_p: multivariate normal (p-variate)
$\underset{\sim}{\theta}$: vector of true parameters
V^{-1}: Variance-covariance matrix.

where V is the $p \times p$ Expected Fisher information matrix defined with (i,j) element as

$$V_{ij} = E\left(-\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}\right) \quad \text{for } i,j \text{ in } 1,2,\ldots p.$$

101023

By Theorem 2.1, every m.l.e. $\hat{\theta}_j$ is asymptotically normal, i.e. $\hat{\theta}_j \sim N(\theta_j, V^{jj})$. → j-th element of diagonal of V^{-1}

Exercise/Example. For the running regression model $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ $i = 1,\ldots,n$, compute derivatives for FIM

Derivatives	Expectation $V_{ij} = E\left(-\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}\right)$
$\frac{\partial^2 l}{\partial \beta_0^2} = -\frac{n}{\sigma^2}$	$V_{11} = \frac{n}{\sigma^2}$
$\frac{\partial^2 l}{\partial \beta_1^2} = -\frac{\sum_{i=1}^n x_i^2}{\sigma^2}$	$V_{22} = \frac{\sum_{i=1}^n x_i^2}{\sigma^2}$
$\frac{\partial^2 l}{\partial \sigma^4} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$	$V_{33} = -\left(\frac{n}{2\sigma^4} - \frac{n\sigma^2}{\sigma^6}\right) = -\left(-\frac{n}{2\sigma^4}\right) = \frac{n}{2\sigma^4}$
$\frac{\partial^2 l}{\partial \beta_0 \partial \beta_1} = -\frac{\sum_{i=1}^n x_i}{\sigma^2}$	$V_{12} = \frac{\sum_{i=1}^n x_i}{\sigma^2}$
$\frac{\partial^2 l}{\partial \beta_0 \partial \sigma^2} = -\frac{1}{\sigma^4}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)$	$V_{13} = 0$
$\frac{\partial^2 l}{\partial \beta_1 \partial \sigma^2} = -\frac{1}{\sigma^4}\sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i)$	$V_{23} = 0$

$A \sim N(0,\sigma^2)$

$$E\big(\underbrace{Y_i - \beta_0 - \beta_1 x_i}_{N(\beta_0+\beta_1 x_i, \sigma^2)}\big)^2 = \sigma^2$$

$E(A^2) = V(A) + (E(A))^2$

We have FIM

$$V = \frac{1}{\sigma^2}\begin{pmatrix} n & \sum_{i=1}^n x_i & 0 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & 0 \\ 0 & 0 & n/2\sigma^2 \end{pmatrix}$$

so the asymptotic variance-covariance matrix is

$$V^{-1} = \sigma^2 \begin{pmatrix} \frac{\sum x_i^2}{n S_{xx}} & -\frac{\bar{x}}{S_{xx}} & 0 \\ -\frac{\bar{x}}{S_{xx}} & \frac{1}{S_{xx}} & 0 \\ 0 & 0 & \frac{2\sigma^2}{n} \end{pmatrix}.$$

with $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$

In summary, we have asymptotic distributions

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \frac{\sum x_i^2}{n S_{xx}}\right), \quad \hat{\beta}_1 \sim N\left(\beta_1, \sigma^2/S_{xx}\right), \quad \hat{\sigma}^2 \sim N\left(\sigma^2, \frac{2\sigma^4}{n}\right).$$

2.4 Multiple linear regression

We consider a regression with $p-1$ explanatory variables and intercept term so

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_{p-1} x_{p-1,i} = x_i^T \beta$$

and $V(Y_i) = \sigma^2$, i.e. $Y_i \sim N(x_i^T \beta, \sigma^2)$.

If we collect observations $y_1, \ldots, y_n$ in vector $\underline{y}$ and the covariate values in a matrix

$$\underline{X} = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p-1,1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p-1,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{p-1,n} \end{pmatrix} \quad \text{and} \quad \underline{\beta} = (\beta_0, \beta_1, \ldots \beta_{p-1})^T$$

we formulate the regression using matrices as follows

$$\underline{Y} \sim N_n(\underline{X}\underline{\beta}, \sigma^2 I)$$

$$E(\underline{Y}) = \underline{X}\underline{\beta}$$

In order to estimate parameters, we build the likelihood

$$L(\underline{\beta}, \sigma^2; \underline{y}) = \left((2\pi)^n \sigma^{2n}\right)^{-1/2} e^{-\frac{1}{2\sigma^2}(\underline{y}-\underline{X}\underline{\beta})^T(\underline{y}-\underline{X}\underline{\beta})}$$

with log-likelihood

$$\underline{y}^T\underline{y} - 2\underline{\beta}^T\underline{X}^T\underline{y} + \underline{\beta}^T\underline{X}^T\underline{X}\underline{\beta}$$

$$\ell(\underline{\beta}, \sigma^2; \underline{y}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\underline{y}-\underline{X}\underline{\beta})^T(\underline{y}-\underline{X}\underline{\beta})$$

Following our template, we compute m.l.e. and require derivatives of ℓ with respect to all parameters.
From (matrix cook book) matrix calculus, we compute the derivatives

$$\frac{\partial \ell}{\partial \underline{\beta}} = \frac{1}{\sigma^2}(\underline{X}^T\underline{y} - \underline{X}^T\underline{X}\underline{\beta}) = \frac{1}{\sigma^2}\underline{X}^T(\underline{y}-\underline{X}\underline{\beta}) \longrightarrow \hat{\underline{\beta}} = (\underline{X}^T\underline{X})^{-1}\underline{X}^T\underline{y}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\underline{y}-\underline{X}\underline{\beta})^T(\underline{y}-\underline{X}\underline{\beta}) \longrightarrow \hat{\sigma}^2 = \frac{1}{n}(\underline{y}-\underline{X}\hat{\underline{\beta}})^T(\underline{y}-\underline{X}\hat{\underline{\beta}})$$

For this regression model, the Fisher information matrix (FIM) is

$$V = \begin{pmatrix} \sigma^{-2}\underline{X}^T\underline{X} & \underline{0} \\ \underline{0} & n/2\sigma^4 \end{pmatrix} \quad \text{so} \quad V^{-1} = \begin{pmatrix} \sigma^2(\underline{X}^T\underline{X})^{-1} & \underline{0}_{p\times 1} \\ \underline{0}_{1\times p} & \frac{2\sigma^4}{n}_{1\times 1} \end{pmatrix}$$

$p\times p$; $(p+1)\times(p+1)$

For large n $\hat{\underline{\beta}} \sim N_p(\underline{\beta}, \sigma^2(\underline{X}^T\underline{X})^{-1})$ and

$\hat{\sigma}^2 \sim N(\sigma^2, 2\sigma^4/n)$.

171023

2.5 Generalized likelihood ratio tests

Some notation

Ω - parameter space of $\underline{\theta}$ (all possible values that $\underline{\theta}$ can take)
$\omega \subset \Omega$ - a subset of the parameter space used to define a hypothesis

We want to test $H_0: \underline{\theta} \in \omega$ (null) against $H_1: \underline{\theta} \in \Omega/\omega$ (alternative)

Example. In linear regression we are interested in testing $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

(Recall model $Y_i \sim N(\beta_0 + \beta_1 x_i; \sigma^2)$)

The parameter space is $\Omega = \underset{\beta_0}{\mathbb{R}} \times \underset{\beta_1}{\mathbb{R}} \times \underset{\sigma^2}{\mathbb{R}^+} = \mathbb{R}^2 \times \mathbb{R}^+$

For the hypothesis $\omega = \underset{\beta_0}{\mathbb{R}} \times \underset{\beta_1}{\{0\}} \times \underset{\sigma^2}{\mathbb{R}^+}$ ← tailored to H_0

Sample space $S = \mathbb{R}^n$

The set of all values $\underline{y}$ that the random variable $\underline{Y}$ can take is the sample space, termed S. To test the hypothesis, the critical region is the subset $R \subset S$ such that we reject H_0 if $\underline{y} \in R$.

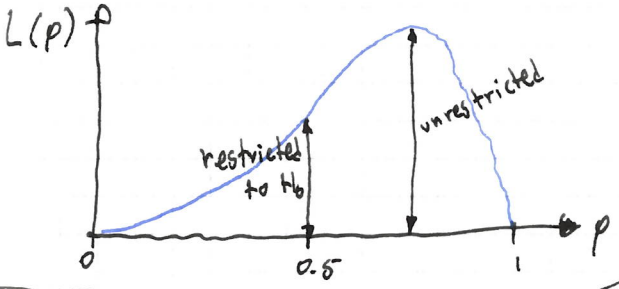
Definition. The generalized likelihood ratio test has critical region $R = \{\underline{y} : \Lambda(\underline{y}) < a_\alpha\}$ where

$$\Lambda(\underline{y}) = \frac{\max_{\underline{\theta} \in \omega} L(\underline{\theta}; \underline{y})}{\max_{\underline{\theta} \in \Omega} L(\underline{\theta}; \underline{y})}$$

— restricted to H_0

— unrestricted

Example Binomial lik. $H_0: p = 0.5$



is the generalized likelihood ratio and a_α is a constant chosen to give a test with significance α

Clearly $0 \leq \Lambda(\underline{y}) \leq 1$ because $\omega \subset \Omega$.

Let $\hat{\underline{\theta}}_0$ be the value of $\underline{\theta}$ that maximizes the likelihood in ω and $\hat{\underline{\theta}}$ is the usual m.l.e. that maximizes over Ω. We write

$$\Lambda(\underline{y}) = \frac{L(\hat{\underline{\theta}}_0; \underline{y})}{L(\hat{\underline{\theta}}; \underline{y})}$$

— restricted to H_0

— unrestricted

Example. Linear regression testing $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

The (unrestricted) m.l.e.s are $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, $\hat{\beta}_1 = S_{xy}/S_{xx}$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

$\hat{\underline{\theta}}$ m.l.e.

The m.l.e.s restricted to H_0 are $\hat{\beta}_{00} = \bar{y}$, $\hat{\beta}_{10} = 0$, $\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$\hat{\underline{\theta}}_0$ restricted

The generalized likelihood ratio $\Lambda(\underline{y})$ is

$$\Lambda(\underline{y}) = \frac{L(\hat{\beta}_{00}, \hat{\beta}_{10}, \hat{\sigma}_0^2; \underline{y})}{L(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2; \underline{y})} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} = \left(\frac{SS_E}{SS_E + SS_R}\right)^{n/2}$$

Reading task! In the notes, I give details of the simplifications, follow them.

We reject $H_0: \beta_1 = 0$ if $\left(\frac{SS_E}{SS_E + SS_R}\right)^{n/2} < a_\alpha$. Equivalently, we reject H_0 if

$$\frac{MS_R}{MS_E} = \frac{SS_R/1}{SS_E/(n-2)} > (n-2)\left(a_\alpha^{-2/n} - 1\right).$$

task

Fisher's F distribution

We have shown that the test using $\Lambda(\underline{y})$ is equivalent to the F-test in regression. This holds for tests in normal theory :)

2.6 Wilks' theorem

The form of the distribution of $\Lambda(\underline{y})$ under H_0 can be known or estimated e.g. by simulation. We can use the following result instead.

Theorem. Suppose that $Y_1, Y_2, \ldots, Y_n$ have a joint distribution depending on parameters $\underline{\theta}$ and consider testing

$$H_0: \underline{\theta} \in \omega \text{ vs. } H_1: \underline{\theta} \in \Omega \setminus \omega.$$

Under regularity conditions, under H_0 and large n

$$-2 \log \Lambda(\underline{y}) \sim \chi^2_s$$

chi-square

where s, the degrees of freedom is the number of constraints imposed by H_0.

Example. Poisson data from lab and $H_0: \mu = 3$.

241023

Some dates coming:

"Mathematical life stories"	Wednesday 8th Nov.	1200-1300	Venu tbc
Mid term test week 7 (online)	Thursday 9th Nov.	1000-1100	Onli
Midterm test Week 12	Wednesday 13th Dec.	0900-1000	onli

---

3 - Generalized linear models

3.1 - Exponential family of distributions

Definition 3.1 The random variable Y with parameters $\underline{\theta} = (\theta, \ldots$ has a distribution in the exponential family if its range does not depend on the parameters and its probability mass function (discrete) or probability density function (continuous) can be written as

$$f_Y(y; \underline{\theta}) = \exp\left(\sum_{j=1}^{p} a_j(y)\, b_j(\theta) + c(\underline{\theta}) + d(y) \right);$$

for the one-parameter exponential family, this reduces to

$$f_Y(y; \theta) = \exp\left(a(y) b(\theta) + c(\theta) + d(y) \right).$$

Example Poisson is exponential

$$f_Y(y; \mu) = \frac{\mu^y e^{-\mu}}{y!} = \exp\left(\underbrace{y}_{a(y)} \underbrace{\log(\mu)}_{b(\mu)} \underbrace{- \mu}_{c(\mu)} \underbrace{- \log(y!)}_{d(y)} \right)$$

If $a(y) = y$ then the distribution is in canonical form and $b(\theta)$ is the natural parameter

For the one-parameter exponential family, we have

Lemma 3.3 Suppose that Y is a distribution in the one-parameter exponential family. Then

$$E(a(Y)) = -\frac{c'(\theta)}{b'(\theta)} \quad \text{and} \quad V(a(Y)) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}.$$

(See examples in exercise session)

We now move towards estimation... if we have data $y_1, \ldots, y_n$ realizations of $Y_1, \ldots, Y_n$ of the exponential family, we build the likelihood as follows

$$L(\theta; \underline{y}) = \prod_{i=1}^{n} \exp\left(a(y_i)\, b(\theta) + c(\theta) + d(y_i)\right)$$

(θ: one parameter)

$$= \exp\left(\sum_{i=1}^{n} a(y_i)\, b(\theta) + n\, c(\theta) + \sum_{i=1}^{n} d(y_i)\right) \quad (*)$$

3.2 The generalized linear model (GLM)

This model unifies a collection of statistical methods involving linear combinations of parameters.

- regression
- logistic regression
- Poisson regression

We start with the likelihood $L(\theta; y)$ for exponential family $(*)$, assuming the distribution is in canonical form so

$$L(\underline{\theta}; \underline{y}) = \exp\left(\sum_{i=1}^{n} y_i\, b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum d(y_i)\right).$$

($\underline{\theta}$: one parameter per observation)

We have θ_i because we do not assume that the distributions share the same parameter. The parameters $\theta_1, \ldots, \theta_n$ are not of interest because there is only one parameter per observation.

In a GLM, we consider parameters $\beta_1, \ldots, \beta_p$ with $p < n$.

Suppose that

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

$\beta = (\beta_1, \ldots \beta_p)^T$

$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ — p covariates

x_{ij} is level of j-th covariate for i-th observation

$\mu_i = g^{-1}(x_i \beta)$

"link function" monotonic differentiable function

$\mu_i = E(Y_i)$

linear predictor η_i

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

Example. Normal $g(\cdot)$ is identity link $g(\mu_i) = \mu_i$

Poisson (counts) $g(\mu_i) = \log(\mu_i)$ log link

$$\eta_i = \log(\mu_i) \quad \text{so} \quad \mu_i = e^{\eta_i} = e^{x_i \beta}$$

Poisson with identity link

$$\eta_i = \mu_i \quad \text{so} \quad \mu_i = \eta_i = x_i \beta$$

Binomial (observations are proportions)

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

logit link — maps $(0,1)$ into $\mathbb{R}$

inverting $\pi_i = \dfrac{1}{1+e^{-\eta_i}} = \dfrac{1}{1+e^{-x_i\beta}}$

other transformations possible for binomial

$$\eta_i = \Phi^{-1}(\pi_i)$$ ← "probit" link

or

$$\eta_i = \log(-\log(1-\pi_i))$$ ← complementary log-log link

3.3 Fitting the model

At this point, model parameters are $\beta_1, \ldots, \beta_p$, and our starting point is the (log) likelihood

$$\ell(\theta; y) = \sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)$$

and the result $E(Y_i) = \mu_i = -\dfrac{c'(\theta_i)}{b'(\theta_i)}$ and use link

$g(\mu_i) = x_i \beta = \sum_{j=1}^{p} x_{ij} \beta_j = \eta_i$ and the $V(Y_i) = \dfrac{b''(\theta_i)\, c'(\theta_i) - c''(\theta_i) b'(\theta_i)}{(b'(\theta_i))^3}$.

3/11/23

"Mathematical life stories" Wednesday 8th Nov. 1200-1300 Peston LT

Extra lab "Wednesdays people" Tuesday 7th Nov. 1600-1800 Maths 302

Midterm test week 7 · Thursday 9th Nov 1000-1100 online

Midterm test week 12 Wednesday 13th Dec. 0900-1000 online

To estimate parameters, we need to compute derivatives of the likelihood $\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} \stackrel{!}{=} 0$ — j-th parameter, there are p parameters; sum over observations

where $l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i)$.

This derivative is computed using the chain rule

$$\frac{\partial l_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} \quad \#\#10$$

$$\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'(\theta_i) = y_i b'(\theta_i) - \mu_i b'(\theta_i) = b'(\theta_i)(y_i - \mu_i)$$

(Expectation)

To do $\frac{\partial \theta_i}{\partial \mu_i}$ use $\frac{\partial \mu_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \equiv 1$ so $\frac{\partial \theta_i}{\partial \mu_i} = 1 \Big/ \frac{\partial \mu_i}{\partial \theta_i}$ and

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial}{\partial \theta_i}\left(-\frac{c'(\theta_i)}{b'(\theta_i)}\right) = \frac{-b'(\theta_i)c''(\theta_i) + c'(\theta_i)b''(\theta_i)}{(b'(\theta_i))^2} = b'(\theta_i)V(Y_i)$$

so $$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b'(\theta_i)V(Y_i)}$$

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot x_{ij}$$

↳ because $\eta_i = x_i\beta = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ij}\beta_j + \ldots + x_{ip}\beta_p$

Collecting the results, we have

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{V(Y_i)} \cdot x_{ij}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \quad \text{and} \quad \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(Y_i)} x_{ij} \cdot \frac{\partial \mu_i}{\partial \eta_i}$$

(depends on parameters; uses the link)

Maximum likelihood estimates are obtained solving $\frac{\partial l}{\partial \beta_0} \stackrel{!}{=} 0, \frac{\partial l}{\partial \beta_1} \stackrel{!}{=} 0, \ldots, \frac{\partial l}{\partial \beta_{p-1}} \stackrel{!}{=} 0$ This is a simultaneous system of equations. In general, they are non-linear and must be solved numerically.

Example - Poisson regression, log-link.

Three models

	# Params
"Null" All Y_i have the same mean	1
"GLM" The Y_i have mean with $g(\mu_i)=\eta_i$	2
"Maximal" Each Y_i has its own mean	9

Model 'GLM' "supported" (i.e. not rejected)

Likelihood ratio test

	H_0		H_1	Statistic $-2\log\Lambda$	P Value
We can test (2)	Null	vs	GLM	15.4	8×10^{-5}
	Null	vs	Maximal	~~17.33~~ 18.4	~~0.89~~ once 0.018
(1)	GLM	vs	Maximal	2.93	0.89

Tried (1), do not reject "GLM", then tried (2) rejected "Null".

We have the following result for computing the Fisher information matrix.

Theorem 3.1. Suppose that random variables $Y_1,\dots,Y_n$ have distributions depending on parameters $\beta_1,\dots,\beta_p$ and their ranges do not depend on parameters. Then

$$E\left(-\frac{\partial^2 l(\underline{\beta};\underline{Y})}{\partial\beta_j\,\partial\beta_k}\right) = E\left(\frac{\partial l(\underline{\beta};\underline{Y})}{\partial\beta_j}\,\frac{\partial l(\underline{\beta};\underline{Y})}{\partial\beta_k}\right), \text{ for } j,k \text{ in } 1,2,\dots p.$$

When computing $E\left(-\frac{\partial^2 l_i}{\partial\beta_j\partial\beta_k}\right) = E\left(\frac{\partial l_i}{\partial\beta_j}\,\frac{\partial l_i}{\partial\beta_k}\right)$

we have

$$= E\left(\frac{Y_i-\mu_i}{V(Y_i)}\cdot x_{ij}\,\frac{\partial\mu_i}{\partial\eta_i}\cdot\frac{Y_i-\mu_i}{V(Y_i)}\,x_{ik}\,\frac{\partial\mu_i}{\partial\eta_i}\right)$$

$$= \frac{x_{ij}x_{ik}}{(V(Y_i))^2}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 E\left((Y_i-\mu_i)^2\right) = \frac{x_{ij}x_{ik}}{V(Y_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$$

adding over all observations we retrieve the j,k element of The FIM as

$$V_{jk} = E\left(-\frac{\partial^2 l}{\partial\beta_j\partial\beta_k}\right) = \sum_{i=1}^{n} \underbrace{\frac{x_{ij}x_{ik}}{V(Y_i)}}_{\text{covariates, variance of dist.}}\underbrace{\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2}_{\text{specific to the link}}.$$

Done as example/exercise in tutorial session.

4/11/23

13/11/23 (Tutorial Monday)

Null $L(\mu; \underline{y}) = \prod_{i=1}^{n} L(\mu; y_i)$ ← All observations share the same mean.

($L(\mu; y_i) = e^{-\mu} \mu^{y_i}/y_i!$)

Maximal $L(\underline{\mu}; \underline{y}) = \prod_{i=1}^{n} L(\mu_i; y_i)$ ← Each observation with own mean.

($L(\mu_i; y_i) = e^{-\mu_i} \mu_i^{y_i}/y_i!$)

Glm $L(\underline{\mu}(\underline{\beta}); \underline{y}) = \prod_{i=1}^{n} L(\mu_i(\underline{\beta}); y_i)$ ← $e^{-e^{\beta_0+\beta_1 x_i}} \left(e^{\beta_0+\beta_1 x_i}\right)^{y_i}/y_i! = e^{-\mu_i(\beta)} \mu_i(\beta)^{y_i}/y_i!$

$\beta = (\beta_0, \beta_1)$ $\quad \underbrace{\mu_i = e^{\eta_i} = e^{\beta_0+\beta_1 x_i}}_{\log\text{-link}}$

Maximal

$$L(\underline{\mu}; \underline{y}) = \prod_{i=1}^{n} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!} = \frac{e^{-\sum_{i=1}^{n} \mu_i} \prod_{i=1}^{n} \mu_i^{y_i}}{\prod_{i=1}^{n} y_i!} = \frac{e^{-\mu_1-\mu_2-\dots-\mu_n} \mu_1^{y_1} \mu_2^{y_2} \cdots \mu_n^{y_n}}{\prod_{i=1}^{n} y_i!}$$

$$\ell(\underline{\mu}; \underline{y}) = -\mu_1 - \mu_2 - \dots - \mu_n + y_1 \log \mu_1 + y_2 \log \mu_2 + \dots + y_n \log \mu_n - \log \prod_{i=1}^{n} y_i!$$

$\hat{\mu}_i = y_i$ $\quad \left| \frac{\partial \ell}{\partial \mu_1} = -1 + \frac{y_1}{\mu_1} \stackrel{!}{=} 0 \right.$ $\quad \frac{y_1}{\mu_1} = 1 \Rightarrow \hat{\mu}_1 = y_1$

$$\Lambda(\underline{y}) = \frac{L(\underline{\mu}(\hat{\beta}); \underline{y})}{L_{max}(\hat{\underline{\mu}}; \underline{y})} = \prod_{i=1}^{n} \frac{e^{-\mu_i(\hat{\beta})} \mu_i(\hat{\beta})^{y_i}}{e^{-\hat{\mu}_i} \hat{\mu}_i^{y_i}} = \prod_{i=1}^{n} \frac{e^{-\mu_i(\hat{\beta})} \mu_i(\hat{\beta})^{y_i}}{e^{-y_i} y_i^{y_i}}$$

$$= \prod_{i=1}^{n} \left(e^{y_i - \mu_i(\hat{\beta})} \mu_i(\hat{\beta})^{y_i} / y_i^{y_i} \right) = e^{\sum_{i=1}^{n} (y_i - \mu_i(\hat{\beta}))} \prod_{i=1}^{n} \left(\frac{\mu_i(\hat{\beta})}{y_i} \right)^{y_i}$$

14/11/23 The Fisher Information Matrix can be written as

$$V = X^T W X$$

with $W = \text{diag}(w_1, w_2, \dots, w_n)$ and $w_i = \frac{1}{V(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$.

By Theorem 2.1, for large sample size n, we have

$$\hat{\beta} \sim N_p(\beta, (X^T W X)^{-1}).$$ ← Distribution of m.l.e.s

In the normal case (linear regression SMI) we have $V(Y_i) = \sigma^2$ and $\frac{\partial \mu_i}{\partial \eta_i} = 1$ so $w_i = \frac{1}{\sigma^2}$ and $W = \frac{1}{\sigma^2} I$ so $V = X^T \frac{1}{\sigma^2} I X = X^T X / \sigma^2$.

How to solve numerically the equations and estimate mles?

→ One approach is the Newton-Rhapson method

Start from $\underline{\beta}^{(0)}$

iterate $\underline{\beta}^{(m)} = \underline{\beta}^{(m-1)} - \left[H^{(m-1)}\right]^{-1} \left.\frac{\partial l}{\partial \underline{\beta}}\right|_{\underline{\beta}=\underline{\beta}^{(m-1)}}$ for $m=1,2,3,\ldots$

Two possibilities → use the Hessian $H^{(m-1)}$ that has elements (j,k) $\left.\frac{\partial^2 l}{\partial \beta_j \partial \beta_k}\right|_{\underline{\beta}=\underline{\beta}^{(m-1)}}$ evaluated at the current "estimate" $\underline{\beta}^{(m-1)}$

→ use the expected value of the Hessian, which is the negative of FIM so the iteration becomes

$$\underline{\beta}^{(m)} = \underline{\beta}^{(m-1)} + \left[V^{(m-1)}\right]^{-1} \left.\frac{\partial l}{\partial \underline{\beta}}\right|_{\underline{\beta}=\underline{\beta}^{(m-1)}}$$

"Fisher's scoring"

$V^{(m-1)}$ evaluated at $\underline{\beta}^{(m-1)}$

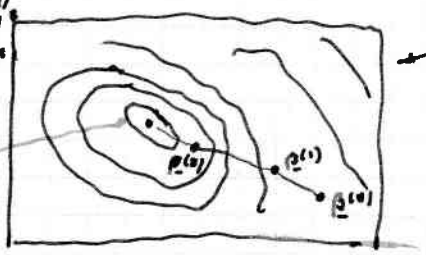
The estimation using Fisher scoring can be shown to be the iterative reweighted least squares. Starting from $\underline{\beta}^{(0)}$ (initial estimate)

Compute $z^{(m)} = X\beta^{(m-1)} + (\underline{y} - \underline{\mu})\frac{\partial \eta}{\partial \mu}$

and update $\beta^{(m)} = (X^T W X)^{-1} X^T W z^{(m)}$ for $m = 1, 2, \ldots$

The iteration continues until $\beta^{(m)}$ and $\beta^{(m-1)}$ are within a specified distance: $\|\underline{\beta}^{(m)} - \underline{\beta}^{(m-1)}\|$ less than, say ε.

Example. Logistic



← contour of $l(\underline{\beta})$

$\underline{\beta}^{(0)}, \underline{\beta}^{(1)}, \underline{\beta}^{(2)}, \ldots$

3.4 Assessing the fit of a model

The adequacy of a model is defined relative to a maximal model (it has the same number of parameters as observations).

Maximal model $\beta_{max} = (\beta_1, \beta_2, \ldots, \beta_n)^T$ compared with another model $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^T$ with $p < n$.

We define the deviance $D = -2 \log \Lambda(\underline{y}) \sim \chi^2_{n-p}$ (Wilks' thm.)

where $\Lambda(\underline{y}) = \frac{L(\hat{\beta}, \underline{y})}{L(\hat{\beta}_{max}, \underline{y})}$ — ratio of likelihoods; χ^2 — chi-square distribution

In general, if we compare two models $q < p$

M_0 (model 0): $\beta_0 = (\beta_1, .. \beta_q)^T$ $D_0 \sim \chi^2_{n-q}$

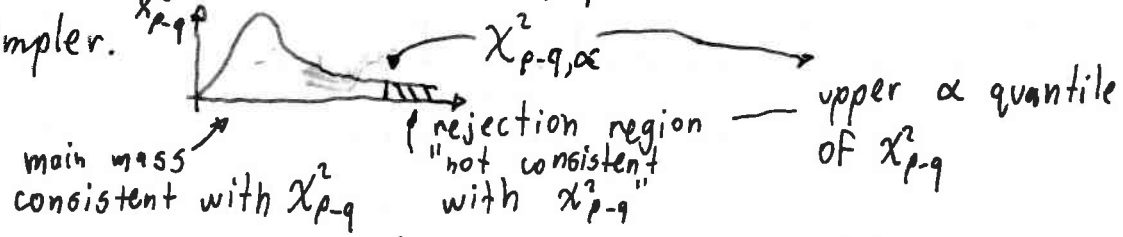
M_1 (model 1): $\beta_1 = (\beta_1, \ldots, \beta_p)^T$ $D_1 \sim \chi^2_{n-p}$

it can be shown that $D_0 - D_1 \sim \chi^2_{p-q}$

We have H_0: data consistent with model M_0 (param β_0)

H_1: " " " " M_1 (param β_1)

if $D_0 - D_1$ is consistent with χ^2_{p-q}, we'd choose M_0 because it is simpler.



if $D_0 - D_1$ is in the rejection region, we would reject H_0 (M_0) in favour of H_1 (M_1) because M_1 provides a better description of the data.

Ex. Poisson 311023. test H_0: GLM vs H_1: maximal → do not reject H_0 } we choose GLM
then H_0: null vs H_1: GLM → reject H_0 }

3.5 Inspecting and checking models

Residuals are simple for linear models (a là SMI), they are the difference between observed and fitted $e_i = y_i - \hat{\mu}_i$.

In GLMs, the variance of observations is not constant, so we can rescale the residual to obtain Pearson residual

$$e_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

Here $V(\hat{\mu}_i)$ is the variance of Y_i in terms of fitted values $\hat{\mu}_i$.

The deviance residuals are

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

where d_i is the deviance term for the i-th observation.
The Anscombe residuals have distributional assumptions so that we can compare with normal quantiles. As example, for Poisson distribution $e_i^A = \frac{3(y_i^{2/3} - \hat{\mu}_i^{2/3})}{2\hat{\mu}_i^{1/6}}$.

4 Binary data

4.1 Binary response

The response variable is binary

$$Y_i \sim \text{Bin}(r_i, \pi_i) \quad \text{for } i = 1, \dots, n$$

and link $g(\pi_i) = \underline{x}_i\underline{\beta}$ so that $\pi_i = g^{-1}(\underline{x}_i\underline{\beta})$
and parameters $\underline{\beta} = (\beta_1, \dots, \beta_p)^T$ and $\underline{x}_i = (x_{i1}, \dots x_{ip})$.

We have likelihood

$$L(\underline{\pi}; \underline{y}) = \prod_{i=1}^{n} \binom{r_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{r_i - y_i} \quad \text{and}$$

log-likelihood

$$\ell(\underline{\pi}; \underline{y}) = \sum_{i=1}^{n} \log\binom{r_i}{y_i} + \sum_{i=1}^{n} y_i \log \pi_i + \sum_{i=1}^{n} (r_i - y_i)\log(1-\pi_i).$$

221128 0900

Rat data set

Doses of poison (x) and delivery method (w) administered to a number of rats (r), and survival numbers were registered (y). We have $n = 12$ observations and will try several models.

p

Model 1 $\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha_j + \beta_j x_k$ 4 $(\alpha_1, \alpha_2, \beta_1, \beta_2) = \underline{\theta}$

j indexes method
k indexes dose

2 methods < 2

6 doses $<$ 1.3, 1.6, 2, 2.5, 3, 3.5

$\sim w + x : w$ R

Model 2 $\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha_j + \beta x_k$ 3 $(\alpha_1, \alpha_2, \beta) = \underline{\theta}$

$\sim w + x$ R

Model 3 $\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha + \beta x_k$ 2 $(\alpha, \beta) = \underline{\theta}$

$\sim x$ R

X matrix

M1

$$\begin{pmatrix} 1 & 0 & 1.3 & 0 \\ 1 & 0 & 1.6 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 0 & 2.5 & 0 \\ 1 & 0 & 3 & 0 \\ 1 & 0 & 3.5 & 0 \\ 0 & 1 & 0 & 1.3 \\ 0 & 1 & 0 & 1.6 \\ 0 & 1 & 0 & 2 \\ 0 & 1 & 0 & 2.5 \\ 0 & 1 & 0 & 3 \\ 0 & 1 & 0 & 3.5 \end{pmatrix}$$

M2

$$\begin{pmatrix} 1 & 0 & 1.3 \\ 1 & 0 & 1.6 \\ 1 & 0 & 2 \\ 1 & 0 & 2.5 \\ 1 & 0 & 3 \\ 1 & 0 & 3.5 \\ 0 & 1 & 1.3 \\ 0 & 1 & 1.6 \\ 0 & 1 & 2 \\ 0 & 1 & 2.5 \\ 0 & 1 & 3 \\ 0 & 1 & 3.5 \end{pmatrix}$$

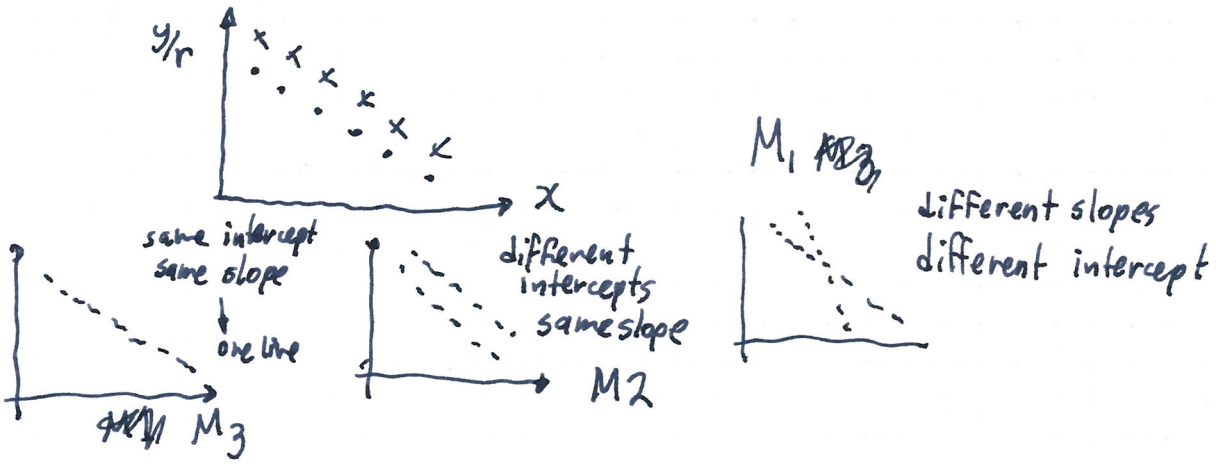
M3

$$\begin{pmatrix} 1 & 1.3 \\ 1 & 1.6 \\ 1 & 2 \\ 1 & 2.5 \\ 1 & 3 \\ 1 & 3.5 \\ 1 & 1.3 \\ 1 & 1.6 \\ 1 & 2 \\ 1 & 2.5 \\ 1 & 3 \\ 1 & 3.5 \end{pmatrix}$$

} method 1

} method 2

Two covariates < Dose x continuous
Method w categorical



Model	log-likelihood	p		Value	Decision
M3	-19.498	2			
M2	-17.690	3	M_2 vs M_1	0.503	Not reject H_0
M1	-17.467	4			
Maximal	-15.606	12	M_3 vs M_2	0.057	Reject H_0
			M_1 vs maximal	0.881	Not reject H_0

We select **M2**

Parameters (from R)

M1

int → $\hat{\alpha}_1$

w2 → $\hat{\alpha}_2 = \text{int} + \text{w2}$

W1:x → $\hat{\beta}_1$

W2:x → $\hat{\beta}_2$

M2

int → $\hat{\alpha}_1$

w_2 → $\hat{\alpha}_2 = \text{int} + w_2$

x → $\hat{\beta}$

M3

int → $\hat{\alpha}$

x → $\hat{\beta}$

221123 1300

The maximal model has n parameters π_i, which are estimated as observed proportions $\hat{\pi}_{i,max} = y_i/r_i$ ← observed proportions

and we have

$$l(\hat{\underline{\pi}}_{max}; \underline{y}) = \sum_{i=1}^{n} \log\binom{r_i}{y_i} + \sum_{i=1}^{n} y_i \log(\hat{\pi}_{i,max}) + \sum_{i=1}^{n} (r_i - y_i) \log(1 - \hat{\pi}_{i,max})$$

to compare the fit achieved with, say, GLM vs the maximal model, we compute the deviance

$$D = -2\left(l(\hat{\underline{\pi}}; \underline{y}) - l(\hat{\underline{\pi}}_{max}; \underline{y})\right)$$

$$= 2\sum_{i=1}^{n} \underbrace{\left[y_i \log\left(\frac{y_i}{r_i \hat{\pi}_i}\right) + (r_i - y_i) \log\left(\frac{r_i - y_i}{r_i(1-\hat{\pi}_i)}\right)\right]}_{\text{deviance contribution of observation } i \ (d_i)} = \sum_{i=1}^{n} d_i$$

By Wilks' theorem, $D \sim \chi^2_{n-p}$ (for large n).

Using the deviance, we perform the test

H_0: maximal model and model M (GLM) are similar in fit

H_1: models are not similar in fit

If pvalue $> \alpha$, we don't reject Ho and conclude that M and the maximal model are similar in fit. We choose M because it has fewer parameters (more parsimonious).

If pvalue $< \alpha$, we reject Ho. The proposed model is worse in fit, we should not choose it.

Usually $\alpha = 0.05$.

Example 4.1

We model Y_{jk} = number of rats surviving using a binomial distribution with logit link. $Y_{jk} \sim Bin(r_{jk}, \pi_{jk})$

Doses x_k: 1.3, 1.6, 2, 2.5, 3, 3.5

Method: solid, liquid

r_{jk}: rats given dose x_k and method j

j indexes method $j=1,2$

k indexes dose $k=1,2,\ldots,6$

Model 1: $\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha_j + \beta_j x_k$ $\quad p=4$ $\quad \underline{\theta} = (\alpha_1, \alpha_2, \beta_1, \beta_2)$

Model 2: $\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha_j + \beta x_k$ $\quad p=3$ $\quad \underline{\theta} = (\alpha_1, \alpha_2, \beta)$

Model 3: $\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha + \beta x_k$ $\quad p=2$ $\quad \underline{\theta} = (\alpha, \beta)$

Simpler models

(For the analysis, wait till 1500-1700)

Let us look at the model M3 and interpret results.

$$\log\left(\underbrace{\frac{\pi_{jk}}{1-\pi_{jk}}}_{\text{odds}}\right) = \alpha + \beta x_k \Leftrightarrow \pi_{jk} = \frac{e^{\alpha+\beta x_k}}{1+e^{\alpha+\beta x_k}} \equiv \frac{1}{1+e^{-(\alpha+\beta x_k)}}$$

β:

$$OR = \frac{\text{odds when } x_k = c+1}{\text{odds when } x_k = c} = \frac{e^{\alpha+\beta(c+1)}}{e^{\alpha+\beta c}} = e^{\beta}$$

Odds ratio Interpretation. e^{β} is the relative change in the odds of survival when the dose is increased by one unit, holding everything else fixed.

If $\beta > 0$ then $e^{\beta} > 1$ odds increase

$\beta = 0$ $\quad e^{\beta} = 1$ odds remain unchanged

$\beta < 0$ $\quad e^{\beta} < 1$ odds decrease

"For every increase of one unit in poison, the survival odds increase/decrease by a factor of e^{β}"

Because $\frac{d}{dx_k}\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \beta$ "rate of change of log odds as function of poison"

α: The intercept has an easy interpretation in probabilities, assumming dose of venom $x_k = 0$.

$$\pi_{jk} = \frac{\exp(\alpha)}{1+\exp(\alpha)}$$ ← Probability of survival for a rat given placebo (zero dose)

4.3 Pearson's goodness of fit statistic

This statistic is an alternative measure of goodness of fit that is asymptotically equivalent to the deviance. This statistic is defined by

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$ recall that $\hat{\mu}_i = r_i \hat{\pi}_i$, $V(\hat{\mu}_i) = \hat{\mu}_i\left(1 - \frac{\hat{\mu}_i}{r_i}\right)$

$$= \sum_{i=1}^{n} \frac{(y_i - r_i\hat{\pi}_i)^2}{r_i\hat{\pi}_i(1-\hat{\pi}_i)} \cdot 1 \qquad 1 = (1-\hat{\pi}_i) + \hat{\pi}_i$$

$$= \sum_{i=1}^{n} \left(\frac{(y_i - r_i\hat{\pi}_i)^2}{r_i\hat{\pi}_i}\right) + \sum_{i=1}^{n} \frac{(r_i - y_i - r_i(1-\hat{\pi}_i))^2}{r_i(1-\hat{\pi}_i)}$$

observed, expected

that is the statistic has the form $X^2 = \sum \frac{(O-E)^2}{E}$

221123 (1500) Rat data set.

Doses of poison (x) (mg) were given to rats using two delivery methods (w). There were six doses (1.3, 1.6, 2, 2.5, 3, 3.5) and deliveries (solid, liquid) coded as 1, 2. There were $n = 12$ batches rats, and for each batch of r_{jk} rats, the number of surviving individuals y_{jk} was recorded.

The basic model is $Y_{jk} \sim \text{Bin}(r_{jk}, \pi_{jk})$, and we will explore different forms of including the effect of poison with delivery method, using logistic link.

Model 1 $\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha_j + \beta_j x_k$ — $p = 4$ — $\underline{\theta} = (\alpha_1, \alpha_2, \beta_1, \beta_2)$

j indexes delivery $j = 1, 2$; x_k: doses

$\sim w + x : w$ R

Model 2 $\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha_j + \beta x_k$ — $p = 3$ — $\underline{\theta} = (\alpha_1, \alpha_2, \beta)$

$\sim w + x$ R

Model 3 $\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \alpha + \beta x_k$ — $p = 2$ — $\underline{\theta} = (\alpha, \beta)$

$\sim x$ R

Design-model matrix X

M1

1	0	1.3	0
1	0	1.6	0
1	0	2	0
1	0	2.5	0
1	0	3	0
1	0	3.5	0
0	1	0	1.3
0	1	0	1.6
0	1	0	2
0	1	0	2.5
0	1	0	3
0	1	0	3.5

M2

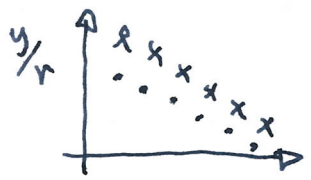
1	0	1.3
1	0	1.6
1	0	2
1	0	2.5
1	0	3
1	0	3.5
0	1	1.3
0	1	1.6
0	1	2
0	1	2.5
0	1	3
0	1	3.5

M3

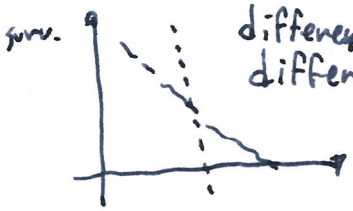
1	1.3
1	1.6
1	2
1	2.5
1	3
1	3.5
1	1.3
1	1.6
1	2
1	2.5
1	3
1	3.5

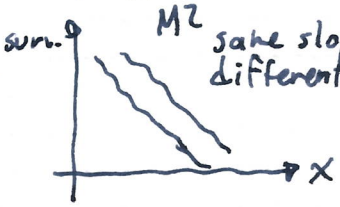
} Delivery solid (coded as 1)

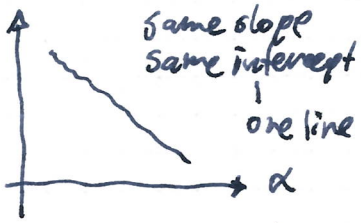
} Delivery liquid (2)



x continuous
w cathegorical







	log-likelihood	p
M3	-19.498	2
M2	-17.690	3
M1	-17.467	4
Maximal	-15.606	12

H_0: M1 vs H_i maximal — $D = 3.72$ — $PV = 0.98$ — Do not reject H_0, choose M1

H_0: M_2 vs H_1: M_1 — $D = 0.44$ — $PV = 0.503$ — Do not reject H_0, choose M2

H_0: M_3 vs H_1: M_2 — $D = 3.61$ — $PV = 0.05$ — Reject H_0, Choose M2

Final model [M2]

Parameters (from R)

M1

Int	→	$\hat{\alpha}_1$
w2		$\hat{\alpha}_2 = \text{Int} + w2$
w1:x	→	$\hat{\beta}_1$
w2:x	→	$\hat{\beta}_2$

M2

Int	→	$\hat{\alpha}_1$
x	→	$\hat{\beta}$
w2		$\hat{\alpha}_2 = \text{Int} + w2$

M3

Int	→	$\hat{\alpha}$
x	→	$\hat{\beta}$

281123

Pearson's statistic can be seen as an approximation to the deviance D. For large samples, the distribution of Pearson's $X^2 \sim \chi^2_{n-p}$.

4.4 Overdispersion

If there is more variability in the data than what we would expect under the assumed mode, one approach is to assume $\text{Var}(Y_i) = \psi V(\mu_i)$.

↑ unknown dispersion parameter $\psi > 0$

The mle of ψ is commonly

$$\psi = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

5 Count data / Poisson regression

5.1 Setup For Poisson regression

We have count data (non-negative integer) $Y_i \sim \text{Poisson}(\mu_i)$ for $i = 1, \ldots, n$. We have $g(\mu_i) = x_i^T \beta$.

We have the likelihood

$$L(\underline{\mu}; \underline{y}) = \prod_{i=1}^{n} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}$$

and log-likelihood

$$l(\underline{\mu}; \underline{y}) = \sum_{i=1}^{n} y_i \log \mu_i - \sum_{i=1}^{n} \mu_i - \sum_{i=1}^{n} \log y_i!$$

The above is the maximal model, i.e. one parameter (μ_i) for every observation y_i. The mles are $\hat{\mu}_1 = y_1, \hat{\mu}_2 = y_2, \ldots, \hat{\mu}_n = y_n$.

Under the link $g(\mu_i) = x_i^T \beta$, then the data is modelled with p parameters, namely the dimensionality of $\underline{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$. The mles of β_i are computed using the method we looked at in Ch. 3. Using $\hat{\beta}$, we compute $\hat{\underline{\mu}}$, the vector of mle estimates of $\underline{\mu}$.
The deviance for comparing glm fit against the maximal model is

$$D = -2\left(l(\hat{\underline{\mu}}; \underline{y}) - l(\hat{\underline{\mu}}_{max}; \underline{y}) \right)$$

Wilk's theorem for testing fit of glm against maximal fit.

The (log) likelihood for the maximal model is

$$l(\hat{\underline{\mu}}_{max}; \underline{y}) = l(\underline{y}; \underline{y}) = \sum_{i=1}^{n} y_i \log y_i - \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log y_i!$$

parameter data; param. data; Care when $y_i = 0$ use $P_r(0) = e^{-\mu_i}$, $\log P_r(0) = -\mu_i$

(1400) The deviance is

$$D = 2 \sum_{i=1}^{n} \underbrace{\left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i + \hat{\mu}_i \right)}_{d_i} \sim \chi^2_{n-p}.$$

Wilks' thm. for large n.

d_i: deviance for the ith observation

Example 5.1 $Y_i \sim \text{Poisson}(\mu_i)$ $\quad \eta_i = \log(\mu_i) = \beta_0 + \beta_1 x_i$; $n = 9$.

Fitted Poisson model

$$\hat{\mu}_i = e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}$$

Interpretation of parameters

$$\log(\mu_x) = \beta_0 + \beta_1 x$$
$$\log(\mu_{x+1}) = \beta_0 + \beta_1 (x+1)$$

look at the difference

$$\log(\mu_{x+1}) - \log(\mu_x) = \beta_1$$
$$\log(\mu_{x+1}/\mu_x) = \beta_1$$
$$\frac{\mu_{x+1}}{\mu_x} = e^{\beta_1}$$

Each increase of one unit of x is associated with a change in (predicted) y by a factor of e^{β_1}.

If $\beta_1 > 0 \Rightarrow e^{\beta_1} > 1$ increase

$\beta_1 < 0 \Rightarrow e^{\beta_1} < 1$ decrease

5.2 Models for contingency tables

The contingency table created by cross-classification of counts considering explanatory variables.

Two explanatory variables Ex. 5.2 < type of cancer / site

variable A → J categories

B → K categories

Some notation:

Y_{jk} is the frequency observed in cell (j,k)

$$Y_{j\cdot} = \sum_{k=1}^{K} Y_{jk} \leftarrow \text{total for row } j$$

$$Y_{\cdot k} = \sum_{j=1}^{J} Y_{jk} \leftarrow \text{total for column } k$$

$$Y_{\cdot\cdot} = \sum_{j=1}^{J}\sum_{k=1}^{K} Y_{jk} = \sum_{k=1}^{K}\sum_{j=1}^{J} Y_{jk} = \sum_{j,k} Y_{jk} = N \leftarrow \text{table total}$$

The simplest model is $Y_{jk} \sim \text{Poisson}(\mu_{jk})$ independently.

μ_{jk} → Parameter for cell (j,k). There are $J \cdot K$ cells and this is the number of μ_{jk}

$$P(\underline{Y} = \underline{y}) = \prod_{j=1}^{J}\prod_{k=1}^{K} P(Y_{jk} = y_{jk})$$

$$= \prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\mu_{jk}^{y_{jk}} e^{-\mu_{jk}}}{y_{jk}!}.$$

031223

This week - last Wednesday 0900-1100 Lab Math 302

Next week (week 12)

Lab Monday (last lab) (all) 1100-1300 Bancroft 1.15a

Lec Tuesday (final lecture) 1300-1500 Bancroft 2.40

No lab Wednesday

Midterm test Wednesday 0900-1000 online

Exam preparation session Wed. 061223 1400-1600 Fogg LT

In practice, there are constraints on the Ys such as that the overall total N is fixed. We know that $N \sim \text{Poisson}(\mu_{\cdot\cdot})$ with $\mu_{\cdot\cdot} = \sum_{j,k} \mu_{jk}$ so that

$$\Pr(N=n) = \frac{\mu_{\cdot\cdot}^n e^{-\mu_{\cdot\cdot}}}{n!} \qquad (\text{with mle } \hat{\mu}_{\cdot\cdot} = n)$$

We are going to condition the probability $\Pr(\underline{Y}=\underline{y})$:

$$\Pr(\underline{Y}=\underline{y}) = \underbrace{\prod_{j=1}^{J}\prod_{k=1}^{K} \frac{\mu_{jk}^{y_{jk}} e^{-\mu_{jk}}}{y_{jk}!}}_{\text{joint}} \cdot \underbrace{\frac{\mu_{\cdot\cdot}^n e^{-\mu_{\cdot\cdot}}}{n!}}_{\text{marginal}} \Big/ \underbrace{\frac{\mu_{\cdot\cdot}^n e^{-\mu_{\cdot\cdot}}}{n!}}_{\text{conditional}}$$

$$= \underbrace{\frac{n!}{\prod_{j=1}^{J}\prod_{k=1}^{K} y_{jk}!} \prod_{j=1}^{J}\prod_{k=1}^{K} \left(\frac{\mu_{jk}}{\mu_{\cdot\cdot}}\right)^{y_{jk}}}_{\text{Multinomial}} \cdot \underbrace{\frac{\mu_{\cdot\cdot}^n e^{-\mu_{\cdot\cdot}}}{n!}}_{\text{Poisson}}$$

let $\frac{\mu_{jk}}{\mu_{\cdot\cdot}} = \theta_{jk}$ — Probability that an individual is in cell j,k.

$$= \frac{n!}{\prod_{j=1}^{J}\prod_{k=1}^{K} y_{jk}!} \prod_{j=1}^{J}\prod_{k=1}^{K} \theta_{jk}^{y_{jk}} \cdot \frac{\mu_{\cdot\cdot}^n e^{-\mu_{\cdot\cdot}}}{n!}$$

Question. Compute $\theta_{\cdot\cdot} = 1$. — The probabilities of MN add to one.

mle maximal $\hat{\theta}_{jk} = \frac{y_{jk}}{y_{\cdot\cdot}}$

Example. Maximal likelihood for contingency data of lab.

5.3 Log-linear models

For contingency tables, the hypothesis can be formulated as multiplicative models for the expected cell frequencies. This suggests using logarithm as the natural link function between expected cell frequencies and linear combinations of parameters.

$$\Pr(\underline{Y}=\underline{y}) = \frac{\Pr(\underline{Y}=\underline{y})}{\Pr(N=n)} \Pr(N=n) = \Pr(\underline{Y}=\underline{y} \mid N=n) \Pr(N=n)$$

Consider the multinomial part of the model. The independence hypothesis is $\theta_{jk} = \theta_{j\cdot}\,\theta_{\cdot k}$ (θ_{jk}: joint; $\theta_{j\cdot}$: marginal row; $\theta_{\cdot k}$: column)

We have new set of parameters $\theta_{1\cdot}, \theta_{2\cdot}, \ldots, \theta_{J\cdot}$ row; $\theta_{\cdot 1}, \theta_{\cdot 2}, \ldots, \theta_{\cdot K}$ column — satisfy $\sum_{j=1}^{J} \theta_{j\cdot} = 1$, $\sum_{k=1}^{K} \theta_{\cdot k} = 1$ (constraints otherwise models are overparameterized.)

Under the independence hypothesis, the expected frequency of cell j,k is

$$E(Y_{jk} \mid N=n) = n\theta_{jk} = n\theta_{j\cdot}\theta_{\cdot k} \quad \text{and we can write}$$

$$\eta_{jk} = \log(E(Y_{jk} \mid N=n)) = \log(n) + \log(\theta_{j\cdot}) + \log(\theta_{\cdot k})$$

$$= \underset{\text{overall effect}}{\mu} + \underset{\text{row effect}}{\alpha_j} + \underset{\text{effect of column}}{\beta_k}$$

The maximal model $E(Y_{jk} \mid N=n) = n\theta_{jk}$ can be written as

$$\eta_{jk} = \log(E(Y_{jk} \mid N=n)) = \mu + \alpha_j + \beta_k + \gamma_{jk}$$

We need to parameterize carefully

$\sum_{j=1}^{J} \alpha_j = 0$ and $\sum_{k=1}^{K} \beta_k = 0$ for the independence model

Another example of parameterization is to set

$\alpha_1 = 0$ and $\beta_1 = 0$ ← This is done in R.

together with intercept, we have $1 + (J-1) + (K-1) = J+K-1$ params.

For the maximal model there are additional constraints

$\sum_{j=1}^{J} \gamma_{jk} = 0$ and $\sum_{k=1}^{K} \gamma_{jk} = 0$ so that we have

$$\underbrace{1 + (J-1) + (K-1)}_{\text{Indep. model } J+K-1 \text{ params}} + (J-1)(K-1) = JK \text{ parameters}$$

5.4 Fitting the models

The multinomial likelihood (log) is

$$\ell(\underline{\theta};\underline{y}) = \log(n!) + \sum_{j,k} y_{jk}\log(\theta_{jk}) - \sum_{j,k}\log(y_{jk}!).$$

This is the maximal model, it has constraint $\sum_{j,k}\theta_{jk} = 1$.

We add the constraint using Lagrange multiplier

$$t(\underline{\theta},\xi;y) = \underbrace{\ell(\underline{\theta};\underline{y})}_{\text{likelihood objective to maximize}} - \xi\underbrace{\left(\sum_{j,k}\theta_{jk} - 1\right)}_{\text{constraint}}$$

t: Lagrangian; ξ: Lagrange multiplier

Estimates are $\hat{\theta}_{jk} = \frac{y_{jk}}{n}$ ← Maximal model

To maximize, compute derivatives

$$\left\{\frac{\partial t}{\partial \theta_{jk}}\right\}_{j,k} \text{ and } \frac{\partial t}{\partial \xi}$$

in total JK

and form a system of simultaneous equations

$$\left\{\frac{\partial t}{\partial \theta_{jk}} \stackrel{!}{=} 0\right\}, \frac{\partial t}{\partial \xi} \stackrel{!}{=} 0$$

The solution set of the simultaneous system are candidates to maximize/minimize / extreme points of the objective function.

Example. Independence model, 2x2 table

$$L(\underline{\theta};\underline{y}) = \frac{n!}{\prod_{jk} y_{jk}!}\theta_{11}^{y_{11}}\theta_{12}^{y_{12}}\theta_{21}^{y_{21}}\theta_{22}^{y_{22}} \qquad \text{indep } \theta_{jk} = \theta_{j\cdot}\theta_{\cdot k}$$

$$= \frac{n!}{\prod_{j,k} y_{jk}!}\theta_{1\cdot}^{y_{11}}\theta_{\cdot 1}^{y_{11}}\theta_{1\cdot}^{y_{12}}\theta_{\cdot 2}^{y_{12}}\theta_{2\cdot}^{y_{21}}\theta_{\cdot 1}^{y_{21}}\theta_{2\cdot}^{y_{22}}\theta_{\cdot 2}^{y_{22}}$$

$$= \frac{n!}{\prod_{j,k} y_{jk}!}\theta_{1\cdot}^{y_{1\cdot}}\theta_{2\cdot}^{y_{2\cdot}}\theta_{\cdot 1}^{y_{\cdot 1}}\theta_{\cdot 2}^{y_{\cdot 2}}$$

11/22/23 Two constraints $\theta_{1\cdot} + \theta_{2\cdot} = 1$ and $\theta_{\cdot 1} + \theta_{\cdot 2} = 1$

$$t(\underline{\theta},\xi_1,\xi_2;\underline{y}) = \ell(\underline{\theta};\underline{y}) - \xi_1(\theta_{1\cdot}+\theta_{2\cdot}-1) - \xi_2(\theta_{\cdot 1}+\theta_{\cdot 2}-1)$$

$$\frac{\partial t}{\partial \theta_{1\cdot}}, \frac{\partial t}{\partial \theta_{2\cdot}}, \frac{\partial t}{\partial \theta_{\cdot 1}}, \frac{\partial t}{\partial \theta_{\cdot 2}}, \frac{\partial t}{\partial \xi_1}, \frac{\partial t}{\partial \xi_2}$$

$$\hat{\theta}_{1.} = \frac{y_{1.}}{y_{1.}+y_{2.}}, \quad \hat{\theta}_{2.} = \frac{y_{2.}}{y_{1.}+y_{2.}}, \quad \hat{\theta}_{.1} = \frac{y_{.1}}{y_{.1}+y_{.2}}, \quad \hat{\theta}_{.2} = \frac{y_{.2}}{y_{.1}+y_{.2}}$$

[21223]

For estimating a general independence model of a $J \times K$ contingency table, we form the lagrangian

$$t(\underline{\theta}, \xi_1, \xi_2; \underline{y}) = \underbrace{\log(n!) + \sum_{j=1}^{J} y_{j.} \log \theta_{j.} + \sum_{k=1}^{K} y_{.k} \log(\theta_{.k}) - \sum_{j,k} \log(y_{jk}!)}_{\text{likelihood } \ell(\underline{\theta}; \underline{y}) \text{ under indep.}} - \xi_1 \Big(\sum_{j=1}^{J} \theta_{j.} - 1 \Big) - \xi_2 \Big(\sum_{k=1}^{K} \theta_{.k} - 1 \Big)$$

constraint marginal by row; constraint by column

Estimates are $\hat{\theta}_{j.} = \frac{y_{j.}}{n}$ and $\hat{\theta}_{.k} = \frac{y_{.k}}{n}$

Recall $n = \sum_{j,k} y_{jk} = y_{..} = \sum_{j} y_{j.} = \sum_{k} y_{.k}$

The deviance is

$$D = 2 \sum_{j,k} y_{jk} \log\left(\frac{y_{jk}}{e_{jk}}\right) \quad \text{with} \quad e_{jk} = \frac{y_{j.} y_{.k}}{n}$$

← Expected entry of table under indep.

that is to be compared against χ^2 with $(J-1)(K-1)$ degrees of freedom.

We can also use Pearson's statistic

$$X^2 = \sum_{j,k} \frac{(y_{jk} - e_{jk})^2}{e_{jk}},$$ also compared with the same χ^2 dist'n.

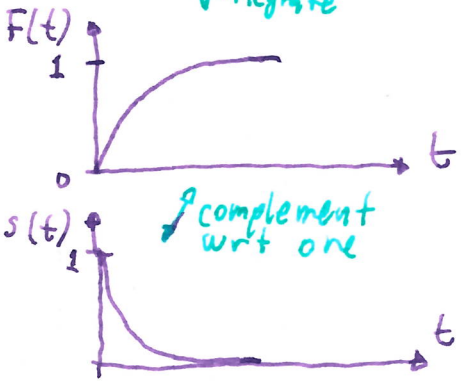
6 - Survival data

6.1) We are interested in modelling values from a positive random variable T. These values can be lifetimes, number of cycles before failure.

We can characterize survival data with descriptors at a population level, the simplest of which is the

probability density function of T which is $f(t)$.

We can also use the cumulative distribution $F(t) = Pr(T < t)$.

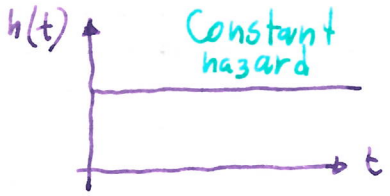


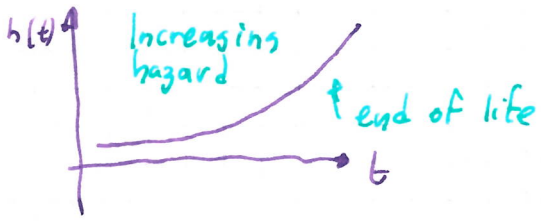
We can study survival times using the $S(t) = Pr(T > t) = 1 - F(t)$. This is known as the survival function.

The hazard function is the conditional probability of T given survival up to time t.

instantaneous failure

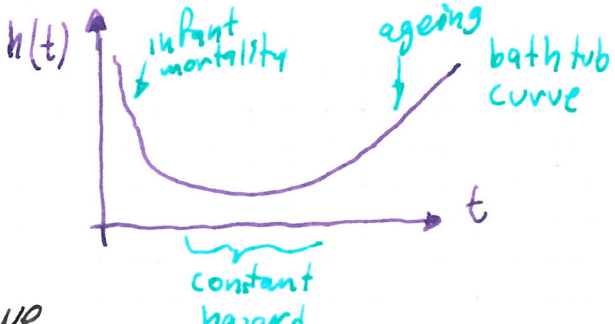
$$h(t) = \lim_{\Delta t \to 0} \frac{Pr(t < T < t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{S(t)}$$





$F(t)$ ← Proportion that has died at time t.

$S(t)$ ← Proportion that has survived at time t.



For a given distribution $f(t)$, we can derive survival, hazard functions. The task of the modeller is to use a distribution flexible enough to represent the data. For the exponential distribution we have <u>constant hazard</u>.

We have other descriptors, like the integrated hazard function $H(t) = \int_0^t h(u)\,du$.

From each of $f(t), F(t), S(t), h(t), H(t)$ you can recover the others.

Example. Exponential distribution

$f(t) = \lambda e^{-\lambda t}$ for $t \geq 0$ density

$F(t) = 1 - e^{-\lambda t}$ for $t \geq 0$ cdf

$S(t) = e^{-\lambda t}$ for $t \geq 0$ survival

$h(t) = \lambda$ for $t \geq 0$ hazard (constant)

$H(t) = \lambda t$ for $t \geq 0$ cum. hazard

6.2 Exponential regression

We consider data assumed from exponential, that is $T_i \sim Exp(\lambda_i)$ for $i = 1, .., n$ all independent.

Maximal The maximal model has all λ_i as parameters with likelihood $L(\underline{\lambda}; \underline{t}) = \prod_{i=1}^{n} \lambda_i e^{-\lambda_i t_i}$. The ml estimates are

$$\hat{\lambda}_1 = 1/t_1, \hat{\lambda}_2 = 1/t_2, \ldots, \hat{\lambda}_n = 1/t_n.$$

$E(T_i) = 1/\lambda_i$ mles are
$E(1/T_i) \neq \lambda_i$ not unbiased in general

Null On the other extreme, the null model has a single parameter, i.e. $\lambda = \lambda_1 = \lambda_2 = \ldots = \lambda_n$, that has maximum likelihood estimate $\hat{\lambda} = 1/\bar{t} = n / \sum_{i=1}^{n} t_i = \left(\frac{1}{n}\sum_{i=1}^{n} t_i\right)^{-1}$

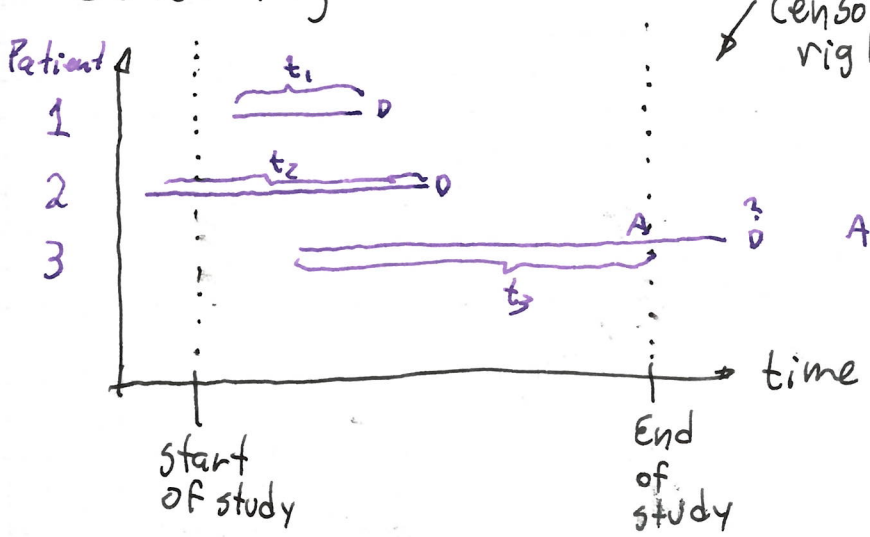
Glm Assume that we have covariates available for prediction and we build

$$g(\lambda_i) = x_i^T \beta$$

g: link function; $x_i^T \beta$: linear predictor; β: vector of parameters (size p)

The maximum likelihood estimates $\hat{\beta}$ are obtained using the numerical technique seen in Ch. 3, and we use the deviance D to compare the glm fit against the maximal model as usual.

6.3 Censoring



Alive at end of study t_3 is censored: the patient alive with age t_3.

Survival times are censored when the individual is still alive at the end of the study and we only know the age when we stopped observing.
Together with each time t_i, we have a variable

$$\delta_i = \begin{cases} 1 & \text{if patient died at time } t_i \\ 0 & \text{if patient alive at time } t_i \end{cases}$$

The likelihood is composed of two parts: the usual density for patients that died, and the survival for those still alive

$$L(\underline{\lambda}; \underline{t}) = \prod_{i=1}^{n} f(t_i)^{\delta_i} \cdot S(t_i)^{1-\delta_i}.$$

which for the exponential model leads to the log-likelihood

$$l(\underline{\lambda}; \underline{t}) = \sum_{i=1}^{n} \delta_i \log(\lambda_i) - \sum_{i=1}^{n} \lambda_i t_i.$$

(t_i: times observed; δ_i: for censoring $\delta_i = 0$)

What we looked at in this course:

Likelihood $L(\underline{\theta}; \underline{y}) = Pr(\underline{Y} = \underline{y}; \underline{\theta})$ ← Discrete r.v.

($\underline{\theta}$: Parameter, $\underline{y}$: Data)

$L(\underline{\theta}; \underline{y}) = f(\underline{y}; \underline{\theta})$ ← continuous r.v.

Log-likelihood $l(\underline{\theta}; \underline{y}) = \log L(\underline{\theta}; \underline{y})$

Regular models - range of r.v. does not depend on parameters

mle

$$\frac{\partial l}{\partial \underline{\theta}} \stackrel{!}{=} 0 \Rightarrow \hat{\underline{\theta}} = \hat{\underline{\theta}}(\underline{y})$$

$\hat{\underline{\theta}}$ ← mle

mle (constrained) $\{g\} \equiv \{g_i(\underline{\theta}) = 0\}$ ← Constraints

$t = l + \sum_i \xi_i g_i$ ← Lagrangian

$$\frac{\partial t}{\partial \underline{\theta}} \stackrel{!}{=} 0, \quad \frac{\partial t}{\partial \xi_i} \stackrel{!}{=} 0 \Rightarrow \hat{\underline{\theta}} = \hat{\underline{\theta}}(\underline{y})$$

ξ_i ← Lagrange multiplier

Theorem: $\hat{\underline{\theta}} \sim N_p(\theta, V^{-1})$

$V = E\left(-\frac{\partial}{\partial \underline{\theta}} \frac{\partial^T}{\partial \underline{\theta}} l\right)$ ← Fisher information matrix (FIM)

$\hat{V} = -\frac{\partial}{\partial \underline{\theta}} \frac{\partial^T}{\partial \underline{\theta}} l \Big|_{\theta=\hat{\theta}}$ ← Observed FIM

Likelihood ratio test \ Wilks' theorem to compare L_{A} (H_0) vs L_{B} (H_1)

$$-2 \log \frac{L_A}{L_B} = -2(l_A - l_B) \sim \chi^2_s$$

$s = p_B - p_A$

Deviance D

Pearson's X^2

Th. $V = E\left(\frac{\partial l}{\partial \underline{\theta}} \frac{\partial^T l}{\partial \underline{\theta}}\right)$ for FIM

GLM

- A distribution from the exponential family ({a, b, c, d}; E(Y), V(Y), canonical link)
- A link $\mu = E(Y)$

$g(\mu) = x_i^T \beta = \eta$

Linear Predictor LP

Estimate iteratively to mle $\hat{\beta}$

Residuals ← Anscombe, Pearson, Deviance

Using Wilks' to compare models

Model	(log) lik	# params
null	l_{null}	1
M_A	l_A	p_A
M_B	l_B	p_B
maximal	l_{max}	n

start complex, & move to simple(r) by non-rejections

Many possibilities with LP

$x_i^T \beta$ ← regression style

α_j ← categorical

$\alpha_j + \beta_j x_K$ ← categorical with covariate

1 ← to null

α_i ← to maximal

⋮

In the lab { Bin, Poi, Exp, N ← Distribution; log, logit, identity, cloglog, probit ← Link

→ Computation of deviances, comparison between models,
→ interpretation of parameters, which parameterization to use, categorical and/or continuous covariates.

Survival data $f(t) \leftrightarrow F(t) \leftrightarrow S(t) \leftrightarrow h(t)$ Censored data (δ_i, t_i)