Course overview

LM (TMA4267)
Linear models, general linear model

GLM
Generalized linear models
Binary, binomial
Poisson
Gamma
Multinomial (VGLM)

LMM
Linear mixed models, linear models including random effects, dependent data
REML

GLMM
Generalized linear mixed models
Laplace approximation of marginal likelihood
REML

Introduction, binary regression

Recall
Multiple regression:
Assumptions: For $i = 1, 2, \ldots, n$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + e_i$$

where $e_i \overset{iid}{\sim} N(0, \sigma^2)$.

In matrix notation

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon}, \quad \underline{\varepsilon} \sim N(0, \sigma^2 I_n).$$

MLE of $\underline{\beta}$

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

Binary regression example: $Y_i \in \{0, 1\}$

Aim: How does a binary outcome depend on covariates of interest $x_{i1}, x_{i2}, \ldots, x_{ik}$.

Model: For $i = 1, 2, \ldots, n$

$Y_i \overset{\text{indep.}}{\sim} \text{Bernoulli}(p_i)$,

where

$$\underbrace{\ln \frac{p_i}{1-p_i}}_{= \text{logit } p_i} = \underbrace{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}}_{\text{linear predictor } \eta_i} + \underbrace{\varepsilon_i}_{\substack{\text{no error term and} \\ \text{dispersion param.}}}.$$

Note: The logit function $\text{logit } p = \ln \frac{p}{1-p}$ maps $p \in (0, 1)$ to $\mathbb{R}$.

The inverse:

$$\ln \frac{p}{1-p} = \eta$$

$$p = (1-p) e^{\eta}$$

$$(1+e^{\eta})p = e^{\eta}$$

$$p = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{1+e^{-\eta}} = \text{logit}^{-1}(\eta)$$

Thus

$$p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik})}}$$

Odds interpretation:

$$\frac{p_i}{1-p_i} = \text{Odds}(Y_i = 1).$$

The model assumes that

$$\ln \underbrace{\frac{p_i}{1-p_i}}_{\text{odds}} = \beta_0 + \beta_1 x_{i1} + \ldots$$

} additive effects on log odds of event

$$\Updownarrow$$

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_{i1}+\ldots+\beta_k x_{ik}}$$

} multiplicative effects on odds of event

One unit change in e.g. x_{i1} changes the odds by a factor (an odds ratio) of e^{β_1}.

Example: $\beta_1 = 0.1 \Rightarrow$ odds ratio is

$$e^{\beta_1} = e^{0.1} \approx 1.11$$

i.e. the odds increase by 11%.

Parameter estimation:

Maximise the likelihood (numerically)

$$L(\underline{\beta}) = P(Y_1 = y_1 \cap \dots \cap Y_n = y_n)$$

$$= \prod_{i=1}^{n} P(Y_i = y_i)$$

$$= \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1 - y_i}$$

where for $i = 1, \dots, n$

$$p_i = \text{logit}^{-1}(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})$$

Exponential family

Def: Pdf or pmf on the form

$$f(y|\theta) = \exp\left(\frac{y\theta - b(\theta)}{\phi} w + c(y, \phi, w) \right)$$

(θ: depends only on θ; ϕ: don't depend on θ)

and

$$\ln f(y|\theta) = \frac{y\theta - b(\theta)}{\phi} w + c(y, \phi, w)$$

where θ is a parameter (the natural or canonical), $b(\theta)$ is a twice

differentiable function, ϕ is a dispersion parameter, w is a known constant.

Support not dependent on θ.

Ex.: $Y \sim \text{Bernoulli}(\pi)$

Pmf

$$f(y|\pi) = \pi^y (1-\pi)^{1-y}$$

$$\ln f(y|\pi) = y \ln \pi + (1-y) \ln(1-\pi)$$

$$= y \underbrace{\left[\ln \pi - \ln(1-\pi) \right]}_{\theta} + \ln(1-\pi)$$

$$= y \underbrace{\ln \frac{\pi}{1-\pi}}_{\theta} - \underbrace{\left(-\ln(1-\pi) \right)}_{b(\theta)}$$

$\theta = \text{logit}\, \pi$

$\pi = \text{logit}^{-1} \theta$

$$= \frac{1}{1+e^{-\theta}} = \frac{e^{\theta}}{1+e^{\theta}}$$

$$1-\pi = \frac{e^{-\theta}}{1+e^{-\theta}} = \frac{1}{1+e^{\theta}}$$

$$= y\theta + \ln \frac{1}{1+e^{\theta}}$$

$$= y\theta - \ln(1+e^{\theta})$$

that is, $\theta = \ln \frac{\pi}{1-\pi}$ is the natural parameter

and $b(\theta) = \ln(1+e^{\theta})$, $c(y, w, \phi) = 0$

Theorem: If Y belongs to the exponential fam.

$$EY = b'(\theta) \qquad (*)$$

$$\mathrm{Var}\, Y = \frac{\phi}{w} b''(\theta)$$

Ex. cont: $b'(\theta) = \dfrac{e^\theta}{1+e^\theta} = \pi = EY$

$$\frac{\phi}{w} b''(\theta) = \frac{e^\theta(1+e^\theta) - e^\theta e^\theta}{(1+e^\theta)^2}$$

$$= \frac{e^\theta}{(1+e^\theta)^2} = \underbrace{\frac{e^\theta}{1+e^\theta}}_{\pi} \cdot \underbrace{\frac{1}{1+e^\theta}}_{1-\pi} = \pi(1-\pi)$$

$$= \mathrm{Var}\, Y$$

Proof of (*): Y has mgf

$$M_Y(u) = E e^{uY} = \int e^{\frac{y\theta - b(\theta)}{\phi} w + yu + c(y,\phi,w)} \, dy$$

$$= e^{\frac{b(\theta + \frac{\phi}{w}u) - b(\theta)}{\phi} w + c(y,\phi,w)}$$

$$= e^{\frac{w}{\phi}\left(b(\theta+\frac{\phi}{w}u)-b(\theta)\right)} \int e^{\frac{y(\theta+\frac{\phi}{w}u)-b(\theta+\frac{\phi}{w}u)}{\phi}w + c(y,\phi,w)} dy$$

$$= 1$$

Recall:

$$\frac{d}{du} M_X(u) = \frac{d}{du} E(e^{uX})$$

$$= E \frac{d}{du} e^{uX} = E\, X e^{uX}\Big|_{u=0} = EX$$

$$\frac{d^2}{du^2} M_X(u)\Big|_{u=0} = E(X^2)$$

We want to find $\mathrm{Var}X = E(X^2) - (EX)^2$

Def.: The cumulant generating function of a random variable X is

$$K_X(u) = \ln M_X(u).$$

Lemma:

$$K_X'(u) = \frac{M_X'(u)}{M_X(u)}\Big|_{u=0} = \frac{M_X'(0)}{M_X(0)} = \frac{EX}{1}$$

$$K''_X(u) = \frac{M''_X(u) M_X(u) - \left(M'_X(u)\right)^2}{\left(M_X(u)\right)^2}$$

$$\overset{u=0}{=} E(X^2) - (EX)^2 = Var(X)$$

$$K'''(0) = \ldots = E\left[(X-EX)^3\right] \quad \text{(exercise 1b)}$$

$$K^{(4)}(0) = \ldots \neq E\left((X-EX)^4\right) \quad \text{(see wikipedia)}$$

Y has cumulant generating

$$K_Y(u) = \frac{w}{\phi}\left(b\left(\theta + \frac{\phi}{w}u\right) - b(\theta)\right)$$

and

$$K'_Y(u) = \frac{w}{\phi}\left(b'\left(\theta + \frac{\phi}{w}u\right)\cdot\frac{\phi}{w}\right) \overset{u=0}{=} b'(\theta)$$

$$K''_Y(u) = b''\left(\theta + \frac{\phi}{w}u\right)\frac{\phi}{w} \overset{u=0}{=} \frac{\phi}{w}b''(\theta)$$

which completes the proof.

Exponential family cont.

Exponential distribution:

$$f(y|\lambda) = \lambda e^{-\lambda y}$$

$$= e^{-\lambda y + \ln\lambda}$$

$$= e^{y\theta - b(\theta)}$$

so $\theta = -\lambda$, $b(\theta) = -\ln\lambda = -\ln(-\theta) = -\ln\theta$

$$EY = b'(\theta) = -\frac{1}{-\theta}(-1) = -\frac{1}{\theta} = \frac{1}{\lambda}$$

$$\text{Var}Y = \frac{\phi}{w} b''(\theta) = \frac{1}{\theta^2} = \frac{1}{(-\lambda)^2} = \frac{1}{\lambda^2}$$

Poisson:

$$f(y|\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = e^{y\ln\lambda - \lambda - \ln y!}$$

$$= e^{y\theta - b(\theta) - c(y,\phi,w)}$$

where $\theta = \ln\lambda$, $b(\theta) = \lambda = e^{\theta}$

$$EY = b'(\theta) = e^{\theta} = \lambda$$

$$\text{Var}Y = b''(\theta) = e^{\theta} = \lambda$$

Normal:

$$f(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

$$= e^{\frac{y\mu - \mu^2/2}{\sigma^2} - \underbrace{\left(\frac{y^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2)\right)}_{c(y,\phi,w)}}$$

$$= e$$

that is $\theta = \mu$, $b(\theta) = \mu^2/2 = \theta^2/2$, $\phi = \sigma^2$

$$EY = b'(\theta) = \mu$$

$$\text{Var } Y = \frac{\phi}{w} b''(\theta) = \frac{\phi}{w} = \sigma^2$$

Generalized linear models in general

For $i = 1, 2, \ldots, n$ each observation y_i belongs to the exponential family and the mean

$$E(y_i) = \mu_i$$

is linked to a linear predictor

$$\eta_i = x_i^T \beta, \quad x_i = \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \\ x_{ik} \end{bmatrix}$$

via a strictly monotonic, twice differentiable link function g or response function $h = g^{-1}$ such that

$$\mu_i = h(\eta_i)$$

or

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

If $g(\mu_i)$ is the canonical parameter then g is the canonical link function.

Multiple linear regression

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$$

As a glm: For $i = 1, 2, \ldots, n$

$$Y_i \sim N(\mu_i, \sigma^2)$$

where

$$\mu_i = \eta_i \quad \text{(identity link function)}$$

and

$$\eta_i = x_i^T \beta$$

Parameter estimation: Maximise likelihood

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

or the log likelihood

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

Score function: Want to solve the eq.

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

Matrix calculus

Def.: If y is a scaler and $\underline{x}$ is a $p \times 1$ column vector, then

$$\frac{\partial y}{\partial \underline{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \vdots \end{bmatrix}.$$

Rule 1: Thus, if $\underline{a}$ is another $p \times 1$ column vector,

$$\frac{\partial}{\partial \underline{x}}(\underline{a}^T\underline{x}) = \frac{\partial}{\partial \underline{x}}(a_1x_1 + \ldots + a_px_p)$$

$$= \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \underline{a}.$$

Analagous to

$$\frac{d}{dx}(ax) = a. \quad (*)$$

Def.: If $\underline{Y}$ is a $q \times 1$ column vector, then

$$\underbrace{\frac{\partial}{\partial \underline{x}}}_{p \times 1} \underbrace{\underline{Y}^T}_{1 \times q} = \underbrace{\begin{bmatrix} \frac{\partial Y_1}{\partial x_1} & \frac{\partial Y_2}{\partial x_1} & \cdots \\ \frac{\partial Y}{\partial x_2} & & \\ \vdots & & \end{bmatrix}}_{p \times q}$$

See wikipedia for other conventions

Rule 2: Thus, if A is $q \times p$ matrix, then

$$\frac{\partial}{\partial \underline{x}} \left(A \underline{x} \right)^T = \frac{\partial}{\partial \underline{x}} \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p & a_{21}x_1 + a_{22}x_2 + \cdots \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & \\ \vdots & \end{bmatrix} = A^T.$$

Again, analogous to (*)

If A is a $p \times p$-matrix, then

$$\frac{\partial}{\partial \underline{x}} \left(\underline{x}^T A \underline{x} \right) \overset{p=2}{=} \frac{\partial}{\partial \underline{x}} \left(a_{11}x_1^2 + a_{12}x_1x_2 + a_{21}x_1x_2 + a_{22}x_2^2 \right)$$

$$= \begin{bmatrix} 2a_{11}x_1 + a_{12}x_2 + a_{21}x_2 \\ a_{12}x_1 + a_{21}x_1 + 2a_{22}x_2 \end{bmatrix}$$

$$= (A + A^T)\underline{x}.$$

Rule 3: If A is symmetric

$$\frac{\partial}{\partial \underline{x}}\left(\underline{x}^T A \underline{x}\right) = 2A\underline{x}.$$

Analogous to $\frac{d}{dx}(ax^2) = 2ax$

Recall: If $\underline{g}(x)$ is a $q \times 1$ column-vector-valued function and f and x are scalars, then

$$\frac{\partial}{\partial x} f(\underline{g}(x)) = \frac{\partial f}{\partial g_1}\cdot\frac{\partial g_1}{\partial x} + \dots + \frac{\partial f}{\partial g_q}\cdot\frac{\partial g_q}{\partial x} = \underbrace{\underbrace{\frac{\partial \underline{g}^T}{\partial x}}_{1\times q}\cdot\underbrace{\frac{\partial f}{\partial \underline{x}}}_{q\times 1}}_{1\times 1}$$

Rule 4 (chain rule):

$$\frac{\partial}{\partial \underline{x}} f[\underline{g}(\underline{x})] = \dots = \underbrace{\underbrace{\frac{\partial \underline{g}^T}{\partial \underline{x}}}_{p\times q}\cdot\underbrace{\frac{\partial f}{\partial \underline{g}}}_{q\times 1}}_{p\times 1}$$

Multiple regression cont.:

$$s(\underline{\beta}) = \frac{\partial l}{\partial \underline{\beta}} = \frac{\partial}{\partial \underline{\beta}}\left(-\frac{1}{2\sigma^2}(\underline{y}-X\underline{\beta})^T(\underline{y}-X\underline{\beta})\right)$$

$$= -\frac{1}{2\sigma^2}\cdot\frac{\partial}{\partial \underline{\beta}} f(\underline{g}(\underline{\beta})) \qquad \underline{g}(\underline{\beta}) = \underline{y} - X\underline{\beta}, \quad f(\underline{s}) = \underline{s}^T\underline{s}$$

$$= \frac{\partial g^T}{\partial \underline{\beta}} \cdot \frac{\partial f}{\partial \underline{g}}$$

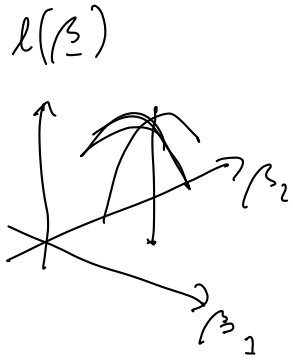
$$= -\frac{1}{2\sigma^2} \cdot \left(-X^T\right) 2(\underline{y} - X\underline{\beta}) \quad \text{(Rule 4, 3 and 2)}$$

At the MLE of $\underline{\beta}$,

$$\underline{s}(\underline{\beta}) = \underline{0}$$

$$X^T\underline{y} - X^TX\underline{\beta} = \underline{0}$$

$$\hat{\underline{\beta}} = (X^TX)^{-1}X^T\underline{y}$$



Observed Fisher information

$$H(\underline{\beta}) = -\underbrace{\frac{\partial^2}{\partial \underline{\beta}\,\partial \underline{\beta}^T}}_{p\times 1 \quad 1\times p} l(\underline{\beta}) = -\frac{\partial}{\partial \underline{\beta}} s^T(\underline{\beta})$$

$$= -\frac{\partial}{\partial \underline{\beta}} \left(\frac{1}{2\sigma^2} 2X^T(\underline{y} - X\underline{\beta}) \right)^T = \frac{X^TX}{\sigma^2} \quad \text{(Rule 2)}$$

Note how

$$H^{-1}(\underline{\beta}) = \sigma^2 (X^TX)^{-1} = \text{Var}(\hat{\underline{\beta}})$$

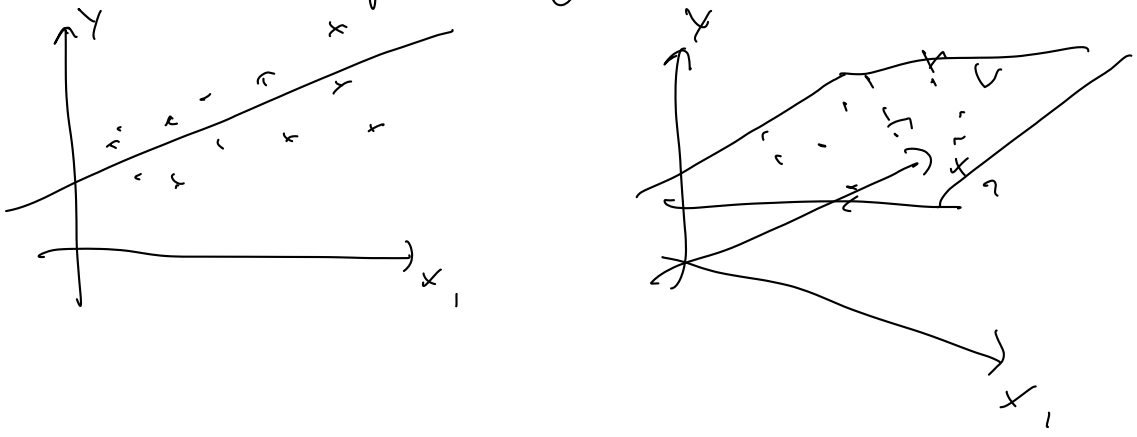
In this case the (expected) Fisher information

$$F(\underline{\beta}) \overset{\text{def}}{=} E\left(-\frac{\partial^2}{\partial \underline{\beta}\,\partial \underline{\beta}^T} l(\underline{\beta}; Y) \right) = H(\underline{\beta}),$$

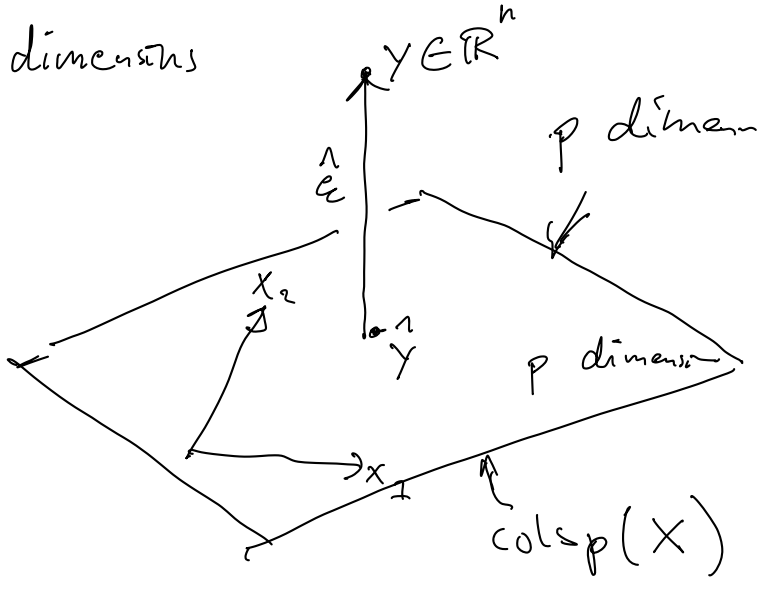
the observed Fisher information.

Distributions associated with the linear model

Data centric, applied, geometric view:



Linear algebra, vector-space view:



$$\underline{y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$$

$$= \begin{bmatrix} \underline{x}_1 & \underline{x}_2 & \underline{x}_p \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_p \end{bmatrix} + \underline{\varepsilon}$$

$$= \underline{x}_1\beta_1 + \underline{x}_2\beta_2 + \ldots + \underline{x}_p x_p + \underline{\varepsilon}$$

We seek the linear combination of the colums in X, $\hat{\underline{y}} = X\hat{\underline{\beta}} \in \text{colsp}(X)$ minimising $\|\underline{y} - \hat{\underline{y}}\|^2 = (\underline{y} - \hat{\underline{y}})^T(\underline{y} - \hat{\underline{y}}) = (\underline{y} - X\hat{\underline{\beta}})^T(\underline{y} - X\hat{\underline{\beta}})$

The residual vector $\hat{\underline{\varepsilon}} = \underline{y} - \hat{\underline{y}}$ must therefore be orthogonal to all colums in X. All the dot products $\underline{x}_1^T\hat{\underline{\varepsilon}}, \underline{x}_2^T\hat{\underline{\varepsilon}}, \ldots, \underline{x}_p^T\hat{\underline{\varepsilon}}$ are therefore all zero, that is,

$$\begin{bmatrix} \underline{x}_1^T \\ \end{bmatrix}$$

$$\begin{bmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix} \underbrace{(\underline{y} - X\hat{\underline{\beta}})}_{\hat{\underline{\varepsilon}}} = \underline{0}$$

$$X^T(\underline{y} - X\hat{\underline{\beta}}) = 0$$

$$\hat{\underline{\beta}} = (X^TX)^{-1}X^T\underline{y} \quad \text{(the MLE)}$$

The projection is given by

$$\hat{\underline{y}} = X\hat{\underline{\beta}} = \underbrace{X(X^TX)^{-1}X^T}_{H}\underline{y} = H\underline{y}$$

H has rank p

The residuals is another projection

$$\hat{\underline{\varepsilon}} = \underline{y} - \hat{\underline{y}} = \underline{y} - H\underline{y} = \underbrace{(I - H)}\underline{y}$$

rank $n-p$

Hence,

$$\left\| \frac{\hat{\underline{\varepsilon}}}{\sigma} \right\|^2 = \frac{SSE}{\sigma^2} = \underbrace{\underbrace{z_1^2 + z_2^2 + \dots + z_{n-p}^2}_{n-p} + \underbrace{0^2 + 0^2 + \dots + 0^2}_{p}}_{} \sim \chi^2_{n-p}$$

$z_1, z_2, \dots, z_n$ are new coordinates in a suitable orthonormal basis (a rotation of the coordinate system)

and

$$E\left(\frac{SSE}{\sigma^2}\right) = n - p$$

such that

$$\hat{\sigma}^2 = \frac{SSE}{n-p} \text{ is unbiased for } \sigma^2.$$

$\hat{\varepsilon}$ and $\hat{\beta}$ are jointly multivariate normal with

$$\mathrm{Cov}(\hat{\varepsilon}, \hat{\beta}) = \mathrm{Cov}\Big(\underbrace{\big(I - X(X^TX)^{-1}X^T\big)}_{A} y, \underbrace{(X^TX)^{-1}X^T}_{B} y\Big)$$

$$= A\,\mathrm{Cov}(y, y)\,B^T$$

$$= \sigma^2\big(I - X(X^TX)^{-1}X^T\big)X(X^TX)^{-1}$$

$$= \sigma^2\big(X(X^TX)^{-1} - X(X^TX)^{-1}X^TX(X^TX)^{-1}\big)$$

$$= 0.$$

Do the residuals $\hat{\varepsilon} = y - \hat{y}$ have constant variance? No:

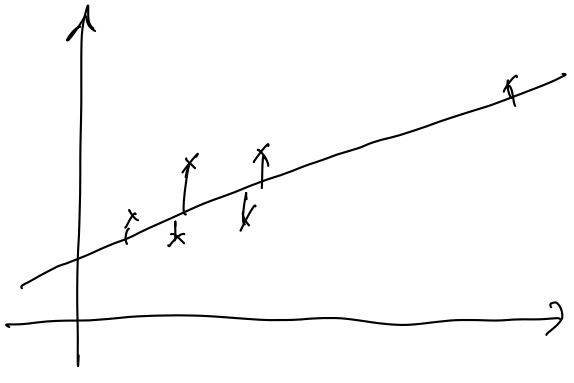
$$\mathrm{Var}(\hat{\varepsilon}) = \mathrm{Var}\big((I - H)y\big)$$

$$= \sigma^2 (I - H)(I - H)$$

$$= \sigma^2 (I - 2H + H^2)$$

$$= \sigma^2 (I - 2H + H)$$

$$= \sigma^2 (I - H)$$



Standardized residuals

$$r_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \overset{\text{approx}}{\sim} N(0, 1)$$

If the the model has an intercept, then since $\hat{\varepsilon}$ is orthogonal to all colums in X including the first column such that

$$\varepsilon^T \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = 0$$

$$\hat{\varepsilon}_1 + \hat{\varepsilon}_2 + \ldots + \hat{\varepsilon}_n = 0$$

T-tests:

$$Var(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = \sigma^2 \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ & \end{bmatrix}$$

It follows that

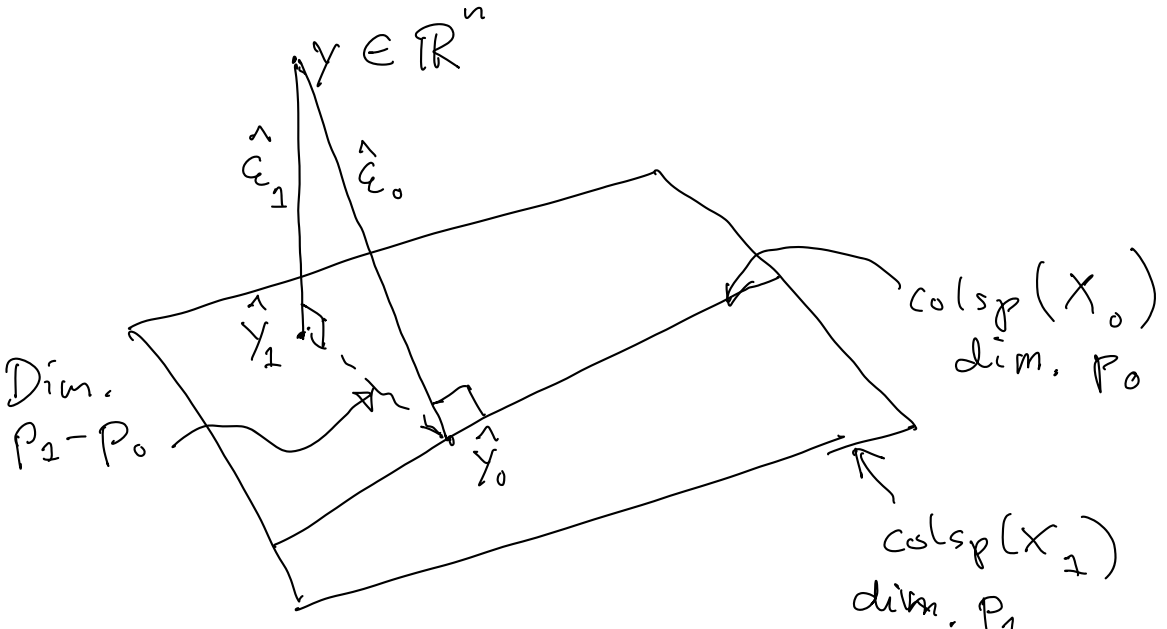
$$T = \frac{\overbrace{\frac{\hat{\beta}_j}{\sigma / \sqrt{c_{ii}}}}^{\sim N(0,1)}}{\sqrt{\underbrace{\frac{SSE}{\sigma^2}}_{\sim \chi^2_{n-p}} / (n-p)}} = \ldots = \frac{\hat{\beta}_j}{\widehat{SE}(\beta_j)}$$

is t-distributed with $n-p$ d.f. under $H_0: \beta_j = 0$

The F-distribution

Testing: $H_0: y = X_0\beta_0 + \varepsilon$ vs $H_1: y = X_1\beta_1 + \varepsilon$:

Vector-space view:



Orthogonality of $\hat{y}_1 - \hat{y}_0$ and $\hat{\varepsilon}_1$ implies that

$$\left\| \frac{\hat{y}_1 - \hat{y}_0}{\sigma} \right\|^2 \sim \chi^2_{p_1 - p_0}$$

and indep. of

$$\underbrace{\left\| \frac{\hat{\varepsilon}_1}{\sigma} \right\|^2}_{\frac{SSE_1}{\sigma^2}} \sim \chi^2_{n-p_1}$$

under H_0. From the Phytagorean theorem we have

$$\|\hat{\varepsilon}_0\|^2 = \|\hat{\varepsilon}_1\|^2 + \|\hat{y}_1 - \hat{y}_0\|^2$$

such that

$$\|\hat{y}_1 - \hat{y}_0\|^2 = \|\hat{\varepsilon}_0\|^2 - \|\hat{\varepsilon}_1\|^2$$

$$= SSE_0 - SSE_1.$$

Hence,

$$F = \frac{\left(SSE_0 - SSE_1\right)/\sigma^2}{\left(p_1 - p_0\right)} \sim F_{p_1 - p_0,\, n - p_1}$$

$$\frac{SSE_1}{\sigma^2} \Big/ (n - p_1)$$

Reject H_0 for large values of F.

Equivalent to a LRT (likelihood ratio test) (exercise 2b)

Testing linear hypothesis

T-test: Testing $H_0: \beta_i = \beta_{i0}$ vs $H_1: \beta_i \neq \beta_{i0}$.

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{SE(\hat{\beta}_i)} \sim t_{n-p}$$

Note that

$$T^2 = \frac{(\beta_i - \beta_{i0})^2}{Var(\hat{\beta}_i)} \sim F_{1, n-p}$$

General linear hypothesis:

$$H_0: \underset{r \times p}{C} \underset{p \times 1}{\underline{\beta}} = \underset{r \times 1}{\underline{d}}$$

vs

$$H_1: C\underline{\beta} \neq \underline{d}$$

Under H_0 $C\hat{\beta} - d$ is normally distributed with

$$E(C\hat{\beta} - d) = 0$$

and variance matrix

$$\mathrm{Var}(C\hat{\beta} - d) = C\,\mathrm{Var}(\hat{\beta})\,C^T$$

$$= \underbrace{\sigma^2 C (X^T X)^{-1} C^T}_{\Sigma}$$

Hence

$$Z = \Sigma^{-1/2}(C\hat{\beta} - d)$$

has mean zero and

$$\mathrm{Var}(Z) = \Sigma^{-1/2}\,\Sigma\,(\Sigma^{-1/2})^T = I$$

that is $Z \sim N(0, I_r)$

Thus,

$$\|Z\|^2 = Z^T Z = z_1^2 + z_2^2 + \ldots + z_r^2$$

$$= (C\hat{\beta} - d)^T \Sigma^{-1} (C\hat{\beta} - d)$$

$$= \frac{1}{\sigma^2}(C\hat{\beta} - d)^T \left(C(X^T X)^{-1} C^T\right)^{-1} (C\hat{\beta} - d)$$

$$\sim \chi^2_r$$

and

$$F = \frac{\frac{1}{\sigma^2}\left(C\hat{\beta} - d\right)^T \left(C(X^TX)^{-1}C^T\right)^{-1}\left(C\hat{\beta} - d\right)/r}{\frac{SSE_1}{\sigma^2}/(n-p)} \Bigg\} \hat{\sigma}^2$$

$$\sim F_{r,\, n-p}$$

Wald-test

How to get H_0? (Next lecture)

Fitting models with linear equality constraints

August 31

Fitting $y = X\beta + \varepsilon$ under the constraint

$$C\beta = d$$

Fahrmeir p. 172–175: Lagrange:

$$\hat{\beta}^{(R)} = \underbrace{\hat{\beta}}_{\text{unrestricted MLE}} - \underbrace{(X^TX)^{-1}C^T\left(C(X^TX)^{-1}C^T\right)^{-1}\left(C\hat{\beta} - d\right)}_{\Delta_{H_0}}$$

It follows that

$$SSE_0 = \|\hat{\varepsilon}_0\|^2 = \|\hat{\varepsilon} + X\Delta_{H_0}\|^2$$

$$= \|\hat{\varepsilon}_1\|^2 + \|X\Delta_{H_0}\|^2$$

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

which leads to

$$F = \frac{(SSE_0 - SSE_1)/r}{SSE_1/(n-p)} \sim F_{r,n-p}$$

$\sim$ ... = previous Wald-test-statistic

Alternative approach via substitution

Partitioning C and $\underline{\beta}$ by columns and rows we can write

$$C\underline{\beta} = \underline{d}$$

as

$$\underbrace{\begin{bmatrix} C_0 & C_r \end{bmatrix}}_{r\times(p-r)\quad r\times r} \begin{bmatrix} \underline{\beta}_0 \\ \underline{\beta}_r \end{bmatrix} = \underline{d}$$

$\underline{\beta}_r$ ← "redundant"

where C_r is invertible.

Using blockwise matrix multiplication

$$C_0\underline{\beta}_0 + C_r\underline{\beta}_r = \underline{d}$$

such that

$$\underline{\beta}_r = C_r^{-1}(C_0\underline{\beta}_0 - \underline{d})$$

Partitioning X the same way, we can write

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

as

$$\underline{Y} = \underbrace{\begin{bmatrix} X_0 & X_r \end{bmatrix}}_{n\times(p-r)\ \ n\times r} \begin{bmatrix} \underline{\beta}_0 \\ \underline{\beta}_r \end{bmatrix} + \underline{\varepsilon}$$

$$= X_0\underline{\beta}_0 + X_r\underline{\beta}_r + \underline{\varepsilon}$$

$$= X_0\underline{\beta}_0 + X_r C_r^{-1}\left(C_0\underline{\beta}_0 - \underline{d}\right) + \underline{\varepsilon}$$

$$= \underbrace{\left(X_0 + X_r C_r^{-1} C_0\right)}_{\text{modified design matrix } X^*}\underline{\beta}_0 + \underbrace{X_r C_r^{-1}\underline{d}}_{\text{offset } \underline{a}} + \underline{\varepsilon}$$

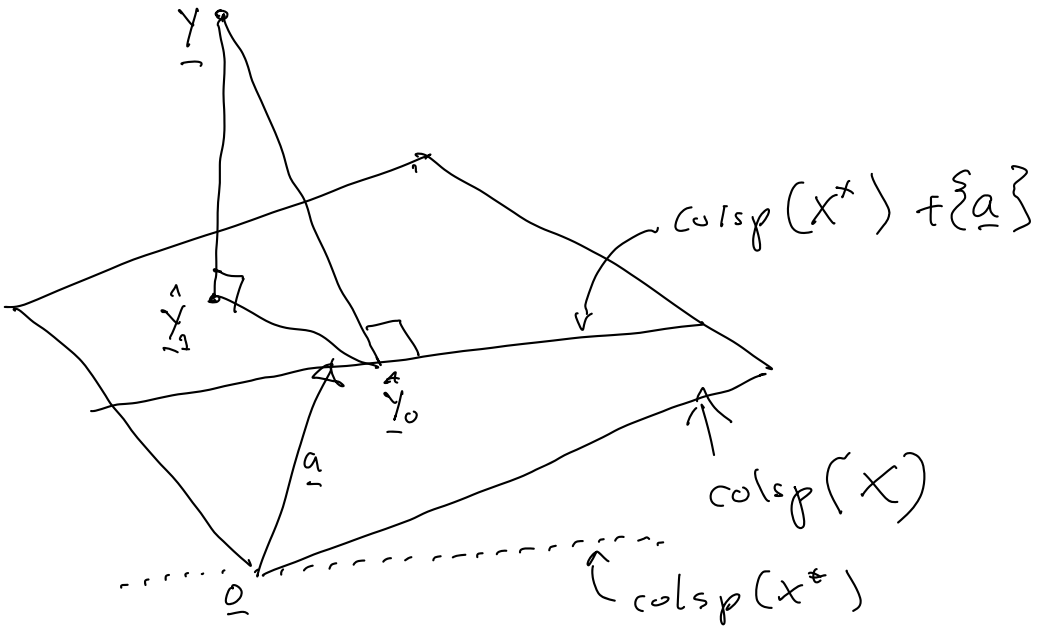
2.9

7.9

Fitted values $\hat{\underline{Y}}_0$ under H_0 given by projection of $\underline{Y}$ onto affine subspace

$$\left\{X\underline{\beta} \mid \underline{\beta}\in\mathbb{R}^p,\ C\underline{\beta} = \underline{d}\right\} = \text{colsp}(X^*) + \{\underline{a}\}$$

↑ Minkowski sum



Thus

$$\hat{\underline{y}}_0 = \text{Proj}^{C}_{\text{colsp}(X^*)}\left(\underline{y} - \underline{a}\right) + \underline{a}$$

$$\hat{\underline{\beta}}_0 = \left(X^{*T}X^*\right)^{-1}X^{*T}\left(\underline{y} - \underline{a}\right)$$

and

$$\hat{\underline{\beta}}_r = -C_r^{-1}\left(C_0\hat{\underline{\beta}}_0 - \underline{d}\right)$$

and the same F-statistic as before.

This can be generalized to fitting GLMs under linear equality constraints facilitating LRTs (as opposed to less reliable Wald-tests).

Categorical covariates (factors) and interactions

One-way anova, k groups: For $j = 1, 2, \dots, k$,
and $i = 1, 2, \dots, n_j$

$$Y_{ij} = \mu + \underbrace{\alpha_j}_{\text{effect of belonging to group } j} + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

Alternative notation (see Wood, p. 198).
For $i = 1, 2, \dots, n$ $\quad \left(n = \sum_{j=1}^{k} n_j\right)$

$$Y_i = \mu + \alpha_{j(i)} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Unknown parameters: $\mu, \alpha_1, \alpha_2, \dots, \alpha_k, \sigma^2$

In matrix notation

$$\underline{Y} = \underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix}}_{X} \begin{matrix} \} j=1 \\ \\ \} j=2 \\ \\ \} j=3 \\ \\ \end{matrix} \underbrace{\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}}_{\underline{\beta}} + \underline{\varepsilon}$$

Columns are not indep. (model is not identifiable)

Omit second column $\Leftrightarrow$ imposing the constraint $\alpha_1 = 0$.
(treatment contrasts)

(Symbolic (Wilkinson-Rogers) notation in R:

$y \sim \underbrace{\text{glucose}}_{\text{factor}}$)

Thus, for

$j = 1, \quad E(y_i) = \mu$

$j = 2, \quad E(y_i) = \mu + \alpha_2$

$j = 3, \quad E(y_i) = \mu + \alpha_3$

i.e. α_2 is the difference in $E(y_i)$ in group 2 relative to group 1 (the control/placebo treatment)

Interactions (with numerical covariate)

"Alternative" notation: For $i = 1, 2, \ldots, n$

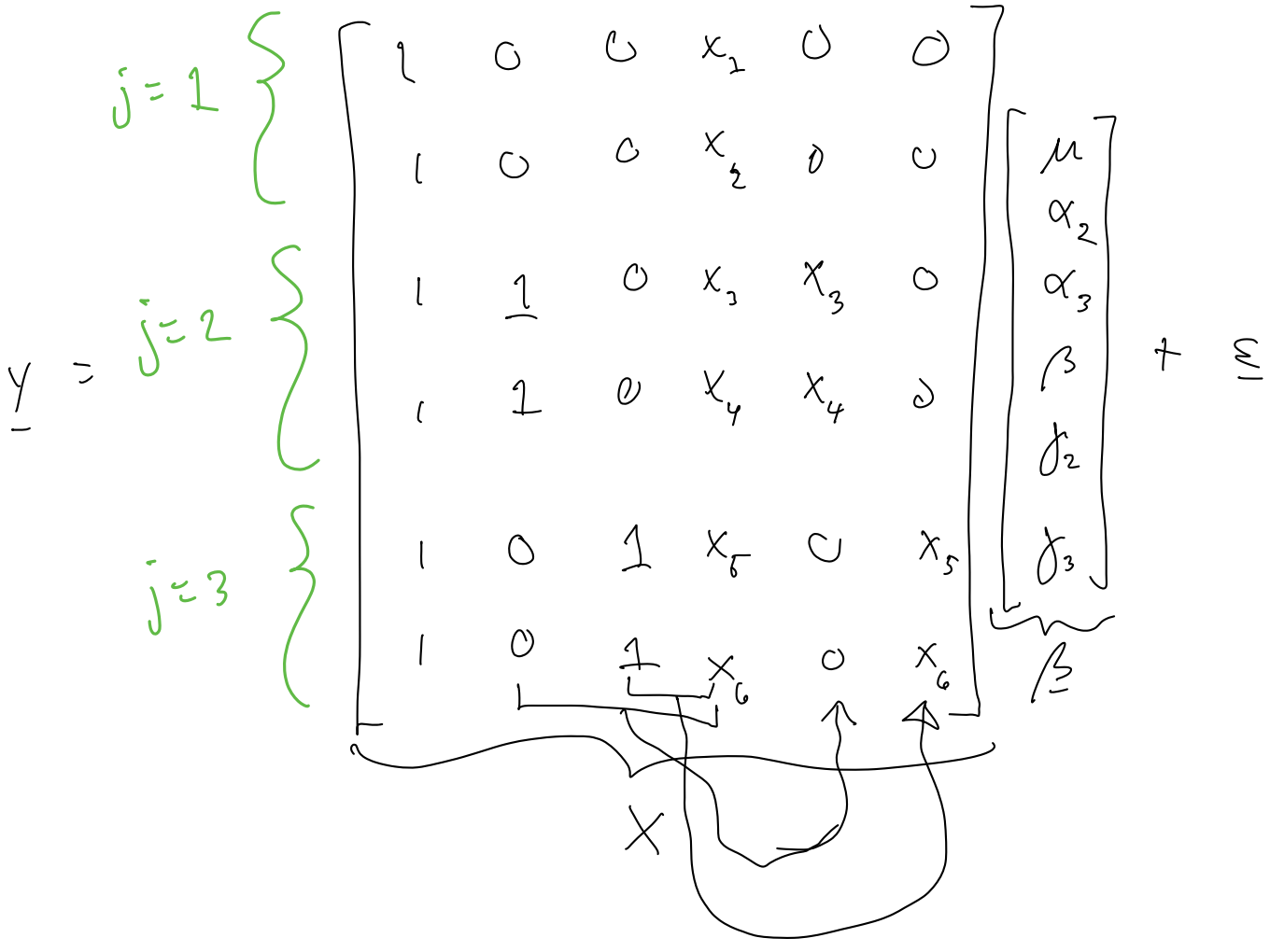
$$y_i = \mu + \underbrace{\alpha_{j(i)}}_{\substack{\text{Main effect} \\ \text{of belonging} \\ \text{to group } j}} + \underbrace{\beta \cdot x_i}_{\substack{\text{Main effect} \\ \text{of num.} \\ \text{covariate } x_i}} + \underbrace{\gamma_{j(i)} \cdot x_i}_{\substack{\text{Interaction} \\ \text{between} \\ j(i) \text{ and } x_i}} + \varepsilon_i$$

$$= \underbrace{(\mu + \alpha_{j(i)})}_{\substack{\text{"Intercept"} \\ \text{in group } j}} + \underbrace{(\beta + \gamma_{j(i)})}_{\substack{\text{"Slope" in} \\ \text{group } j}} x_i + \varepsilon_i$$

Symbolic notation

$$\log(\text{diam}) \sim \text{glucose} + \log(\text{cons}) + \text{glucose} : \log(\text{cons})$$

Impose the constraints $\alpha_1 = 0$ and $\gamma_1 = 0$ (if using treatment contrasts), the model in matrix notation becomes



Thus, for

$j = 1$, $E(y_i) = \mu + \beta x_i$

$j = 2$, $E(y_i) = \mu + \alpha_2 + (\beta + \gamma_2) x_i$

$j = 3,\ E(y_i) = \mu + \alpha_3 + (\beta + \gamma_3) x_i$

Difference in intercept and slope relative to the control group $j = 1$

Sum to zero contrasts: Impose the constraints $\sum_{j=1}^{k} \alpha_j = 0$, $\sum_{j=1}^{k} \gamma_j = 0$

such that e.g.

$\alpha_3 = -\alpha_1 - \alpha_2$, $\gamma_3 = -\gamma_1 - \gamma_2$

R: contrasts(glucose) <- "contr.sum"

glucose: factor

Model in matrix notation:

$$\underset{\sim}{y} = \begin{matrix} j=1 \\ \\ j=2 \\ \\ \\ j=3 \\ \end{matrix} \overbrace{\begin{bmatrix} 1 & 0 & x_1 & x_1 & 0 \\ 1 & 0 & x_2 & x_2 & 0 \\ 1 & 0 & x_3 & 0 & x_3 \\ 1 & 0 & x_4 & 0 & x_4 \\ 1 & -1 & -1 & x_5 & -x_5 & -x_5 \\ 1 & -1 & -1 & x_6 & -x_6 & -x_6 \end{bmatrix}}^{X} \underbrace{\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta \\ \gamma_1 \\ \gamma_2 \end{bmatrix}}_{\underset{\sim}{\beta}} + \underset{\sim}{\varepsilon}$$

Second alternative notation:

$$\log(\text{diam}_i) = \mu + \alpha_{\text{glucose}(i)} + \beta \log(\text{conc}_i) + \gamma_{\text{glucose}(i)} \cdot \log(\text{conc}_i) + \varepsilon_i$$

GLMs: Linear predictor constructed the same way.

Binary regression (Ch. 5)

Sept. 1

Ungrouped data: $(y_i, \underline{x}_i)$, $i = 1, 2, \ldots, n$

Model:

$$y_i \sim \text{Bernulli}(\pi_i)$$

where π_i depends on a linear predictor $\eta_i = \underline{x}_i^T \underline{\beta}$ via response function

$$\pi_i = h(\eta_i)$$

$$g(\pi_i) = \eta_i$$

Logit link

$$g(\pi) = \ln \frac{\pi}{1-\pi}, \quad h(\eta) = \frac{e^\eta}{1+e^\eta} = \frac{1}{1+e^{-\eta}}$$

Odds interpretation

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0} e^{\beta_1 x_{i1}} \cdots e^{\beta_k x_{ik}}$$

Unit change in x_{i1} changes the odds by a factor (odds ratio) of e^{β_i}

Probit link

Response function

$$\pi = \Phi(\eta) = \int_{-\infty}^{\eta} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \underbrace{\text{pnorm}(\eta)}_{\text{R-function}}$$

Link function

$$\Phi^{-1}(\pi) = \eta$$

$$\text{probit}(\pi) = \eta \qquad = \text{qnorm}(p_i)$$

Example: Model:

Age at time of menarche

$$T_i \sim N(\mu, \sigma^2)$$

We then have

$$\pi_i = P(Y_i = 1) = P(T_i < \underbrace{t_i}_{\text{age at medical examination}})$$

$$= P\left(\underbrace{\frac{T_i - \mu}{\sigma}}_{\sim N(0,1)} < \frac{t_i - \mu}{\sigma}\right)$$

$$= \Phi\left(\frac{t_i - \mu}{\sigma}\right)$$

$$\Phi^{-1}(\pi_i) = \text{probit}\,\pi_i = \frac{t_i - \mu}{\sigma}$$

$$= \underbrace{-\frac{\mu}{\sigma}}_{\beta_0} + \underbrace{\frac{1}{\sigma}}_{\beta_1} t_i$$

} a glm.

glm(y ~ age , family = binomial(link = "probit"))

It follows that

$$\sigma = \frac{1}{\beta_1} \qquad \mu = -\frac{\beta_0}{\beta_1}$$

Hence, the MLEs of μ and σ are

$$\hat{\sigma} = \frac{1}{\hat{\beta}_1} = \frac{1}{0.86} = 1.16 \text{ years}, \quad \hat{\mu} = -\frac{\hat{\beta}_0}{\hat{\beta}_1} = \frac{-11.37}{0.86}$$

$$= 13.22 \text{ years}$$

Exercise 5. Compute the S.E.s using the delta-method

$$Z = f(X, Y) \approx f(\mu_X, \mu_Y) + \frac{\partial f}{\partial x}(X - \mu_X) + \frac{\partial f}{\partial y}(Y - \mu_Y)$$

$$\text{Var } Z \approx \left(\frac{\partial f}{\partial x}\right)^2 \text{Var } X + \left(\frac{\partial h}{\partial y}\right)^2 \text{Var } Y + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}\text{Cov}(X, Y)$$

7.9

9.9

Complementary log-log link

Response func.

$$\pi = h(\eta) = 1 - e^{-e^{\eta}}$$

and link func.

$$g(\pi) = \underbrace{\ln(-\ln(1-\pi))}_{\text{cloglog}} = \eta$$

Project 1 : Show that this corresponds to the latent variable model

$T_i \sim \text{Weibull}(\alpha, \beta)$ and $\pi_i = P(T_i > t_i)$

Example 1: Survival analysis: Suppose that the hazard for the ith subject is

$$\lambda_i(t) = \underbrace{\lambda_0(t)}_{\text{baseline hazard}} e^{\beta_1 x_i}$$

$$= \lim_{\Delta t \to 0} \frac{P(T_i \le t + \Delta t \mid T_i > t)}{\Delta t}$$

At the end of the experiment at time τ each subject is dead with probability

$$\pi_i = P(Y_i = 1) = P(T_i < \tau)$$

$$= 1 - e^{-\int_0^\tau \lambda_i(t)\,dt} \quad \text{(known from survival analysis)}$$

$$= 1 - e^{-e^{\beta_1 x_i} \underbrace{\int_0^\tau \lambda_0(t)\,dt}_{= e^{\beta_0}}}$$

$$= 1 - e^{-e^{\underbrace{\beta_0 + \beta_1 x_i}_{\eta_i}}}$$

$$= 1 - e^{-e^{\eta_i}}$$

i.e.

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

Interpretation of β_1: One unit change in x_i changes the hazard $\lambda_i(t)$ by a factor of e^{β_1}, e.g. if $\beta_1 = -0.1$ the $e^{\beta_1} \approx 0.91$, that is, a 9% decrease.

11:1

Example 2: Presence-absence data in ecology:

Suppose the number of individuals of a given species $Z_i \sim \text{Poisson}(\lambda_i)$ at given sampling sites $i = 1, 2, \ldots, n$, but that we only observe/record the binary response

$$Y_i = \begin{cases} 1 & \text{if } Z_i \geq 1 \quad \text{(presence)} \\ 0 & \text{if } Z_i = 0 \quad \text{(absent)} \end{cases}$$

Assume that $\ln \lambda_i = \eta_i = \underset{\sim}{x}_i^T \underset{\sim}{\beta}$

Then

$$\pi_i = P(Y_i = 1) = P(Z_i \geq 1)$$

$$= 1 - P(Z_i = 0)$$

$$= 1 - \frac{\lambda_i^0 e^{-\lambda_i}}{0!}$$

$$= 1 - e^{-e^{\eta_i}}$$

that is,

$$\text{cloglog}(\pi_i) = \eta_i = \underset{\sim}{x}_i^T \underset{\sim}{\beta}$$

Grouped data: Each observation $i = 1, 2, \ldots, n$ involves n_i Bernoulli trials such that

$$y_i \sim \text{bin}(n_i, \pi_i)$$

Notation:

$$\bar{y}_i = \frac{y_i}{n_i}$$

$$\bar{y}_i \sim \frac{\text{bin}(n_i, \pi_i)}{n_i}$$

R: glm(cbind(y, $n - y$) ~ x, binomial)

cbind(y, $n - y$): $n \times 2$ matrix

$$\begin{bmatrix} y_1 & n_1 - y_1 \\ y_2 & n_2 - y_2 \\ \vdots & \vdots \end{bmatrix} \} n \text{ rows}$$

If $\text{Var}(y_i) > n_i \pi_i (1 - \pi_i)$ we have overdispersion.

Log-likelihood (non-grouped data)

$$l(\underline{\beta}) = \sum_{i=1}^{n} l_i(\underline{\beta})$$

$$P(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

where the contributions

$$l_i(\underline{\beta}) = y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)$$

$$= y_i \ln \frac{\pi_i}{1-\pi_i} + \ln(1-\pi_i)$$

canonical link func.

$$= y_i \eta_i + \ln \frac{1}{1+e^{\eta_i}}$$

$$= y_i \eta_i - \ln(1+e^{\eta_i})$$

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

$$1-\pi_i = \frac{1}{1+e^{\eta_i}}$$

$$\eta_i = \underline{x}_i^T \underline{\beta}$$

Score function

$$\underline{S}(\underline{\beta}) = \frac{\partial}{\partial \underline{\beta}} l(\underline{\beta}) = \frac{\partial}{\partial \underline{\beta}} \sum_{i=1}^{n} l_i(\underline{\beta}) = \sum_{i=1}^{n} \underbrace{\frac{\partial}{\partial \underline{\beta}} l_i(\underline{\beta})}_{\underline{S}_i(\underline{\beta})}$$

Contribution from i-th observation

$$\underline{S}_i(\underline{\beta}) = \frac{\partial}{\partial \underline{\beta}} l_i(\underline{\beta})$$

$$= \frac{\partial}{\partial \underline{\beta}} \left(y_i \eta_i - \ln(1+e^{\eta_i}) \right)$$

$$= y_i \frac{\partial \eta_i}{\partial \underline{\beta}} - \frac{e^{\eta_i}}{1+e^{\eta_i}} \cdot \frac{\partial \eta_i}{\partial \underline{\beta}}$$

$$= (y_i - \pi_i) \frac{\partial}{\partial \underline{\beta}} (\underline{x}_i^T \underline{\beta})$$

$$= (y_i - \pi_i) \underline{x}_i$$

$$= (y_i - \pi_i) \begin{bmatrix} 1 \\ x_{i1} \\ x_{i2} \\ \vdots \end{bmatrix}$$

Total score function

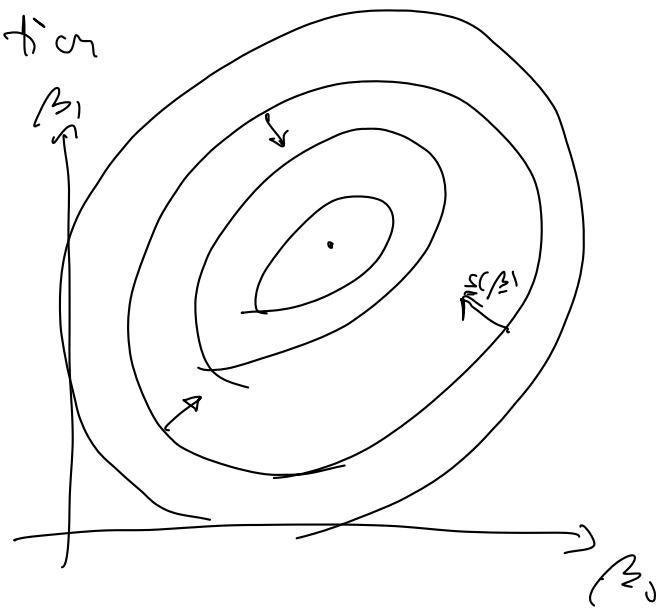
$$\underbrace{\underline{S}(\underline{\beta})}_{p \times 1} = \sum_{i=1}^{n} \underline{S}_i(\underline{\beta}) = \sum_{i=1}^{n} (y_i - \pi_i) \underbrace{\underline{x}_i}_{p \times 1}$$

Keep in mind that π_i is a function of η_i which in turn is a function of $\underline{\beta}$, i.e. $\pi_i = \pi_i(\eta_i(\underline{\beta}))$.

Need to solve the equation

$$\underbrace{\underline{S}(\underline{\beta})}_{p \times 1} = \underbrace{\underline{0}}_{p \times 1}$$

A system of p non-linear eqs. in p unknowns.



Next time: algorithm for solving this numerically

Newton's method

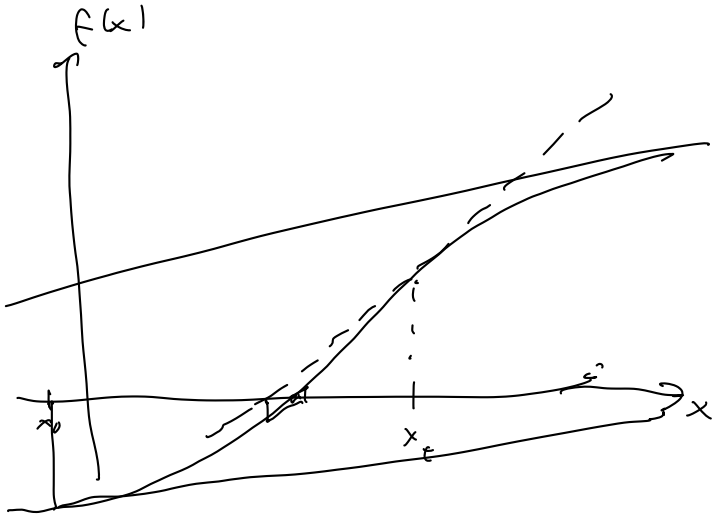
In one dimension. Find the root of

$$f(x) = 0$$

Idea: Current best guess at the solution x_t.

Set linear approximation around x_t equal to zero and solve for x:

$$f(x) \simeq f(x_t) + f'(x_t)(x - x_t) = 0$$



$$x_{t+1} = x_t - \frac{f(x_t)}{f'(x_t)}$$

Multiple dimensions: Current best guess is $\underline{\beta}_t$

Approximate $\underline{s}(\underline{\beta})$ by

$$\underline{s}(\underline{\beta}) \simeq \underline{s}(\underline{\beta}_t) + \frac{\partial}{\partial \underline{\beta}} \underline{s}^T(\underline{\beta}_t)(\underline{\beta} - \underline{\beta}_t)$$

$$= \underline{s}(\underline{\beta}_t) + \frac{\partial}{\partial \underline{\beta}} \frac{\partial}{\partial \underline{\beta}^T} l(\underline{\beta}_t)(\underline{\beta} - \underline{\beta}_t)$$

$$= \underline{s}(\underline{\beta}_t) - H(\underline{\beta}_t)(\underline{\beta} - \underline{\beta}_t) = \underline{0}$$

Solving for $\underline{\beta}$ we obtain

$$\underline{\beta}_{t+1} = \underline{\beta}_t + \left(H^{-1}(\underline{\beta}_t)\right) \underline{s}(\underline{\beta}_t) \qquad \text{(Newton's method)}$$

In practice we replace $H(\underline{\beta}_t)$ by $F(\underline{\beta}_t)$ (expected Fisher information)

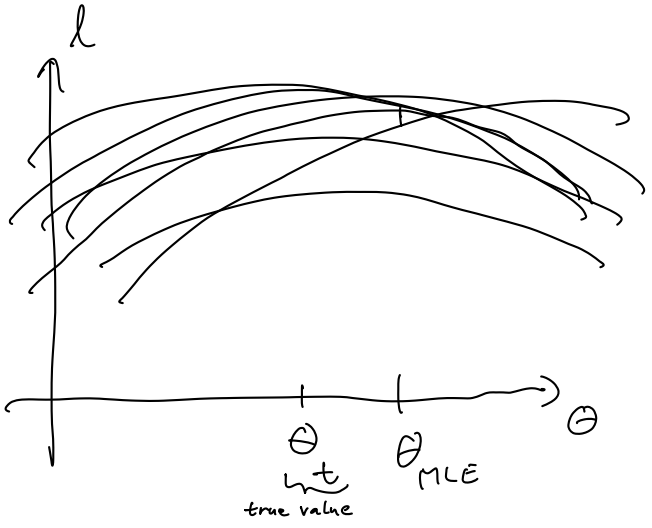
$$\underline{\beta}_{t+1} = \underline{\beta}_t + F^{-1}(\underline{\beta}_t)\,\underline{s}(\underline{\beta}_t)$$ } Fisher scoring algorithm

Recall that

$$F(\underline{\beta}) = E(H(\underline{\beta})) = E\left(-\frac{\partial^2 \ell}{\partial \underline{\beta}\, \partial \underline{\beta}^T}\right)$$

Note: $\ell(\underline{\beta}; Y)$ is a random variable

Some general properties of the log-likelihood



random

$$\ell(\theta; Y) = \ln L[\theta; Y]$$
$$= \ln f(Y; \theta)$$

joint pdf of all data Y

1. $$E\left(\frac{\partial \ell}{\partial \underline{\theta}}\bigg|_{\theta=\theta_t}\right) = E\left(\underline{s}(\underline{\theta}_t)\right) = \underline{0}$$

Proof:

$$E\left(\frac{\partial \ell}{\partial \underline{\theta}}\right) = \int \frac{\partial}{\partial \underline{\theta}} \ln f(y)\, f(y)\, dy$$

$$= \int \frac{1}{f(y)} \frac{\partial f}{\partial \underline{\theta}} \cdot f(y)\, dy$$

$$= \frac{\partial}{\partial \underline{\theta}} \int f(y)\, dy$$

$$= \frac{\partial}{\partial \underline{\theta}} 1 = \underline{0}$$

2. $\underbrace{\mathrm{Var}\left(\frac{\partial \ell}{\partial \underline{\theta}}\right)}_{p \times p} \overset{\text{by def.}}{=} E\left(\frac{\partial \ell}{\partial \underline{\theta}} - E\left(\frac{\partial \ell}{\partial \underline{\theta}}\right)\right)\left(\frac{\partial \ell}{\partial \underline{\theta}} - E\left(\frac{\partial \ell}{\partial \underline{\theta}}\right)\right)^T$

$\underset{p \times 1}{\frac{\partial \ell}{\partial \underline{\theta}}}$

$$\overset{\text{prop. 1}}{=} E\left(\frac{\partial \ell}{\partial \underline{\theta}} \frac{\partial \ell}{\partial \underline{\theta}^T}\right)$$

3. $$\underbrace{\mathrm{Var}\left(\left.\frac{\partial \ell}{\partial \underline{\theta}}\right|_{\underline{\theta}=\underline{\theta}_t} \right)}_{p\times p} = \underbrace{E\left(\left. -\frac{\partial^2 \ell}{\partial \underline{\theta}\, \partial \underline{\theta}^T}\right|_{\underline{\theta}=\underline{\theta}_t} \right)}_{p\times p} = F(\underline{\theta}_t)$$

Proof:

$$\frac{\partial}{\partial \underline{\theta}} E\left(\frac{\partial \ell}{\partial \underline{\theta}^T}\right) = \frac{\partial}{\partial \underline{\theta}} \underline{0}^T = 0$$

$\frac{\partial}{\partial \underline{\theta}}$: $p\times 1$; $\underline{0}^T$: $1\times p$; 0: $p\times p$

$$\frac{\partial}{\partial \underline{\theta}} \int \frac{\partial}{\partial \underline{\theta}^T} \ln f(y)\cdot f(y)\, dy = 0$$

$$\int \frac{\partial}{\partial \underline{\theta}} \left(\frac{\partial}{\partial \underline{\theta}^T} \ln f(y) \cdot f(y) \right) dy = 0$$

$$\int \left(\frac{\partial}{\partial \underline{\theta}} \frac{\partial}{\partial \underline{\theta}^T} \ln f(y)\, f(y) + \underbrace{\frac{\partial}{\partial \underline{\theta}} f(y)}_{= \frac{\partial}{\partial \underline{\theta}} \ln f(y)\cdot f(y)} \frac{\partial}{\partial \underline{\theta}^T} \ln f(y) \right) dy = 0$$

$$-F(\underline{\theta}) + E\left(s(\underline{\theta})\, s^T(\underline{\theta}) \right) = 0$$

$$F(\underline{\theta}) = \mathrm{Var}\left(\underline{s}(\underline{\theta})\right)$$

Fisher information for binary glm

Recall: $\pi_i = \pi_i(\eta_i(\underline{\beta}))$

$$\underline{s}(\underline{\beta}) = \sum_{i=1}^{n} \underline{s}_i(\underline{\beta}) = \sum_{i=1}^{n} (y_i - \pi_i)\underline{x}_i$$

Expected Fisher information via property 3:

$$F(\underline{\beta}) = \mathrm{Var}(\underline{s}(\underline{\beta})) = \mathrm{Var}\left(\sum_{i=1}^{n} (y_i - \pi_i)\underline{x}_i\right)$$

($\underline{x}_i$: $p \times 1$)

indep obs.

$$= \sum_{i=1}^{n} \mathrm{Var}(\underbrace{\underline{x}_i \cdot y_i}_{p \times 1})$$

$$= \sum_{i=1}^{n} \underline{x}_i \underbrace{\mathrm{Var}(y_i)}_{\pi_i(1-\pi_i)} \underline{x}_i^T$$

$$= \sum_{i=1}^{n} \pi_i(1-\pi_i)\underbrace{\underline{x}_i \underline{x}_i^T}_{p \times 1 \times 1 \times p}$$

$p \times p$

Observed Fisher information:

$$H_i(\underline{\beta}) = -\frac{\partial}{\partial \underline{\beta} \partial \underline{\beta}^T} l_i(\beta_i) = -\frac{\partial}{\partial \underline{\beta}} \underline{s}_i^T(\underline{\beta})$$

$$= -\frac{\partial}{\partial \underline{\beta}} \left(y_i - \frac{e^{\eta_i}}{1+e^{\eta_i}} \right) \cdot \underline{x}_i^T$$

$$= \frac{e^{\eta_i}(1+e^{\eta_i}) - e^{\eta_i} e^{\eta_i}}{(1+e^{\eta_i})^2} \frac{\partial \eta_i}{\partial \underline{\beta}} \underline{x}_i^T$$

$$= \pi_i (1-\pi_i) \frac{\partial}{\partial \underline{\beta}} (\underline{x}_i^T \underline{\beta}) \cdot \underline{x}_i^T$$

$$= \pi_i (1-\pi_i) \underline{x}_i \underline{x}_i^T$$

$$\frac{e^{\eta_i}}{(1+e^{\eta_i})^2} = \frac{1}{1+e^{\eta_i}} \cdot \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

$1-\pi_i$ $\quad$ π_i

$$H(\underline{\beta}) = \sum_{i=1}^{n} \pi_i (1-\pi_i) \underline{x}_i \underline{x}_i^T$$

Expected Fisher info. via def.:

$$F(\underline{\beta}) = E\, H(\underline{\beta}) = E\left(\sum_{i=1}^{n} \pi_i (1-\pi_i) \underline{x}_i \underline{x}_i^T \right)$$

and so

$$T(\underline{\beta}) = H(\underline{\beta})$$

Because we are using the canonical link functin.

9.9

14.9

Is $F(\underline{\beta})$ always positive definite and invertible? (p. 283 in Fahrmeir)

Binary regression:

$$F(\underline{\beta}) = \sum_{i=1}^{n} \pi_i(1-\pi_i)\underline{x}_i\underline{x}_i^T = X^T W X$$

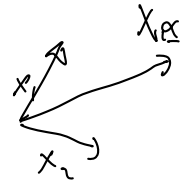
where $W = \text{diag}(\underline{\pi}(1-\underline{\pi}))$

$$= \begin{bmatrix} \pi_1(1-\pi_1) & & \\ & \pi_2(1-\pi_2) & \\ & & \ddots \end{bmatrix}$$

Recall: If X has full rank ($p \leq n$ indep columns), then

$$\|X\underline{\alpha}\| = \underline{\alpha}^T X^T X \underline{\alpha} > 0 \quad \text{for all } \underline{\alpha} \neq \underline{0}$$

($X\underline{\alpha}$: lin. comb. of columns of X)

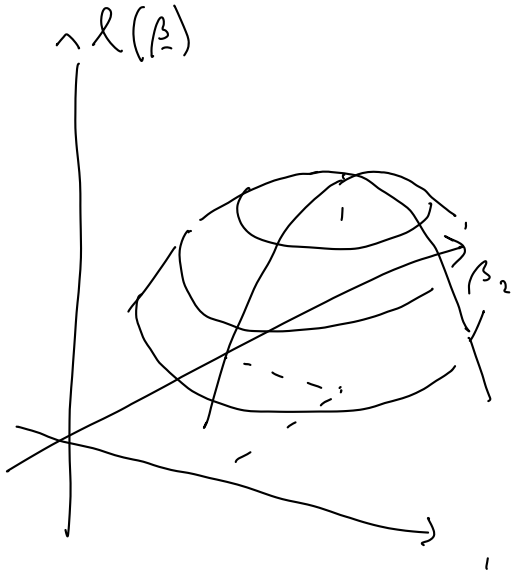


i.e. $X^T X$ is positive definite and invertible.

Let $Z = \text{diag}\left(\sqrt{\underline{\pi}(1-\underline{\pi})}\right) X$, i.e. each row $i = 1, \ldots, n$ of X multiplied by $\sqrt{\pi_i(1-\pi_i)}$. For $\underline{\beta} \in \mathbb{R}^p$, $0 < \pi_i < 1$ and $\sqrt{\pi_i(1-\pi_i)} > 0$ for all $i = 1, 2, \ldots, n$, and $\text{rowsp}(Z) = \text{rowsp}(X) \Rightarrow \text{rank}(Z) = \text{rank}(X)$,

$\Rightarrow F(\underline{\beta}) = Z^T Z$ is positive definite and invertible.

If $H(\beta) = F(\beta) \Rightarrow l(\beta)$ is concave (with a single maxima)



But $F(\beta)$ may tend to positive semi-definite as some $\beta_i \to +/-\infty$

Recommended exercise: For a binary glm with an intercept term included and a logit link-function, show that

$$\frac{1}{n}\sum y_i = \frac{1}{n}\sum \hat{\pi}_i \iff \sum_{i=1}^{n}(y_i - \hat{\pi}_i) = 0$$

where $\hat{\pi}_i = h(\hat{\eta}_i) = \frac{1}{1+e^{-x_i^T\beta}}$ (the fitted probabilities).

Would the same be true for other link functions / models without an intercept?

Fisher scoring algorithm for the linear model

$$y_i \sim N(\mu_i, \sigma^2) \qquad \underbrace{\mu_i = \eta_i}_{\text{identity link}} \quad , \quad \eta_i = \underline{x}_i^T \underline{\beta}$$

$$\ell_i = C - \frac{1}{2\sigma^2}(y_i - \mu_i)^2$$

$$\underline{s}_i = \frac{\partial \ell_i}{\partial \underline{\beta}} = \frac{1}{\sigma^2}(y_i - \mu_i)\frac{\partial \mu_i}{\partial \underline{\beta}}$$

$$= \frac{1}{\sigma^2}(y_i - \mu_i)\underline{x}_i$$

$$\underline{s}(\underline{\beta}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)\underline{x}_i = \ldots = \frac{1}{\sigma^2} X^T(Y - X\underline{\beta})$$

$$F(\underline{\beta}) = \mathrm{Var}(\underline{s}(\underline{\beta})) = \left(\frac{1}{\sigma^2}\right)^2 \sum_{i=1}^{n} \mathrm{Var}(\underline{x}_i y_i)$$

$$= \frac{1}{\sigma^4}\sum x_i \underbrace{\mathrm{Var}(y_i)}\, \underline{x}_i^T$$

$$= \frac{1}{\sigma^2} \sum \underline{x}_i \underline{x}_i^T$$

$$= \frac{1}{\sigma^2} X^T X$$

Fish. scor. algorithm.

$$\underline{\beta}_{t+1} = \underline{\beta}_t + \underbrace{\left(\frac{X^T X}{\sigma^2}\right)^{-1}}_{F^{-1}(\underline{\beta}_t)} \underbrace{\frac{1}{\sigma^2}\left(X^T Y - X^T X \underline{\beta}_t\right)}_{\underline{S}(\underline{\beta}_t)}$$

$$= \cancel{\underline{\beta}_t} + (X^T X)^{-1} X^T Y - \cancel{\underline{\beta}_t}$$

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

Very fast convergence (since $\ell(\underline{\beta})$ is exactly quadratic in $\underline{\beta}$!)

Fisher scoring alg. sometimes diverge

Minimal example: A single observation, no covariates (intercept only)

$y \sim \text{bin}(n, \pi)$, $\text{logit}(\pi) = \eta$, $\eta = \beta_0$

$$l(\beta) = \ln\binom{n}{y} + y \ln \pi + (n-y)\ln(1-\pi)$$

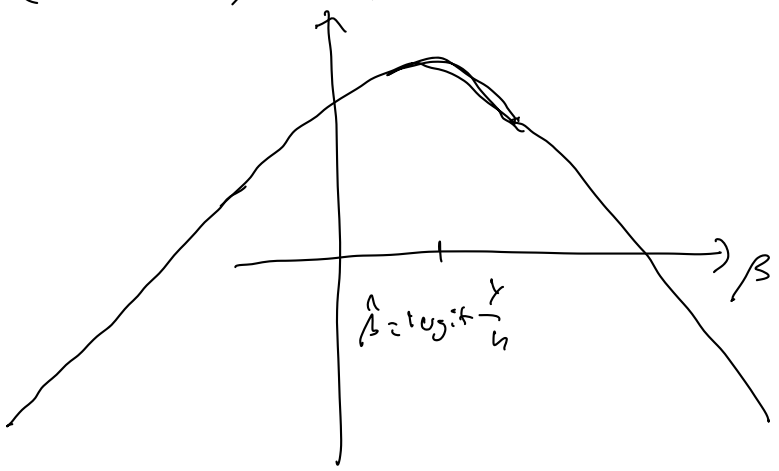
$$= C + y \ln \frac{\pi}{1-\pi} + n \ln(1-\pi)$$

$$= C + y\beta + n \ln\left(1 - \frac{e^\beta}{1+e^\beta}\right) = \ldots =$$

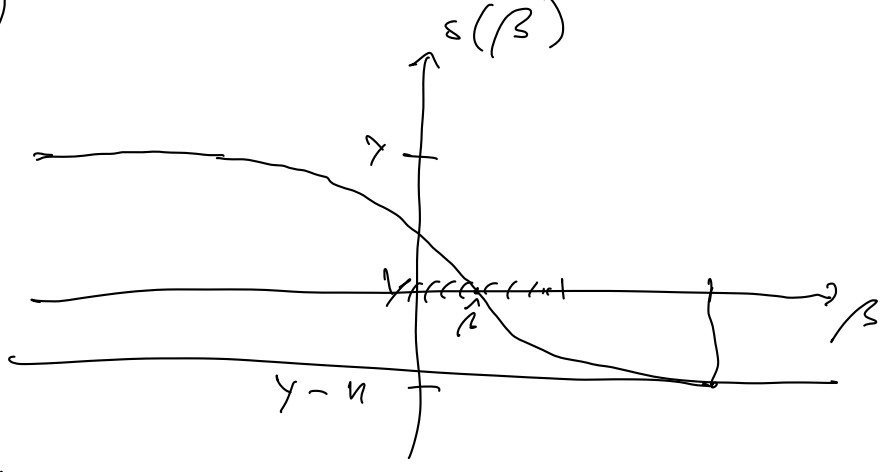
$$= C + y\beta - n \ln(1+e^\beta)$$

$\hat{\pi} = \frac{y}{n}$

$\hat{\beta} = \text{logit}\,\hat{\pi}$



$$s(\beta) = y - n\frac{e^\beta}{1+e^\beta}$$



Newton method = Fisher scoring alg. in this case will overshoot for bad starting values.

Starting values sometimes important, glm(, start=)

Some asymptotic properties of MLEs

With independent observations, as the sample size $n \to \infty$,

$$\hat{\underset{\sim}{\theta}}_{MLE} \sim N\left(\underset{\sim}{\theta}, F^{-1}(\underset{\sim}{\theta})\right)$$

Sketch of proof: Linear approximation of $s(\hat{\theta})$ around θ (the true value of θ):

$$\underbrace{\underset{\sim}{s}(\hat{\underset{\sim}{\theta}})}_{=0} \simeq \underset{\sim}{s}(\underset{\sim}{\theta}) + \underbrace{\left.\frac{\partial^2 \ell}{\partial \underset{\sim}{\theta}\, \partial \underset{\sim}{\theta}^T}\right|_{\theta}}_{-H(\underset{\sim}{\theta})} \cdot (\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta})$$

Hence

$$\hat{\theta} - \theta \simeq H^{-1}(\theta) s(\theta)$$

and

$$\sqrt{n}(\hat{\theta} - \theta) \simeq \sqrt{n}\, H^{-1}(\theta)\, s(\theta)$$

$$= \left(\frac{H(\theta)}{n}\right)^{-1} \frac{\underset{\sim}{s}(\underset{\sim}{\theta})}{\sqrt{n}} \xrightarrow{n \to \infty} \left(\frac{F(\theta)}{n}\right)^{-1} W$$

where $W \sim N\left(0, \frac{F(\theta)}{n}\right)$ (since $E(\underset{\sim}{s}(\underset{\sim}{\theta})) = \underset{\sim}{0}$ and $\mathrm{Var}(\underset{\sim}{s}(\underset{\sim}{\theta})) = F(\theta)$)

Hence,

$$E(\hat{\theta}) \simeq \theta$$

$$\mathrm{Var}\left(\sqrt{n}(\hat{\theta} - \theta)\right) \simeq \left(\frac{F(\theta)}{n}\right)^{-1} \frac{F(\theta)}{n} \left(\frac{F(\theta)}{n}\right)^{-1^T}$$

$$= n F^{-1}(\theta)$$

$$\mathrm{Var}(\hat{\theta}) \simeq F^{-1}(\theta)$$

In practice (for finite sample size) we can use this as an approximation and estimate $Var(\hat{\theta})$

by $\widehat{Var(\hat{\theta})} = \widehat{F^{-1}(\theta)} = F^{-1}(\hat{\theta})$

or by $H^{-1}(\hat{\theta})$ (as suggested by Efron & Hinkley, 1978)

summary(glm( ))

vcov( )

Hypothesis testing for GLMs

$H_0: C\underset{\sim}{\beta} = \underset{\sim}{d}$ vs. $H_1: C\underset{\sim}{\beta} \neq \underset{\sim}{d}$

$r \times p \quad p \times 1 \quad r \times 1$

Fitted model	Type of test
Both H_0 and H_1	Likelihood ratio test (usually most accurate) Under H_0: $2\left(l(\hat{\underset{\sim}{\beta}}_1) - l(\hat{\underset{\sim}{\beta}}_0)\right) \overset{\text{asymt.}}{\sim} \chi^2_r \quad r = p_1 - p_0$ R: anova(mod0, mod1, test="Chisq") or drop1(mod1, test="Chisq")
Only H_1	Wald test: Under H_0, $(C\hat{\underset{\sim}{\beta}}_1 - \underset{\sim}{d})^T \left(C F^{-1}(\hat{\underset{\sim}{\beta}}_1) C^T\right)^{-1} (C\hat{\underset{\sim}{\beta}} - \underset{\sim}{d}) \overset{\text{asymp}}{\sim} \chi^2_r$ For the case $H_0: \beta_i = 0$ vs. $H_1: \beta_i \neq 0$ the Wald test simplifies to $\frac{\hat{\beta}_i^2}{\widehat{Var}(\hat{\beta}_i)} \sim \chi^2_1$ or $Z = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \sim N(0,1)$

summary(mod1) gives you
Wald-tests or "2-tests"

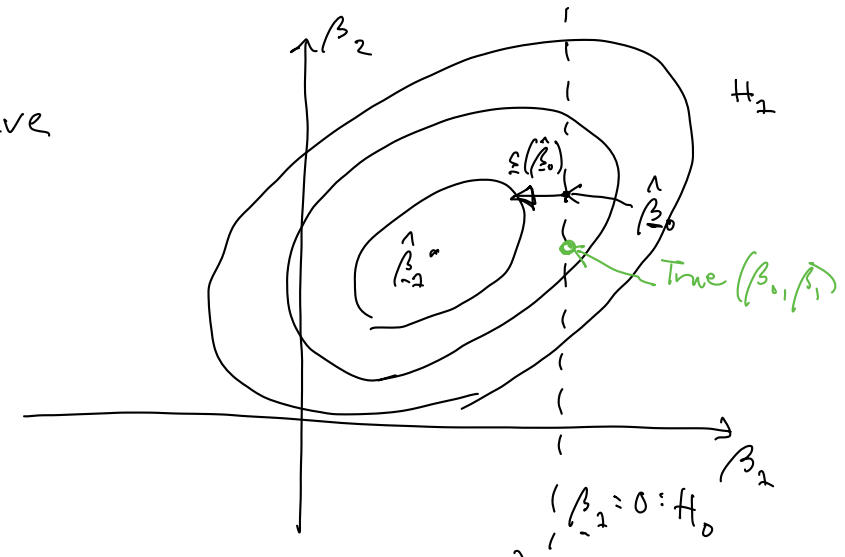
Only H_0

Score test (under H_0, $p-r$ parameters)

$$\underset{\sim}{S}_1(\hat{\beta}_0)^T \; \underset{1}{I}^{-1}(\hat{\beta}_0) \; \underset{\sim}{S}_1(\hat{\beta}_0) \overset{\text{asymp}}{\sim} \chi^2_r$$

$1 \times p \quad p \times p \quad p \times 1$

$\underset{\sim}{S}_1(\hat{\beta}_0)$ will have
r nonzero
elements



anova(mod0, test = "Rao")

Deviance (and testing goodness-of-fit)

Plays the same role as SSE of LMs.

Def.: The deviance D of a fitted model is as

$$D = 2\left(\ell_{saturated} - \ell(\hat{\underline{\beta}})\right)$$

where $\ell(\hat{\underline{\beta}})$ is the maximum log likelihood under the fitted model (p_0 parameters) and $\ell_{saturated}$ is the maximum log likelihood under a model with $p_1 = n$ parameters (n is no. of observations)

If the candidate fitted model is true, then

$$D \overset{asymp}{\sim} \chi^2_{n-p_0}$$

$$E(D) = n - p_0$$

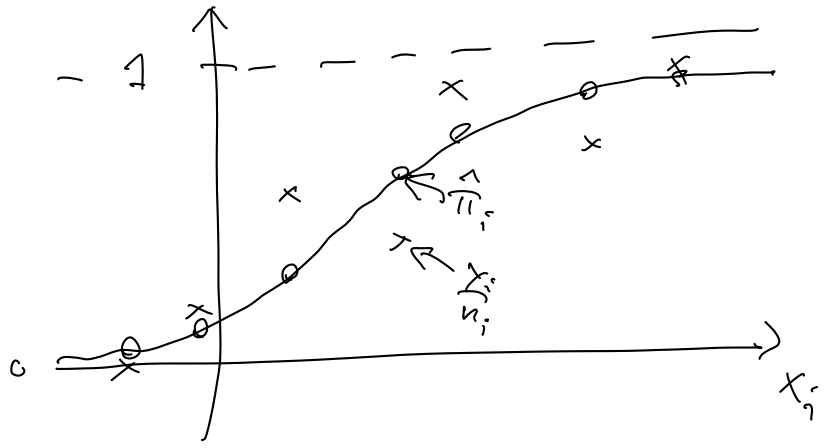
Reject H_0 if $D > \chi^2_{n-p_0,\alpha}$ (goodness-of-fit test

Deviance for binary regression model (grouped data)

$y_i \sim bin(n_i, \pi_i)$

$$D = 2\left[l\left(\frac{y_1}{n_1}, \frac{y_2}{n_2}, \dots, \frac{y_n}{n_n}\right) - l\left(\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_n\right)\right]$$

$$\hat{\pi}_i = h\left(x_i^T \hat{\beta}\right)$$



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

$$= 2\sum_{i=1}^{n} y_i \ln \frac{\bar{y}_i}{\hat{\pi}_i} + (n_i - y_i)\ln\frac{1-\bar{y}_i}{1-\hat{\pi}_i} \qquad \bar{y}_i = \frac{y_i}{n_i}$$

$$= \sum r_i^2$$

deviance residuals

Two types of residuals (binary regression)

Pearson residuals

$$r_i = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$$

Alternative goodness-of-fit statistic

$$\sum r_i^2 \overset{\text{approx}}{\sim} \chi^2_{n-p}$$

Deviance residuals:

$$r_i = \underbrace{\text{sign}(\bar{y}_i - \hat{\pi}_i)} \sqrt{2\left(y_i \ln \frac{\bar{y}_i}{\hat{\pi}_i} + (n_i - y_i) \ln \frac{1 - \bar{y}_i}{1 - \hat{\pi}_i}\right)}$$

$$\text{sign}(x) = \begin{cases} 1 & \text{for } x > 0 \\ 0 & \text{for } x = 0 \\ -1 & \text{for } x < 0 \end{cases}$$

$$E(r_i) \simeq 0, \quad \text{Var}(r_i) \simeq 1$$

Estimating the dispersion parameter (in quasilikelihood models)

$$\hat{\phi} = \frac{D}{n-p}$$

or

$$\hat{\phi} = \frac{1}{n-p} \sum \underbrace{r_i^2}_{= \text{pearson residual}}$$

} used by quasibinomial / quasipoisson

Asymptotic distribution of the LRT-statistic (Wood sec. 4.4)

Want to test

$$H_0: \underset{r\times 1}{\underline{R}}(\underline{\theta}) = \underline{0} \quad \left.\begin{matrix}\\\\\end{matrix}\right\} \; r \text{ nonlinear equality constraints}$$

vs.

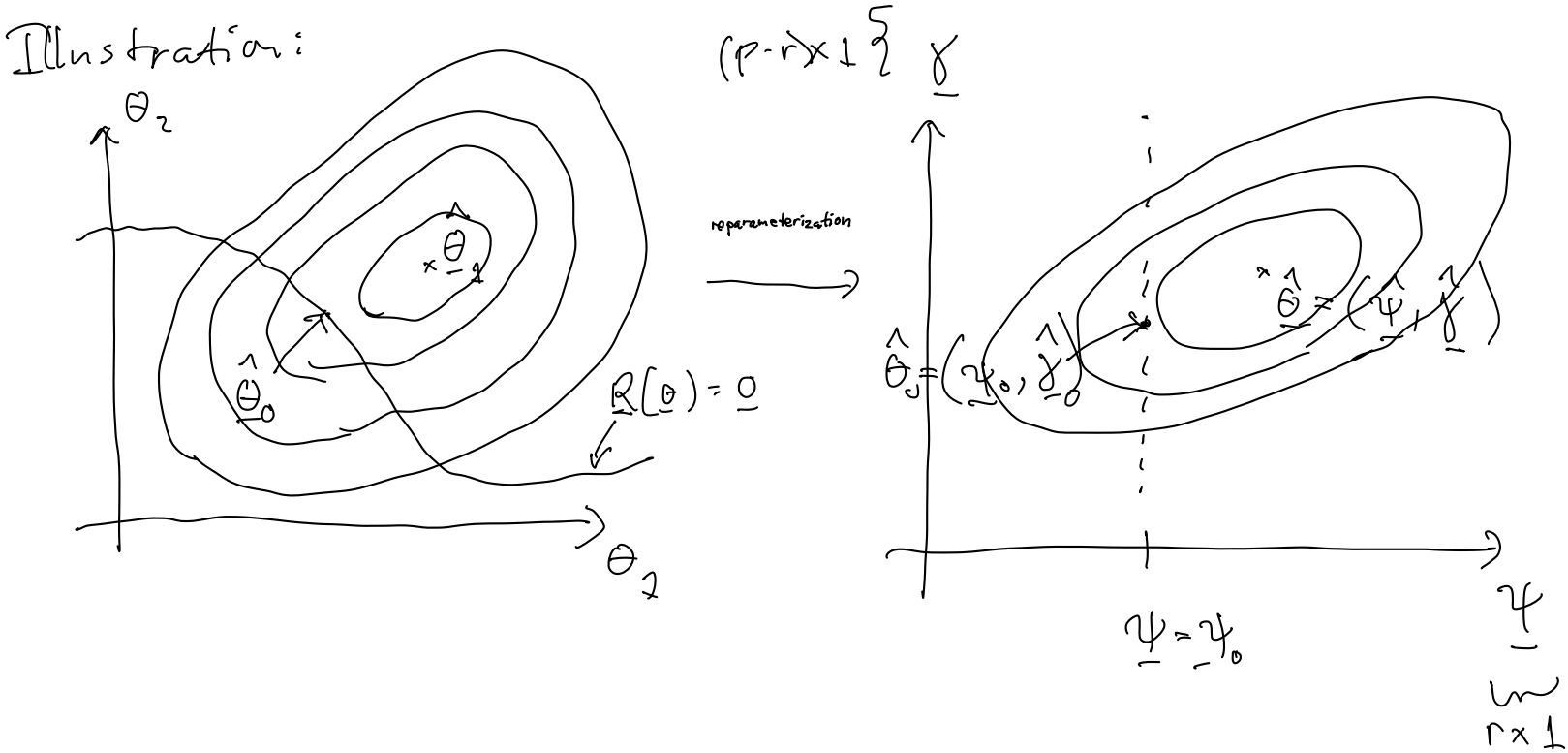
$$H_1: \underline{R}(\underline{\theta}) \neq \underline{0}$$

Theorem: Under H_0

$$LRT = 2\left[\ell(\hat{\underline{\theta}}_1) - \ell(\hat{\underline{\theta}}_0)\right] \overset{\text{asympt.}}{\sim} \chi^2_r$$

as $n \to \infty$ where $\hat{\underline{\theta}}_1$ and $\hat{\underline{\theta}}_0$ are the MLE of $\underline{\theta}$ under H_1 and H_0

Illustration:



Proof: Asymptotic approximation of $\ell(\underline{\theta})$ around the MLE $\hat{\underline{\theta}} = (\hat{\underline{\psi}}, \hat{\underline{\gamma}})$ under H_1:

$$\ell(\underline{\theta}) \simeq \ell(\hat{\underline{\theta}}) - \frac{1}{2}(\underline{\theta} - \hat{\underline{\theta}})^T H (\underline{\theta} - \hat{\underline{\theta}}), \qquad (1)$$

$$-\frac{1}{2}(\underline{\theta} - \hat{\underline{\theta}})^T H (\underline{\theta} - \hat{\underline{\theta}}) \qquad H = -\frac{\partial^2 \ell}{\partial \underline{\theta}\, \partial \underline{\theta}^T}$$

$$L(\underline{\theta}) \simeq L(\hat{\underline{\theta}})\, e$$

$$\propto MVN\left(\hat{\underline{\theta}}, \underbrace{\begin{bmatrix} \Sigma_{\psi\psi} & \Sigma_{\psi\gamma} \\ \Sigma_{\gamma\psi} & \Sigma_{\gamma\gamma} \end{bmatrix}}_{= H^{-1}}\right)$$

Hence, under H_0, $L(\underline{\theta})$ is maximised by $\hat{\underline{\theta}}_0 = (\underline{\psi}_0, \hat{\underline{\gamma}}_0)$ where

$$\hat{\underline{\gamma}}_0 = \hat{\underline{\gamma}} + \Sigma_{\gamma\psi} \Sigma_{\psi\psi}^{-1} (\underline{\psi}_0 - \hat{\underline{\psi}}) \qquad (2)$$

Analogous to the formula $E(\underline{x}_1 | \underline{x}_2) = \underline{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underline{x}_2 - \underline{\mu}_2)$ when $\begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \end{bmatrix} \sim MVN$. (wikipedia)

Since $\Sigma = H^{-1}$, partitions of these matrices satisfies

$$\begin{bmatrix} \Sigma_{\psi\psi} & \Sigma_{\psi\gamma} \\ \Sigma_{\gamma\psi} & \Sigma_{\gamma\gamma} \end{bmatrix} \begin{bmatrix} H_{\psi\psi} & H_{\psi\gamma} \\ H_{\gamma\psi} & H_{\gamma\gamma} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

which leads to 3 eqs. in $\Sigma_{\psi\psi}, \Sigma_{\gamma\psi} = \Sigma_{\psi\gamma}^T, \Sigma_{\gamma\gamma}$:

$$\Sigma_{\psi\psi} H_{\psi\psi} + \Sigma_{\psi\gamma} H_{\gamma\psi} = I \qquad \Sigma_{\psi\psi} H_{\psi\gamma} + \Sigma_{\psi\gamma} H_{\gamma\gamma} = 0$$

$$\Sigma_{\psi\psi} H_{\psi\psi} - \Sigma_{\psi\psi} H_{\psi\gamma} H_{\gamma\gamma}^{-1} H_{\gamma\psi} = I \qquad \Sigma_{\psi\gamma} = -\Sigma_{\psi\psi} H_{\psi\gamma} H_{\gamma\gamma}^{-1}$$

with solution

$$H_{\psi\psi} - H_{\psi\gamma} H_{\gamma\gamma}^{-1} H_{\gamma\psi} = \Sigma_{\psi\psi}^{-1} \quad (4) \qquad \Sigma_{\gamma\psi} = -H_{\gamma\gamma}^{-1} H_{\gamma\psi} \Sigma_{\psi\psi} \quad (3)$$

Substitute (3) into (2)

$$\hat{\underline{\gamma}}_0 = \hat{\underline{\gamma}} - H_{\gamma\gamma}^{-1} H_{\gamma\psi} \Sigma_{\psi\psi} \Sigma_{\psi\psi}^{-1} \left(\underline{\psi}_0 - \hat{\underline{\psi}} \right)$$

$$\hat{\underline{\gamma}}_0 - \hat{\underline{\gamma}} = -H_{\gamma\gamma}^{-1} H_{\gamma\psi} \left(\underline{\psi}_0 - \hat{\underline{\psi}} \right)$$

Using (1), the log likelihood under H_0 is then

$$\ell\left(\underline{\psi}_0, \hat{\underline{\gamma}}_0\right) \simeq \ell\left(\hat{\underline{\psi}}, \hat{\underline{\gamma}}\right) - \frac{1}{2} \underbrace{\begin{bmatrix} \psi_0 - \hat{\psi} \\ \hat{\gamma}_0 - \hat{\gamma} \end{bmatrix}^T H \begin{bmatrix} \psi_0 - \hat{\psi} \\ \hat{\gamma}_0 - \gamma \end{bmatrix}}_{= LRT}$$

and

$$LRT \sim \left(\underline{\psi}_0 - \hat{\underline{\psi}}\right)^T \begin{bmatrix} I \\ -H_{\gamma\gamma}^{-1} H_{\gamma\psi} \end{bmatrix}^T \begin{bmatrix} H_{\psi\psi} & H_{\psi\gamma} \\ H_{\gamma\psi} & H_{\gamma\gamma} \end{bmatrix} \begin{bmatrix} I \\ -H_{\gamma\gamma}^{-1} H_{\gamma\psi} \end{bmatrix} \left(\underline{\psi}_0 - \hat{\underline{\psi}}\right)$$

$$= \dots = (\underline{\psi}_0 - \hat{\underline{\psi}})^T \left(H_{\psi\psi} - H_{\psi\gamma} H_{\gamma\gamma}^{-1} H_{\gamma\psi} \right) (\underline{\psi}_0 - \hat{\underline{\psi}})$$

(4)

$$= (\underline{\psi}_0 - \hat{\underline{\psi}})^T \Sigma_{\psi\psi}^{-1} (\underline{\psi}_0 - \hat{\underline{\psi}})$$

As $n \to \infty$, $\mathrm{Var}(\hat{\underline{\theta}}) \simeq F^{-1}(\underline{\theta}) \simeq H^{-1}(\underline{\theta}) = \Sigma$

and in particular

$$\mathrm{Var}(\hat{\underline{\psi}}) \simeq \Sigma_{\psi\psi}$$

Hence, under H_0,

$$LRT = 2\left(\ell(\hat{\underline{\theta}}_1) - \ell(\hat{\underline{\theta}}_0) \right) \simeq (\underline{\psi}_0 - \hat{\underline{\psi}})^T \Sigma_{\psi\psi}^{-1} (\underline{\psi}_0 - \hat{\underline{\psi}}) \sim \chi^2_r$$

Numerical issues in computation of the deviance

(project 1, ~~prob. 1b~~)

Log likelihood:

$$\ell(\hat{\beta}) = \sum_{i=1}^{n} y_i \ln \hat{\pi}_i + (1-y_i) \ln(1-\hat{\pi}_i)$$

ok for computing $\ell(\hat{\beta})$ under H_0 but under the saturated model

Contribution to the log likelihood

$$\ell_i(\hat{\lambda}_i) = \ln f(y_i; \hat{\lambda}_i)$$

$$= \ln \frac{\hat{\lambda}_i^{y_i} e^{-\hat{\lambda}_i}}{y_i!}$$

$$= y_i \ln \hat{\lambda}_i - \hat{\lambda}_i - \ln y_i!$$

Under the saturated model $\hat{\lambda}_i = y_i$. For observation $y_i = 0$, $\hat{\lambda}_i = 0$ this leads to

$$\ell_i(\hat{\lambda}_i) = 0 \ln 0 - 0 - \ln 0!$$

$$= 0 \cdot \text{NaN} - 0 - \ln 1$$

$$= \text{NaN}$$

But

$$\ln f(0; 0) = \ln 1 = 0$$

dpois(0, 0, log = TRUE)

Nested models

Tests based on change in deviance

$$D_0 - D_1 = \ldots = LRT \sim \chi^2_{p_1 - p_0}$$

why?

Models selection

- Based on hypothesis testing
  - Include only significant covariates
  - Not recommended

- Choose the model with the smallest Akaike information criteria

$$AIC = -2\ell(\hat{\theta}) + 2p$$

$-2\ell(\hat{\theta})$: measures model fit

$2p$: penalize complex models

This is an estimate of Kullback-Leibler distance to the true model (next week)

16.19

21.9

Example:

```
library(ISwR)
data(eba1977)
...
```

} See R-code on course web page

Y_i (cases) Number of lung cancer cases

n_i (pop) sub pop. size

$j(i)$ (age) age categories (6 levels)

$k(i)$ (city) Fredericia, Horsens, Kolding, Vejle

Aim: Is there an elevated risk of lung cancer in Fredericia caused by petrochemical factory?

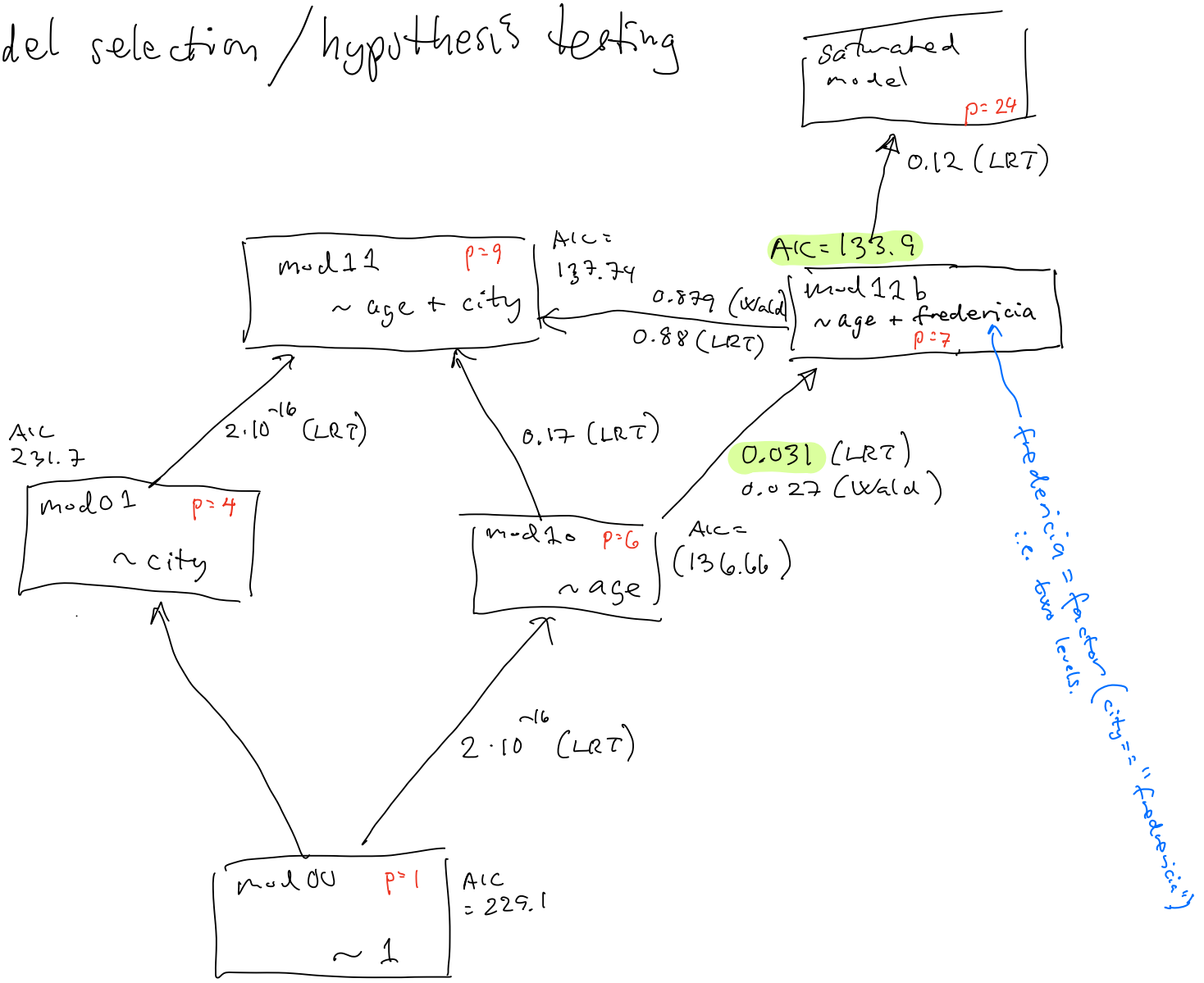
Model (model):

$$y_i \sim \text{bin}(n_i, \pi_i)$$

$$\text{logit}\,\pi_i = \mu + \underbrace{\alpha_{j(i)}}_{\text{effect of age}} + \underbrace{\beta_{k(i)}}_{\text{effect of city}}$$

$$\underline{\beta} = (\mu, \alpha_2, \alpha_3, \ldots, \alpha_6, \beta_2, \beta_3, \beta_4)^T \quad (p = 9 \text{ parameters})$$

Model selection / hypothesis testing



Want to fit the model $\beta_1 \neq \beta_2 = \beta_3 = \beta_4$ (mod12b)

Testing this via a Wald test

$$C\underline{\beta} = \underline{d}$$

$$\begin{bmatrix} 0 & \cdots & 0 & 0 & 1 & -1 & 0 \\ \vdots & \ddots & \vdots & \vdots & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Wald test: $(C\hat{\underline{\beta}} - \underline{d})^T (C F^{-1}(\hat{\underline{\beta}}) C^T)^{-1} (C\hat{\underline{\beta}} - \underline{d}) \sim \chi^2_2$

$\hat{\underline{\beta}}$: from mod11; $F^{-1}(\hat{\underline{\beta}})$: from mod11

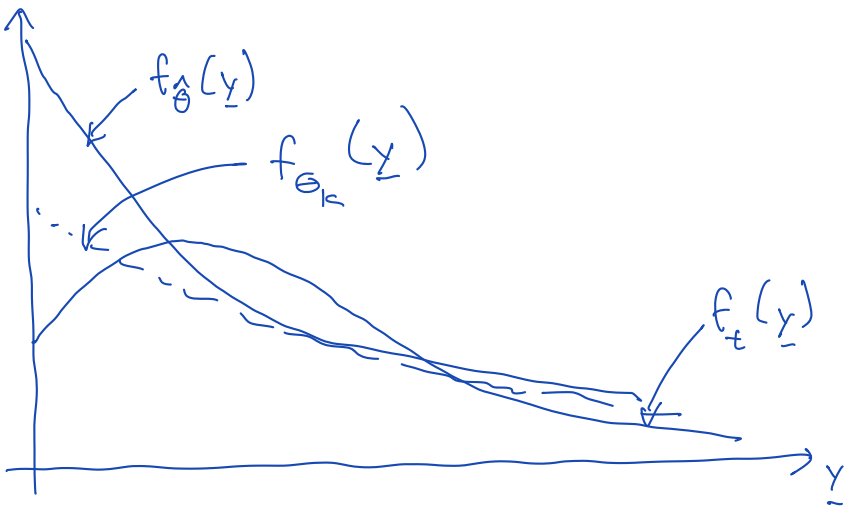
Theory behind AIC (Wood sec. 4.6)

Aim: Choose a candidate model $f_\theta(\underline{y})$ minimising the Kullback-Leibler distance to the true model $f_t(\underline{y})$

$$K(f_\theta, f_t) = \int \left(\ln f_t(\underline{y}) - \ln f_\theta(\underline{y})\right) f_t(\underline{y})\, d\underline{y} \qquad (1)$$

$$= E\left(\ln f_t(\underline{Y}) - \ln f_\theta(\underline{Y})\right), \quad \underbrace{\underline{Y} \sim f_t}_{\text{"}\underline{Y}\text{ distributed as a random variable with pdf (or pmf) } f_t\text{"}}$$

One-dimensional illustration:



Problems: f_t is unknown and θ is only estimated by $\hat\theta = \hat\theta(\underline{Y}_2)$, $\underline{Y}_2 \sim f_t$ and indep. of $\underline{Y}$. However,

$$E\,K(f_{\hat\theta}, f_t) = E\,E\left(\ln f_t(\underline{Y}) - \ln f_{\hat\theta(\underline{Y}_2)}(\underline{Y}) \,\middle|\, \underline{Y}_2\right)$$

can be estimated up to a constant only involving f_t as follows:

First consider $K(f_\theta, f_t)$: This is minimised by some $\theta = \theta_K$ satisfying

$$\frac{\partial}{\partial\theta}\int (\ln f_t(\underline{y}) - \ln f_\theta(\underline{y}) f_t(\underline{y})\, d\underline{y}\Big|_{\theta=\theta_k} = 0$$

$$\int \frac{\partial}{\partial\theta}\ln f_\theta(\underline{Y})\Big|_{\theta=\theta_k} f_t(\underline{y})\, d\underline{y} = 0 \qquad (2)$$

Approximating $\ln f_{\hat\theta}(\underline{Y})$ around θ_k,

$$\ln f_{\hat\theta}(\underline{y}) \simeq \ln f_{\theta_k}(\underline{Y}) + (\hat\theta-\theta_k)^T \frac{\partial}{\partial\theta}\ln f_\theta(\underline{Y})\Big|_{\theta=\theta_k} + \frac{1}{2}(\hat\theta-\theta_k)\frac{\partial^2 \ln f_\theta(\underline{Y})}{\partial\theta\,\partial\theta^T}\Big|_{\theta=\theta_k}(\hat\theta-\theta_k)$$

Substituting this into (1) we then find that

$$K(f_{\hat\theta(\underline{Y})}, f_t)$$

$$\simeq \int\Big(\ln f_t(\underline{y}) - \ln f_{\theta_k}(\underline{Y}) - (\hat\theta-\theta_k)^T \frac{\partial}{\partial\theta}\ln f_\theta(\underline{Y})\Big|_{\theta=\theta_k} + \frac{1}{2}(\hat\theta-\theta_k)\frac{\partial^2 \ln f_\theta(\underline{Y})}{\partial\theta\,\partial\theta^T}\Big|_{\theta=\theta_k}(\hat\theta-\theta_k)\Big) f_t(\underline{y})\, d\underline{y}$$

(1), (2)

$$= K(f_{\theta_k}, f_t) + \frac{1}{2}\underbrace{(\hat\theta-\theta_k)^T F_{\theta_k}(\hat\theta-\theta_k)}_{(*)} \qquad (3)$$

For f_θ sufficiently close to f_t

$$\hat\theta \overset{\text{approx}}{\sim} N(\theta_k, F_{\theta_k}^{-1})$$

such that $(*) \sim \chi^2_p$ and

$$E\,K(f_{\hat\theta}, f_t) \simeq K(f_{\theta_k}, f_t) + \frac{1}{2}p. \qquad (4)$$

Next consider

$$E\left(-\ell(\hat{\theta})\right) = E\left(-\ell(\theta_k) - \left(\ell(\hat{\theta}) - \ell(\theta_k)\right)\right)$$

$$\overset{(3)}{\simeq} E\left(-\ell(\theta_k)\right) - \frac{1}{2}p$$

$$= \int \left(\ln f_t(\underline{y}) - \ln f_{\theta_k}(\underline{y})\right) f_t(\underline{y})\, d\underline{y} - \frac{1}{2}p$$

$$- \int \ln f_t(\underline{y})\, f_t(\underline{y})\, d\underline{y}$$

$$= K(f_{\theta_k}, f_t) - \frac{1}{2}p - \int \ln f_t(\underline{y})\, f_t(\underline{y})\, d\underline{y} \qquad (5)$$

Eliminating $K(f_{\theta_k}, f_t)$ from (4) and (5) we obtain

$$EK(f_{\hat{\theta}}, f_t) \simeq E\left(-\ell(\hat{\theta})\right) + p + \text{constant term only involving } f_t(\underline{y})$$

Thus, an estimate of $EK(f_{\hat{\theta}}, f_t)$ (up to an unknown constant) is

$$EK(f_{\hat{\theta}}, f_t) \simeq -\ell(\hat{\theta}) + p,$$

and an estimate of twice of this is

$$AIC = -2\ell(\hat{\theta}) + 2p.$$

Note the similarity between $K(f_\theta, f_t)$ and the leave-out-one log-likelihood cross-validation score

$$CV = -\sum_{i=1}^{n} \ln f_{\hat{\theta}_{-i}}(y_i)$$

Poisson regression with identity non-canonical link function

Model

$$y_i \sim \text{Poisson}(\lambda_i), \quad \underbrace{\lambda_i = \eta_i}_{\text{identity link}}, \quad \eta_i = x_i^T \beta$$

Likelihood

$$L(\beta) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}.$$

Log-lik.

$$\ell(\beta) = \sum_{i=1}^{n} y_i \ln \lambda_i - \lambda_i - \ln y_i!$$

$$= \sum_{i=1}^{n} y_i \ln \eta_i - \eta_i - \ln y_i!$$

Score function

$$s(\beta) = \sum_{i=1}^{n} y_i \frac{1}{\eta_i} \frac{\partial \eta_i}{\partial \beta} - \frac{\partial \eta_i}{\partial \beta}$$

$$= \sum_{i=1}^{n} \left(\frac{y_i}{\eta_i} - 1 \right) x_i.$$

Recall that $\frac{\partial \eta_i}{\partial \beta} = \frac{\partial}{\partial \beta}(x_i^T \beta) = x_i$

Fisher info

$$F(\beta) = \text{Var}(s(\beta)) = \sum_{i=1}^{n} x_i \frac{\text{Var}(y_i)}{\eta_i^2} x_i^T$$

$$= \sum_{i=1}^{n} x_i \frac{\lambda_i}{\eta_i^2} x_i^T = X^T \text{diag}\left(\frac{1}{\lambda}\right) X$$

x_i: $p \times 1$, $\frac{\lambda_i}{\eta_i^2}$: 1×1, x_i^T: $1 \times p$; product: $p \times p$

$$\text{diag}\left(\frac{1}{\lambda}\right) = \begin{bmatrix} 1/\lambda_1 & 0 & \cdots & 0 \\ 0 & 1/\lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 1/\lambda_n \end{bmatrix}$$

Observed Fisher info:

$$H(\underline{\beta}) = -\frac{\partial^2}{\partial \underline{\beta}\,\partial \underline{\beta}^T} \ell(\underline{\beta}) = -\underbrace{\frac{\partial}{\partial \underline{\beta}}}_{p\times 1}\underbrace{\left(\frac{\partial}{\partial \underline{\beta}^T}\ell(\underline{\beta})\right)}_{1\times p} = -\frac{\partial}{\partial \underline{\beta}} s^T(\underline{\beta})$$

$$= -\sum \frac{\partial}{\partial \underline{\beta}}\left(\frac{y_i}{\eta_i} - 1\right) x_i^T$$

$$= +\sum \frac{y_i}{\eta_i^2}\frac{\partial \eta_i}{\partial \underline{\beta}} x_i^T$$

$$= +\sum \underbrace{\frac{y_i}{\eta_i^2} x_i x_i^T}_{p\times p} = X^T \operatorname{diag}\left(\frac{\underline{y}}{\underline{\eta}^2}\right) X = Z^T Z$$

$$\neq F(\underline{\beta})$$

Even if X has full rank $p < n$, $H(\underline{\beta})$ is possibly only semi-positive definite. Why?

Let

$$Z = \operatorname{diag}\left(\sqrt{\frac{\underline{y}}{\underline{\eta}^2}}\right) X$$

(rows of X gets multiplied by $\sqrt{y_i/\eta_i^2} = 0$ for some obs.)

Z may not have p independent rows (and columns) and $\operatorname{rank}(Z) \leq p$. Thus $H(\underline{\beta}) = Z^T Z$ may only be positive semi-definite and non-invertible since

$$(Z\alpha)^T(Z\alpha) = \|\underbrace{Z\alpha}_{\text{any linear comb. of columns in } Z.}\|^2 \geq 0$$

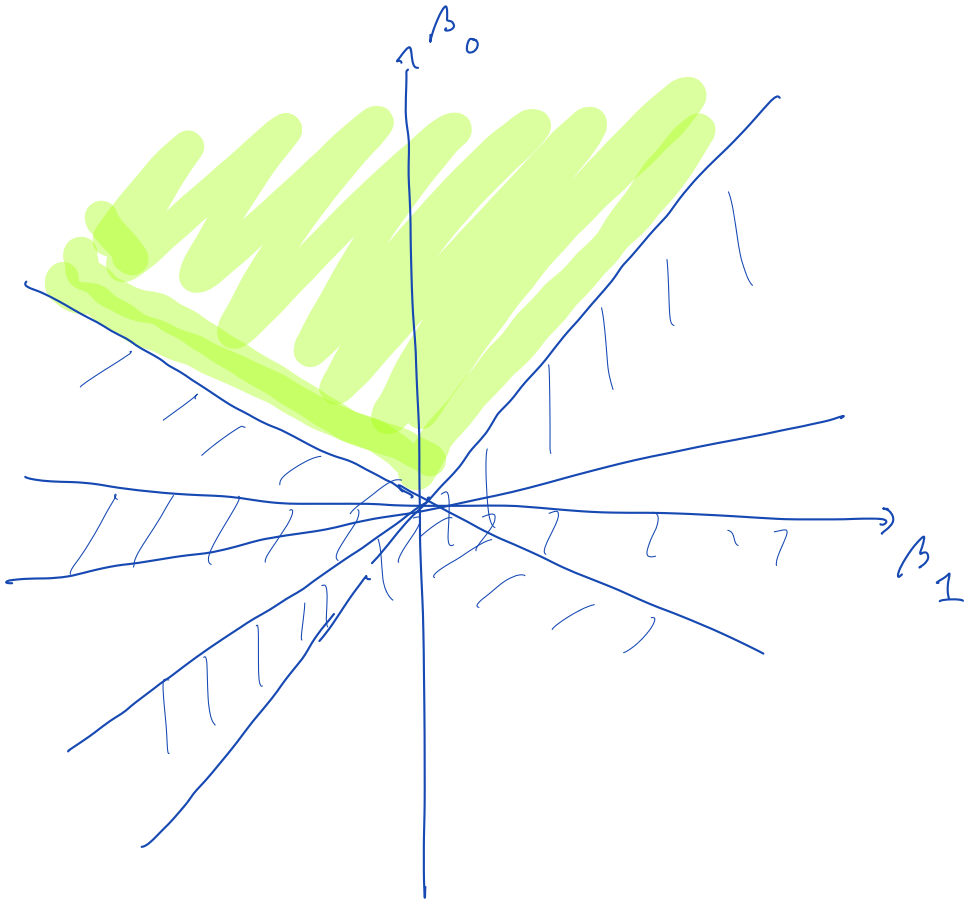
One reason for preferring Fisher scoring alg. over Newton's method.

Constraints on β: For $i = 1, 2, \ldots, n$ and $p = 2$

$$\lambda_i = \eta_i = \beta_0 + \beta_1 x_i \geq 0$$

$$\beta_0 \geq -\beta_1 x_i$$

$$\beta_0 \geq \max_i \left(-\beta_1 x_i\right) = \begin{cases} \beta_1 \max_i \left(-x_i\right) & \text{for } \beta_1 \geq 0 \\ -\beta_1 \max_i \left(x_i\right) & \text{for } \beta_1 \leq 0 \end{cases}$$



GLMs for positive continuous responses (Ch. 5.3)

Gamma

$$f(y|a,b) = \frac{b^a}{\Gamma(a)} y^{a-1} e^{-by}, \quad EY = \frac{a}{b}, \quad \text{Var}\, Y = \frac{a}{b^2}$$

Accomodates right skew.

Belongs to exponential family since

$$\ln f(y|a,b) = a \ln b + (a-1)\ln y - by - \ln\Gamma(a)$$

$$= -by + a\ln b + (a-1)\ln y - \ln\Gamma(a)$$

(θ, $b(\theta)$)

$$= \underbrace{\frac{-\frac{b}{a}y + \ln(b/a)}{1/a}}_{\theta,\ -b(\theta)} - \underbrace{\frac{\ln(1/a)}{1/a} - \ln\Gamma(a) + (a-1)\ln y}_{c(y,\phi,w)}$$

($\phi = 1/a$)

For the parameterization $\mu = \frac{a}{b}$ and $\nu = a$ then

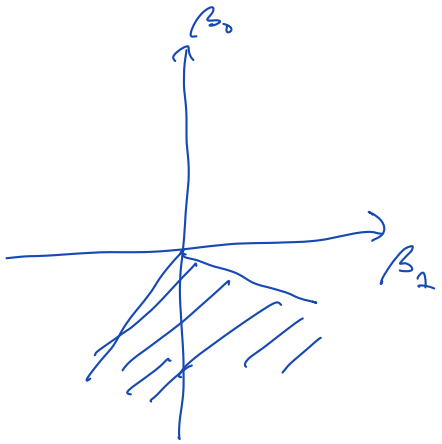
$$\theta = -\frac{1}{\mu} \quad \text{and} \quad b(\theta) = -\ln(-\theta) \qquad \phi = \frac{1}{a} = \frac{1}{\nu}$$

GLM with canonical link function

$$y_i \sim \text{Gamma}(\mu_i, \nu)$$

where

$$-\frac{1}{\mu_i} = \eta_i = \vec{x}_i^T \vec{\beta}$$



But

$$\mu_i > 0$$

$$-\frac{1}{\mu_i} = \eta_i = \vec{x}_i^T \vec{\beta} < 0$$

If $\eta_i = \beta_0 + \beta_1 x_i$, then $\beta_0 < \min_i(-\beta_1 x_i)$

Gamma GLM with non-canonical log-link

$$\log(\mu_i) = \eta_i = \underline{x}_i^T \underline{\beta}$$

$$y_i \sim \text{Gamma}(\mu_i, \nu)$$

ensures that

$$\mu_i = e^{\eta_i} = e^{\underline{x}_i^T \underline{\beta}} > 0 \quad \text{for all } \underline{\beta} \in \mathbb{R}^p$$

Note that $\ln Y_i$ becomes strongly left skewed (exercise 10)

Alternative lognormal GLM with a log-link

$$y_i \sim \text{lognormal}(\mu_i, \sigma)$$

$$\log(\mu_i) = \underline{x}_i^T \underline{\beta}$$

$$\Updownarrow$$

$$\ln y_i = \underline{x}_i^T \underline{\beta} + \varepsilon_i \quad (\text{a LM})$$

5.4. General treatment of GLMs

Log likelihood contribution on the exponential family form:

$$\ell_i(\underline{\beta}) = \ln f(y_i \mid \theta_i, \phi, w_i)$$

$$= \frac{w_i}{\phi}\left(y_i\theta_i - b(\theta_i)\right) + c(y_i, w_i, \phi).$$

Recall that

$$\mu_i = b'(\theta_i) \qquad (1)$$

and

$$\sigma_i^2 = \frac{\phi}{w_i} b''(\theta_i). \qquad (2)$$

We assume that

$$\mu_i = h(\eta_i). \qquad (3)$$

} possibly non-canonical link function.

Combining (1) and (3)

$$b'(\theta_i) = h(\eta_i). \qquad (4)$$

Implicit differentiation leads to

$$\frac{\partial}{\partial \underline{\beta}} b'(\theta_i) = \frac{\partial}{\partial \underline{\beta}} h(\eta_i)$$

$$b''(\theta_i)\frac{\partial \theta_i}{\partial \underline{\beta}} = h'(\eta_i)\frac{\partial \eta_i}{\partial \underline{\beta}}$$

such that

$$\frac{\partial \theta_i}{\partial \underline{\beta}} = \frac{h'(\eta_i)}{b''(\theta_i)}\frac{\partial \eta_i}{\partial \underline{\beta}} = \frac{h'(\eta_i)}{b''(\theta_i)}\underline{x}_i$$

$\eta_i = \underline{x}_i^T\underline{\beta}$

$\frac{\partial \eta_i}{\partial \underline{\beta}} = \underline{x}_i$

Contribution to score vector

$$\underset{\sim}{S}_i(\underset{\sim}{\beta}) = \frac{w_i}{\phi}\left(y_i \frac{\partial \theta_i}{\partial \underset{\sim}{\beta}} - b'(\theta_i)\frac{\partial \theta_i}{\partial \underset{\sim}{\beta}}\right)$$

$$= \frac{w_i}{\phi}(y_i - \mu_i)\frac{h'(y_i)}{b''(\theta_i)}\underset{\sim}{x}_i \qquad \underset{\sim}{x}_i : p\times 1$$

$$= \underset{\sim}{x}_i (y_i - \mu_i)\frac{h'(y_i)}{\sigma_i^2(y_i)}$$

Score vector

$$\underset{\sim}{S}(\underset{\sim}{\beta}) = \sum_{i=1}^{n} \underset{\sim}{S}_i(\underset{\sim}{\beta}) = \underset{p\times n}{X^T} D \Sigma^{-1}(\underset{\sim}{y} - \underset{\sim}{\mu})$$

where $\underset{n\times n}{D} = \operatorname{diag}(h'(\underset{\sim}{y})) = \begin{bmatrix} h'(y_1) & & \\ & h'(y_2) & \\ & & \ddots \\ & & & h'(y_n) \end{bmatrix}$ and $\underset{n\times n}{\Sigma} = \operatorname{diag}\left(\frac{1}{\underset{\sim}{\sigma}^2}\right) = \begin{bmatrix} 1/\sigma_1^2 & & \\ & 1/\sigma_2^2 & \\ & & \ddots \end{bmatrix}$

Fisher information:

$$\underset{p\times p}{\underset{\sim}{F}(\underset{\sim}{\beta})} = \sum_{i=1}^{n} \operatorname{Var}(\underset{\sim}{S}_i(\underset{\sim}{\beta}))$$

$$= \sum_{i=1}^{n} \underset{\sim}{x}_i \operatorname{Var}\left(\frac{h'(y_i)}{\sigma_i^2} y_i\right)\underset{\sim}{x}_i^T$$

$$= \sum_{i=1}^{n} \underset{\sim}{x}_i \underset{\sim}{x}_i^T \frac{(h'(y_i))^2}{\sigma_i^2}$$

$$= X^T W X \quad \text{where} \quad W = D^2\Sigma^{-1}$$

Fisher scoring algorithm

$$\underline{\beta}_{t+1} = \underline{\beta}_t + F(\underline{\beta}_t)^{-1} s(\underline{\beta}_t)$$

$$= (X^T W X)^{-1} X^T W X \underline{\beta}_t + (X^T W X)^{-1} X^T \underbrace{D \Sigma^{-1}}_{W D^{-1}} (\underline{y} - \underline{\mu})$$

$$= (X^T W X)^{-1} X^T W \left(X\underline{\beta}_t + D^{-1}(\underline{y} - \underline{\mu}) \right)$$

$$= (X^T W X)^{-1} X^T W \tilde{\underline{y}} \qquad (\hat{\beta}_{MLE} \text{ for general linear model})$$

where the "working observations"

$$\tilde{y}_i^{(t)} = \eta_i^{(t)} + \frac{y_i - h(\eta_i^{(t)})}{h'(\eta_i^{(t)})}$$

Iterated Reweighted Least Squares (IRLS)

Advantages:

- Instead of initial values $\underline{\beta}$ glm(, start)
  we can instead initialise $\underline{\mu} = h(\underline{\eta})$ glm(, mustart =)

  By default, initial values for $\underline{\mu}$ are computed by the $initialise function component of the family argument, returning

  $$\mu_i = \frac{y_i + 0.5}{n_i + 1}$$

  for binomial GLMs.

- Builds on efficient code computing

  $$\hat{\beta}_{t+1} = (X^T W X)^{-1} W \tilde{y}$$

  via the QR-decomposition of $W^{1/2} X$

  (e.g. via Gram-Schmidt process,

more stable numerically)

instead of via inversion of $X^T W X$ (numerically unstable)

Both have computational complexity $O(np^2)$

General linear model via QR-decomp:

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}, \quad \underline{\varepsilon} \sim N(\underline{0}, \Sigma), \quad \Sigma^{-1} = L^T L$$

$$\Sigma = (L^T L)^{-1}$$

Then

$$L\underline{y} = LX\underline{\beta} + L\underline{\varepsilon}$$

or

$$\underline{y}^* = X^*\underline{\beta} + \underline{\varepsilon}^*$$

where

$$\mathrm{Var}(\underline{\varepsilon}^*) = \mathrm{Var}(L\underline{\varepsilon}) = L\Sigma L^T = L(L^T L)^{-1} L^T = L L^{-1} L^{T^{-1}} L^T = \underline{I}$$

Hence, $\hat{\beta}$ satisfies

$$X^{*T}(y^* - X^*\hat{\underline{\beta}}) = \underline{0} \quad \text{(columns of } X^* \text{ orthogonal to } \hat{\varepsilon}^*)$$

Let

$$QR = X^*$$

denote QR-decomposition of X^* ($Q^T Q = 0$ and R upper triangular) (Gram-Schmidt)

Then

$$(QR)^T(y - QR\hat{\beta}) = 0$$

$$R^T Q^T y = R^T \underbrace{Q^T Q}_{=I} R\hat{\beta}$$

$$R\hat{\beta} = Q^T y$$

Quasi-likelihood (Ch. 5.5)

For real likelihoods from exponential family

$$S_i(\beta) = x_i \frac{h'(y_i)}{\sigma^2(\mu_i)}(y_i - \mu_i) \quad (*)$$

and

$$F_i(\beta) = x_i x_i^T \frac{(h'(y_i))^2}{\sigma^2(\mu_i)}$$

depends on $f(y_i|\mu_i,\ldots)$ only via μ_i and $\sigma_i^2 = \sigma^2(\mu_i)$

Idea: Accomodate overdispersion (and other relationships between σ_i^2 and μ_i^2 by changing σ_i^2 to $\sigma_i^2 = \phi\sigma^2(\mu_i)$. Does not necessarily correspond to any real underlying distribution.

Quasi-binomial:

$$S_i(\beta) = x_i \frac{h'(\eta_i)}{\phi\pi_i(1-\pi_i)} n_i(\bar{y}_i - \pi_i)$$

$$F_i(\beta) = x_i x_i^T \frac{n_i(h'(\eta_i))^2}{\phi\pi_i(1-\pi_i)}$$

$\bar{y}_i = \frac{y_i}{n_i}$

Quasi-Poisson

$$S_i(\beta) = x_i \frac{h'(\eta_i)}{\phi\lambda_i}(y_i - \lambda_i)$$

$$F_i(\beta) = x_i x_i^T \frac{(h'(\eta_i))^2}{\phi\lambda_i}$$

Note: The MLE of β remain unchanged but $\mathrm{Var}(\hat{\beta})$ is inflated by a factor ϕ.

Estimate of ϕ

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n} \frac{y_i - \hat{\mu}_i}{\sqrt{\sigma^2(\hat{\mu}_i)}} \quad \text{or alternatively} \quad \hat{\phi} = \frac{D}{n-p}$$

R: glm(y ~ x , family = quasipoisson(link = "log"))

Corresponds to a quasi-(log)likelihood with contributions defined as

$$Q_i(\mu_i; y_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi \sigma^2(t)}\, dt \qquad \text{(exercise 11)}$$

Note how $\frac{\partial Q}{\partial \beta}$ equals (*). We define quasi-deviance contribution as

$$D_i(\hat{\mu}_i; y_i) = -2\phi \left[Q_i(y_i; y_i) - Q_i(\hat{\mu}_i; y_i) \right]$$

$$= +2\phi\, Q_i(\hat{\mu}_i; y_i)$$

quasibinomial

$$= y_i \ln \frac{y_i}{\hat{\mu}_i} + (n_i - y_i) \ln \frac{1 - y_i}{1 - \hat{\pi}_i}$$

Given two nested models H_0 and H_1

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

Also

$$\frac{D_1}{\phi} \overset{\text{asymp}}{\sim} \chi^2_{n - p_1}$$

Under H_0

$$F = \frac{(D_1 - D_0)/(p_1 - p_0)}{\hat{\phi}}$$

$$= \frac{\left(\frac{D_1 - D_0}{\phi}\right)/(p_1 - p_0)}{\frac{D_1}{\phi}\Big/(n - p_2)} \quad \overset{\text{asymp}}{\sim} \quad F_{p_1 - p_0,\ n - p_2}$$

$\hat{\phi} = \phi$

see Venables & Ripley, eq. 7.10. (springer link)

R: drop1(mod, test = "F")

anova(mod0, mod1, test = "F")

Replaces test: "LRT" for non-quasi-likelihood models.

Alternative approaches:

Parametric models:

Beta-binomial: $y_i | p_i \sim \text{bin}(n_i, p_i)$

$p_i \sim \text{Beta}(\alpha, \beta)$

Negative binomial: $y_i | \lambda_i \sim \text{Poisson}(\lambda_i)$

$\lambda_i \sim \text{Gamma}(\alpha, \beta)$

GLMMs

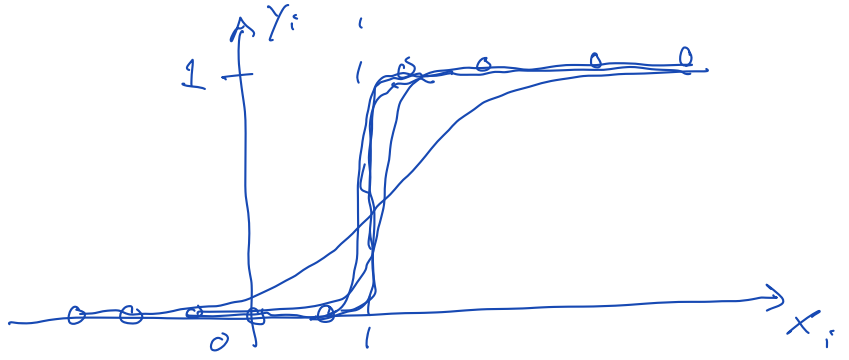
Models with random variables in the linear predictor

Linear separation

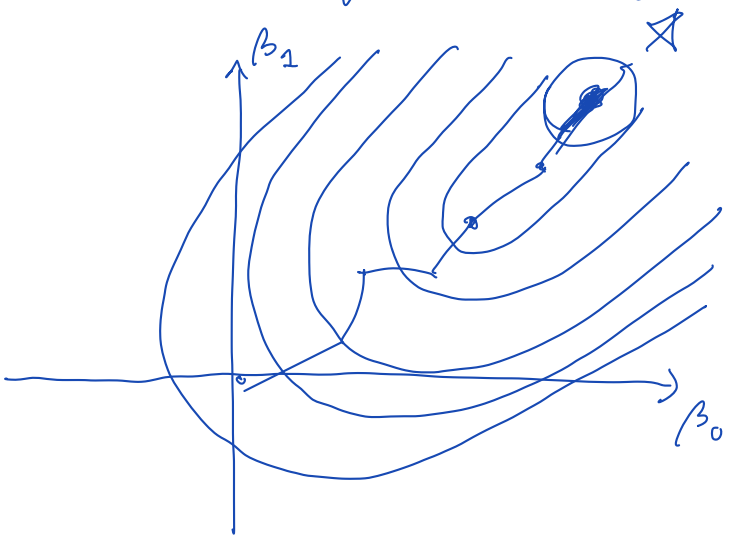
Sometimes encounted when working with sparse data, typically if there are a small number of observations for some levels of a factor.

Minimal example: $y_i \sim \text{bin}(1, \pi_i)$,

$\text{logit}\,\pi_i = \beta_0 + \beta_1 x_i$



The MLEs of β_0 and β_1 goes to infinity. $F(\underline{\beta})$ will have one eigen value very close to zero and $\text{Var}(\hat{\underline{\beta}}) = F^{-1}(\underline{\beta})$ will become extremely large.



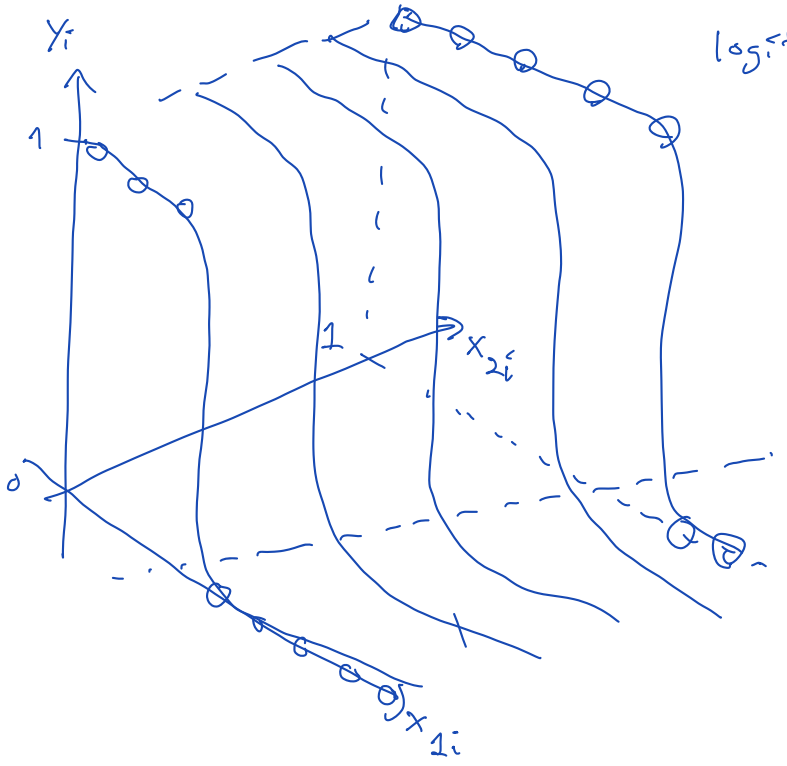
If we do a Wald-test based on $\text{Var}(\hat{\beta})$

$$Z = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$$

becomes very small, the Wald test can't be trusted.

$H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ can still be tested using $LRT = 2\left(\ell(\hat{\underline{\beta}}_1) - \ell(\hat{\underline{\beta}}_0)\right) \sim \chi^2_{p_1 - p_0}$

Example 2: Two covariates



$\text{logit}\,\pi_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$

No linear separation in terms of x_{i1} or x_{i2} alone.

A certain linear combination $L = a_1 x_{i1} + a_2 x_{i2}$ separates y_i into 0's and 1's.

Possible solutions:

Penalized regression: Lasso, ridge

Using GLMMs modelling effect associated with each level as random $\alpha_i \sim N(0, \sigma^2)$

Offset variabler

Additional term in linear prediction involving no unknown parameters

$$\eta_i = \underline{x}_i^T \underline{\beta} + o_i$$

R: glm(y ~ x , offset = o, ...)

1) Useful for likelihood ratio test of linear hypotheses $C\underline{\beta} = \underline{d}$, e.g. $H_0: \beta_1 = 1$ vs $H_1: \beta_1 \neq 1$

e.g. the Poisson regression model

$$y_i \sim \text{Poisson}(\lambda_i), \quad \lambda_i = c x_i^{\beta}$$

$$\ln \lambda_i = \beta_0 + \beta_1 \ln x_i$$

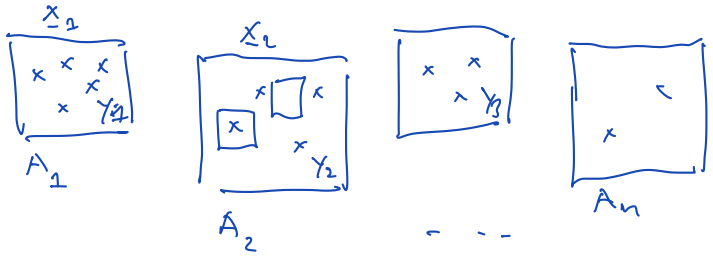
mod0 ⇖ H_0: glm(y ~ 1 , offset = log(x), poisson)

mod1 ⇖ H_1: glm(y ~ log(x), poisson)

anova (mod0, mod1, test = "Chisq")

(Also see exercise 2)

2) Poisson regression and rates



We expect $E(y_i) = A_i e^{\underline{x}_i^T \underline{\beta}}$

Model:

$$\lambda_i = A_i e^{\underline{x}_i^T \underline{\beta}}$$

$$\ln \lambda_i = x_i^T \beta + \ln A_i$$

offset term

Confidence intervals for regression coefficients or functions of β more generally

- Wald: We have

$$P\left(-z_{\alpha/2} < \underbrace{\frac{\hat{\beta}_j - \beta_j}{\widehat{SE}(\hat{\beta}_j)}}_{\text{asymp } \sim N(0,1)} < z_{\alpha/2} \right) \approx 1 - \alpha$$

pivotal quantity

Hence,

$$P\left(\hat{\beta}_j - \widehat{SE}(\hat{\beta}_j) z_{\alpha/2} < \beta_j < \hat{\beta}_j + \widehat{SE}(\hat{\beta}_j) z_{\alpha/2} \right) \approx 1 - \alpha$$

Thus

$$\hat{\beta}_j \pm \widehat{SE}(\hat{\beta}_j) z_{\alpha/2}$$

is a $(1-\alpha)$-conf. int. for β_j

- Profile likelihood confidence intervals

Recall: For a given value β_j of the jth component of $\underline{\beta}$

$$P\left(\underbrace{2\left(l(\hat{\beta}_j, \hat{\underline{\beta}}_{-j}) - l(\beta_j, \hat{\underline{\beta}}_{-j,0}) \right)}_{\text{pivotal quantity.}} < \chi^2_{\alpha,1} \right) \approx 1 - \alpha \quad (*)$$

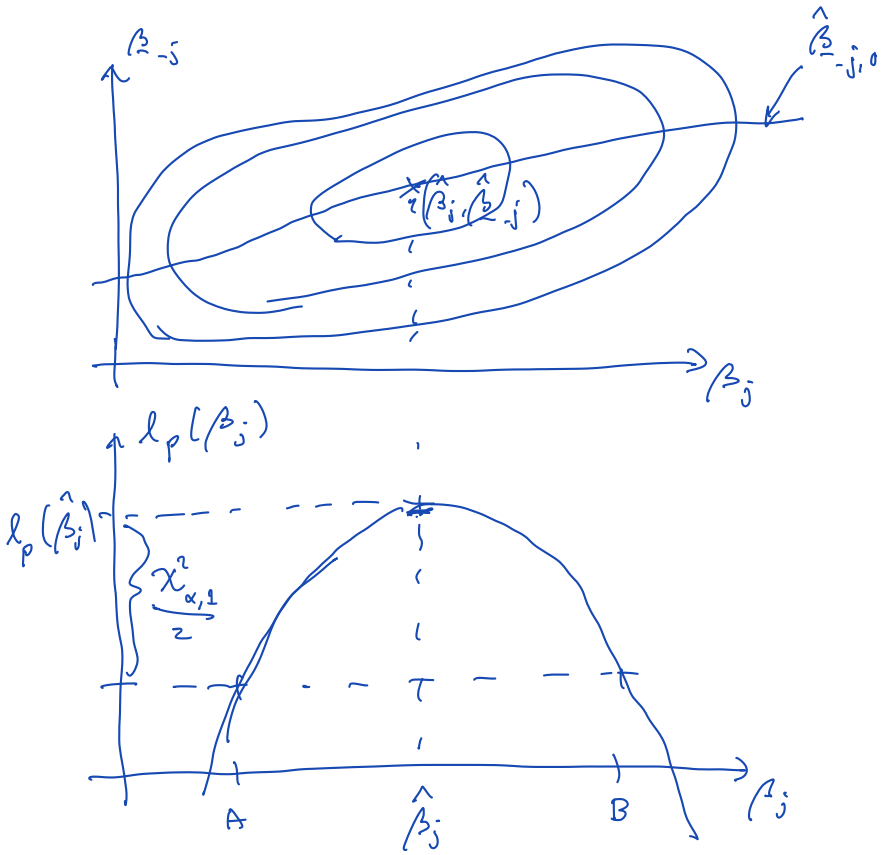
Letting $l_p(\beta_j) = \sup_{\underline{\beta}_{-j}} l(\beta_j, \underline{\beta}_{-j})$ denote the profile likelihood for β_j ("profiling out" $\underline{\beta}_{-j}$)

$(*)$ can be written

$$P\left(2\left(l_p(\hat{\beta}_j) - l_p(\beta_j) \right) < \chi^2_{\alpha,1} \right) \approx 1 - \alpha$$

$$P\left(l_p(\beta_j) > l_p(\hat{\beta}_j) - \frac{\chi^2_{\alpha,1}}{2} \right) \approx 1 - \alpha$$

$$P(A < \beta_j < B) \approx 1-\alpha$$



R: confint (mod) calls MASS::: confint.glm multiple times with $\beta_j x_{ij}$ included as an offset term to compute $l_p(\beta_j)$

Functional invariance:

If (A, B) is the profile (likelihood) conf. int for θ, then $(g(A), g(B))$ is the profile likelihood conf. int. for $g(\theta)$ provided that g is strictly monotonic.

Vector GLM (multinomial regression Ch. 6)

More than two categories $\{1, 2, \ldots, c, c+1\}$

$$\underset{c\times 1}{\underline{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_c \end{bmatrix} \sim \text{Multinomial}\left(m, \underset{c\times 1}{\underline{\pi}}\right)$$

Pmf of multinomial distn.

$$f(\underline{y} \mid \underline{\pi}) = \frac{m!}{y_1!\, y_2! \cdots y_c!\, (m - y_1 - \ldots - y_c)!} \pi_1^{y_1} \pi_2^{y_2} \cdots (1 - \pi_1 - \ldots - \pi_c)^{m - y_1 - \ldots - y_c}$$

$$E(\underline{y}) = m\underline{\pi} \qquad \underset{c\times c}{\text{Var}(\underline{y})} = m \begin{bmatrix} \pi_1(1-\pi_1) & & & \\ -\pi_1\pi_2 & \pi_2(1-\pi_2) & & \\ & & \ddots & \\ -\pi_1\pi_c & & & \pi_c(1-\pi_c) \end{bmatrix}$$

Data:

Response matrix

$$\underset{n\times c}{\begin{bmatrix} \underline{y}_1^T \\ \underline{y}_2^T \\ \vdots \\ \underline{y}_n^T \end{bmatrix}} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1c} \\ y_{21} & & & \\ \vdots & & & \\ y_{n1} & y_{n2} & \cdots & \end{bmatrix}$$

Design matrix

$$\begin{bmatrix} \underline{x}_1^T \\ \underline{x}_2^T \\ \vdots \\ \underline{x}_n^T \end{bmatrix} = \underset{n\times p}{\begin{bmatrix} x_{11} & x_{12} & \cdots \\ x_{21} & & \\ & & \\ & & \end{bmatrix}}$$

6.2. Models for unordered data (nominal multinomial regression)

Tentative model: Let $\pi_{ir} \propto e^{x_i^T \beta_r'}$ and normalize, i.e.

(β_r': specific to each category r.)

$$\pi_{ir} = \frac{e^{x_i^T \beta_r'}}{\sum_{s=1}^{c+1} e^{x_i^T \beta_s'}} \quad \Big\} \; p \times (c+1) \text{ parameters}$$

Adding a constant vector $\underline{c}$ to all the β_r''s doesn't change the π_{ir} (and the distr. of all the data) since

$$\pi_{ir}^* = \frac{e^{x_i^T(\beta_r' + \underline{c})}}{\sum_{s=1}^{c+1} e^{x_i^T(\beta_s' + \underline{c})}} = \pi_{ir}$$

Hence, the model is non-identifiable.

Identifiable model: Work with $\beta_r = \beta_r' - \beta_{c+1}'$. This leads to

$$\pi_{ir} = \frac{e^{x_i^T \beta_r}}{1 + \sum_{s=1}^{c} e^{x_i^T \beta_s}} \quad \text{for } r = 1, 2, \ldots, c$$

$$\pi_{i,c+1} = \frac{1}{1 + \sum_{s=1}^{c} e^{x_i^T \beta_s}} \quad \text{for } r = c+1$$

$\Big\}$ $p \times c$ parameters

- Odds ratio interpretation: A unit change in x_{ij}, changing $\underline{x}_i$ to $\underline{x}_i^* = \underline{x}_i + (0, 0, \underset{\uparrow}{1}, 0, \ldots, 0)^T$ (element j) changes the odds of event r_1 relative to r_2 by a

factor (an odds ratio) of

$$\frac{\pi^*_{ir_1} / \pi^*_{ir_2}}{\pi_{ir_1} / \pi_{ir_2}} = \frac{e^{\underline{x}_i^{*T}(\underline{\beta}_{r_1} - \underline{\beta}_{r_2})}}{e^{\underline{x}_i^{T}(\underline{\beta}_{r_1} - \underline{\beta}_{r_2})}} = e^{(\underline{x}_i^{*} - \underline{x}_i)^T(\underline{\beta}_{r_1} - \underline{\beta}_{r_2})}$$

$$= e^{\beta_{r_1,j} - \beta_{r_2,j}}$$

Example: Diet of alligators: 219 alligators
Response is diet (fish, invertebrates, reptiles, bird, other)

Two covariates
Size: numerical
Lake: 4 levels (see r-code)

Likelihood

$$L(\underline{\beta}) = \prod_{i=1}^{n} f(\underline{y}_i \mid n_i, \underline{\pi}_i)$$

($\underline{\pi}_i$ depends on $\underline{\beta}$)

$\underline{\beta}$: $cp \times 1$

$\underline{y}_i$: $c \times 1$

Log-likelihood

$$\ell(\underline{\beta}) = \sum_{i=1}^{n} \sum_{s=1}^{c+1} y_{is} \ln(\pi_{is}(\underline{\beta}))$$

Deviance

$$D = 2 \sum_{i=1}^{n} \sum_{s=1}^{c+1} y_{is} \ln\left(\frac{y_{is}/n_i}{\hat{\pi}_{is}}\right) \overset{\text{asymp/approx}}{\sim} \chi^2_{\underbrace{nc}_{p_1} - \underbrace{pc}_{p_0}}$$

LRTs / test based on change in deviance work as before.

Goodness-of-fit via observed deviance

- Latent utility interpretation of the model

Assume that individuals ("consumers") assess / perceive the utility of different alternatives ("products") $r = 1, 2, \ldots, c+1$ with some error

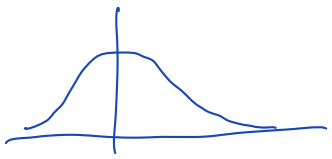
$$u_r = \underbrace{x_i^T \beta_r}_{\eta_r} + \varepsilon_r$$

$$\varepsilon_r \overset{iid}{\sim} \text{Gumbel}(0,1)$$

$$F_{\varepsilon_r}(x) = e^{-e^{-x}}$$

Gumbel(μ, σ) $\quad -\frac{x-\mu}{\sigma}$

$F(x) = e^{-e}$



and choose the alternative with the largest u_r.

Theorem (McFadden 1973): If then follows that

$$P(\text{choosing } r) = \frac{e^{\eta_r}}{\sum_{s=1}^{c+1} e^{\eta_s}}.$$

Proof:

$$P(\text{choosing } r) = P\left(u_r > \max_{s \neq r} u_s\right)$$

$$= P\left(u_r - \max_{s \neq r} u_s > 0 \right)$$

$u_r \sim \text{Gumbel}(\eta_r, 1)$; $\max_{s \neq r} u_s \sim ?$

Cdf of the maximum:

$$F_{\max_{s \neq r} u_s}(x) = P\left(\max_{s \neq r} u_s < x \right)$$

$$= P\left(\bigcap_{s \neq r} u_s < x \right)$$

$$= \prod_{s \neq r} P(u_s < x)$$

$$= \prod_{s \neq r} e^{-e^{-(x - \eta_s)}}$$

$$= e^{-x} \sum_{s \neq r} e^{\eta_s}$$

$$= e^{-e^{-(x - \ln \sum_{s \neq r} e^{\eta_s})}}$$

$$= e^{-e^{-(x - \mu)}}$$

that is, $\text{Gumbel}\left(\ln \sum_{s \neq r} e^{\eta_s}, 1 \right)$

If $X \sim \text{Gumbel}(0,1)$ then

$$F(x) = e^{-e^{-x}}$$

$$f(x) = \dots = e^{-x - e^{-x}}$$

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx - x - e^{-x}}\, dx$$

$u = e^{-x}$

$x = -\ln u$

$dx = -\frac{1}{u}\, du$

$$= \int_{\infty}^{0} e^{-(t-1)\ln u - u} \left(-\frac{1}{u}\right) du$$

$$= \int_0^{\infty} u^{-t+1} \frac{1}{u} e^{-u}\, du = \int_0^{\infty} u^{1-t-1} e^{-u}\, du$$

$$= \Gamma(1-t)$$

Gumbel(μ, σ):

$$M_{\mu+\sigma X}(t) = E e^{t(\mu+\sigma X)} = e^{\mu t} E e^{t\sigma X}$$

$$= e^{\mu t} M_X(t\sigma)$$

$$= e^{\mu t} \Gamma(1-\sigma t)$$

Hence,

$$M_{u_r - \max_{s \neq r} u_s}(t) = e^{\eta_r t} \Gamma(1-t)\, e^{-\left(\ln \sum_{s \neq r} e^{\eta_s}\right) t}\, \Gamma(1+t)$$

$$= e^{\left(\eta_r - \ln \sum_{s \neq r} e^{\eta_s}\right) t}\, \underbrace{\Gamma(1-t)\Gamma(1+t)}_{= B(1+t,\, 1-t)}$$

If $X \sim \text{logistic}(0,1)$ then

$$F(x) = \frac{e^x}{1+e^x}, \quad f(x) = \frac{e^x}{(1+e^x)^2} \qquad F(x) = \frac{e^{x-\mu}}{1+e^{x-\mu}}$$

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} \frac{e^x}{(1+e^x)^2}\, dx \qquad u = \frac{e^x}{1+e^x} \qquad x = \ln\frac{u}{1-u}$$

$$= \int_0^1 \cancel{u(1-u)} \left(\frac{u}{1-u}\right)^t \frac{du}{\cancel{u(1-u)}} \qquad dx = \frac{1-u}{u} \cdot \frac{1-u+u}{(1-u)^2}\, du = \frac{du}{u(1-u)}$$

$$= B(1+t, 1-t)$$

Logistic (μ, σ):

$$M_{\mu+\sigma X}(t) = E e^{t(\mu+\sigma X)}$$

$$= e^{t\mu} E\left(e^{t\sigma X}\right)$$

$$= e^{t\mu} M_X(t\sigma)$$

$$= e^{t\mu} B(1+\sigma t, 1-\sigma t)$$

Thus, $u_r - \max_{s \neq r} u_s \sim \text{logistic}\left(\eta_r - \ln \sum_{s \neq r} e^{\eta_s}, 1\right)$

and

$$P(\text{Choosing } r) = P\left(u_r - \max_{s \neq r} u_s > 0\right)$$

$$= 1 - F_{\text{logistic}(\ ,1)}(0)$$

$$= 1 - \frac{e^{0 - \eta_r + \ln \sum_{s \neq r} e^{\eta_s}}}{1 + e^{0 - \eta_r + \ln \sum_{s \neq r} e^{\eta_s}}}$$

$$= 1 - \frac{\sum_{s \neq r} e^{\eta_s} e^{-\eta_r}}{1 + \sum_{s \neq r} e^{\eta_s} e^{-\eta_r}}$$

$$= 1 - \frac{\sum_{s \neq r} e^{\eta_s}}{e^{\eta_r} + \sum_{s \neq r} e^{\eta_s}}$$

$$= \frac{e^{\eta_r}}{\sum_{s=1}^{c+1} e^{\eta_s}} \quad \text{(proving the theorem)}$$

$$= \frac{e^{\underline{x}_i' \underline{\beta}_r}}{\sum_{s=1}^{c+1} e^{\underline{x}_i' \underline{\beta}_s}} \quad \text{(as in multinomial logit model)}$$

Ordinal regression (6.3)

The response is an ordered categorical variable $Y_i \in \{1, 2, \ldots, c+1\}$ e.g. grades $(A, B, \ldots, F)$, a Likert scale (strongly disagree, disagree, indifferent, ..., strongly agree).

Model: The response Y_i is determined by an underlying latent variable

$$u_i = -\underline{x}_i^T \underline{\beta} + \varepsilon_i$$

where ε_i has a standard distribution with cdf F.

The event $Y_i = r$ occurs if

$$\theta_{r-1} < u_i \leq \theta_r$$

where $-\infty = \theta_0 < \theta_1 < \ldots < \theta_{c+1} = \infty$

Unknown parameters, $\theta_1, \theta_2, \ldots, \theta_c, \beta_1, \beta_2, \ldots, \beta_k$.

Then

$$P(Y_i \leq r) = P(u_i \leq \theta_r)$$

$$= P(-\underline{x}_i^T \underline{\beta} + \varepsilon_i \leq \theta_r)$$

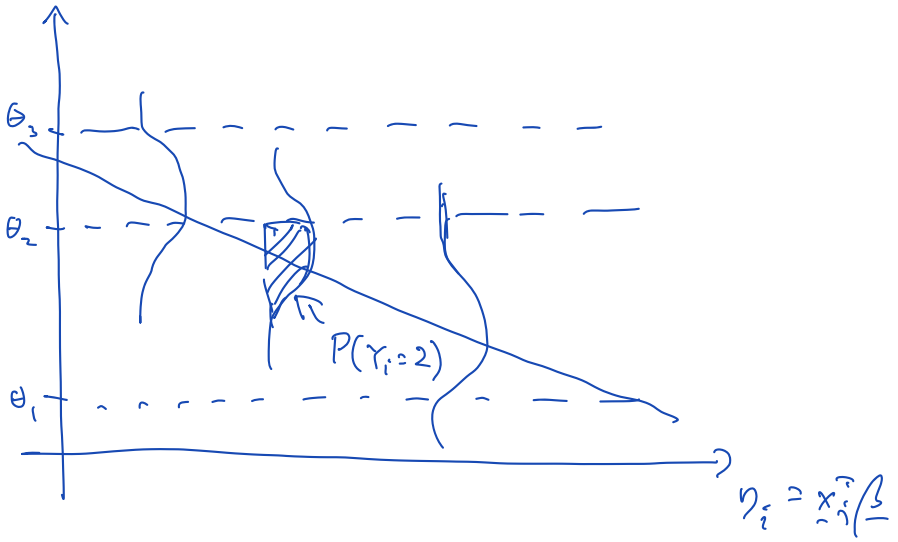
$$= P(\varepsilon_i \leq \theta_r + \underline{x}_i^T \underline{\beta})$$

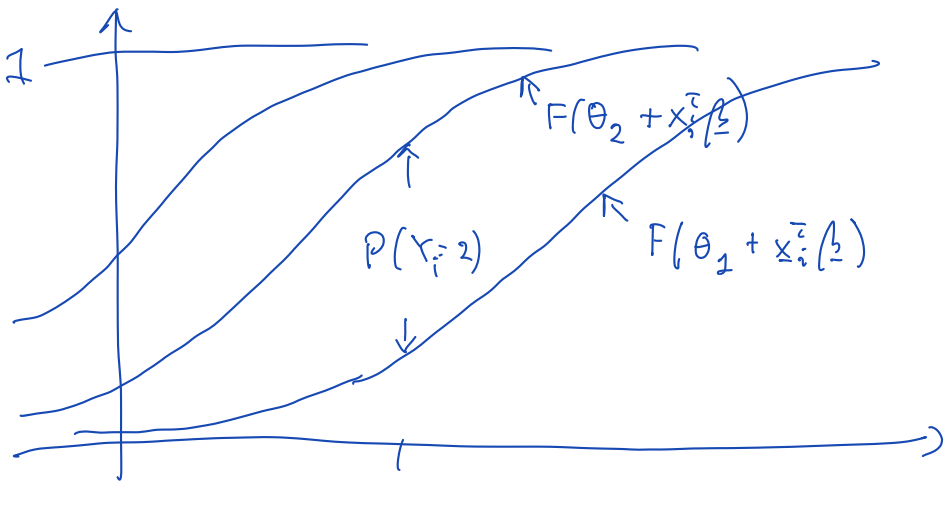
$$= F(\theta_r + \underline{x}_i^T \underline{\beta})$$

and

$$\pi_{ir} = P(Y_i = r) = P(Y_i \leq r) - P(Y_i \leq r-1)$$

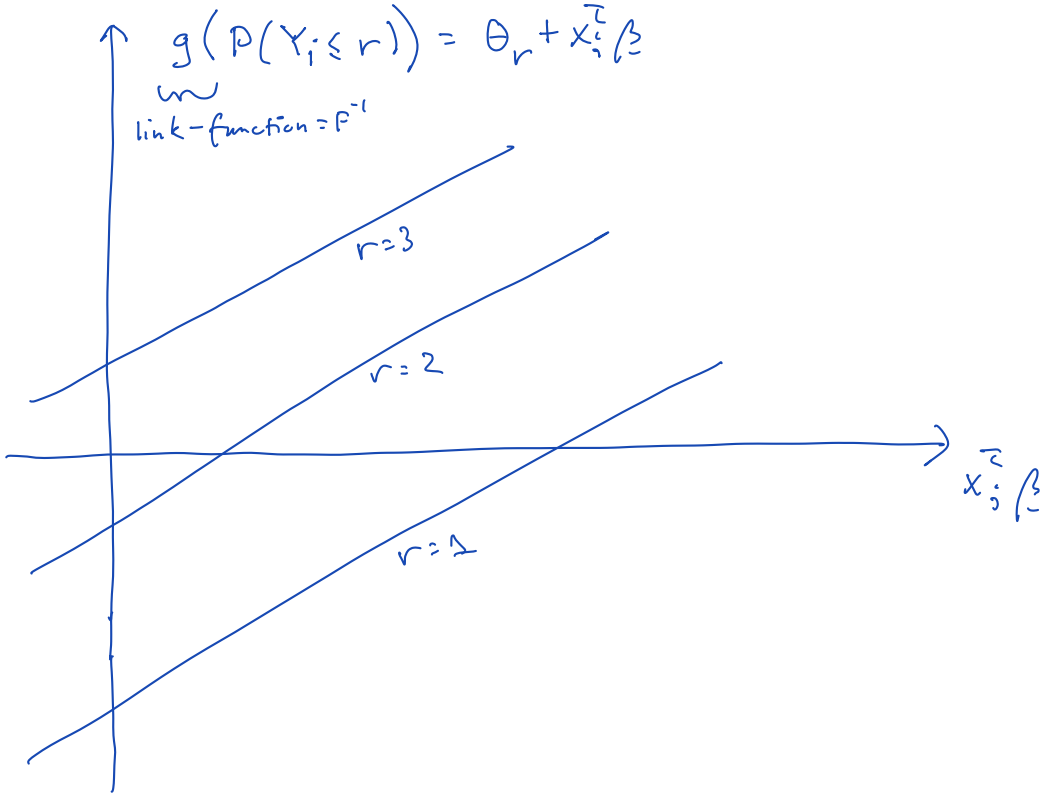
$$= F(\theta_r + \underline{x}_i^T \underline{\beta}) - F(\theta_{r-1} + \underline{x}_i^T \underline{\beta})$$





$\beta > 0 \Rightarrow E(Y_i)$ decreases with x_i

Parallelism assumption:



Proportional odds model:

For $F(x) = \frac{e^x}{1+e^x}$, that is, $\varepsilon_i \sim \text{logistic}(0,1)$

$$P(Y_i \leq r) = F(\theta_r + \underline{x}_i^T \underline{\beta}) = \frac{e^{\theta_r + \underline{x}_i^T \underline{\beta}}}{1 + e^{\theta_r + \underline{x}_i^T \underline{\beta}}}$$

$$\text{Odds}(Y_i \leq r) = \frac{P(Y_i \leq r)}{P(Y > r)} = \ldots = e^{\theta_r + \underline{x}_i^T \underline{\beta}} = e^{\theta_r} e^{\underline{x}_i^T \underline{\beta}}$$

Changing covariate x_{ij} by one unit changes all the odds of the cumulative events $Y_i \leq 1$, $Y_i \leq 2$, ...

by a common factor e^{β_j}, i.e. the odds of those events change proportionally.

Cumulative probit

$F(x) = \Phi(x)$ perhaps a more natural assumption

(discretisation of latent LM)

R: library(VGAM)

vglm(y ~ ... , family = cumulative(link = "probit", parallel = TRUE))

↑

factor(, ordered = TRUE)

Likelihood inference (nominal and ordinal multinomial models 6.4)

General notation:

$$\underset{c\times 1}{\underline{y}_i} \sim \text{Multinomial}(n_i, \underset{c\times 1}{\underline{\pi}_i})$$

Vector response function linking $\underline{\pi}_i$ to $\underline{\eta}_i$,

$$\underset{c\times 1}{\underline{\pi}_i} = \underline{h}(\underset{c\times 1}{\underline{\eta}_i})$$

where

$$\underset{c\times 1}{\underline{\eta}_i} = \underset{c\times p}{X_i}\,\underset{p\times 1}{\underline{\beta}}.$$

Example: Nominal model with k covariates has response function

$$\underset{\sim}{\pi}_i = \underset{\sim}{h}(\underset{\sim}{\eta}_i) = \underbrace{\begin{bmatrix} e^{\eta_{i1}} \\ e^{\eta_{i2}} \\ \vdots \\ e^{\eta_{ic}} \end{bmatrix}}_{c \times 1} \Big/ \left(1 + \sum_{s=1}^{c} e^{\eta_{is}}\right)$$

and

$$\underset{\sim}{\eta}_i = \begin{bmatrix} \eta_{i1} \\ \eta_{i2} \\ \vdots \\ \eta_{ic} \end{bmatrix} = \underbrace{\begin{bmatrix} \underset{\sim}{x}_i^T & 0 & \cdots & 0 \\ 0 & \underset{\sim}{x}_i^T & & \\ \vdots & & \ddots & \\ 0 & & & \underset{\sim}{x}_i^T \end{bmatrix}}_{c \times (k+1)c} \underbrace{\begin{bmatrix} \underset{\sim}{\beta}_1 \\ \underset{\sim}{\beta}_2 \\ \vdots \\ \underset{\sim}{\beta}_c \end{bmatrix}}_{(k+1)c \times 1}.$$

Example: Ordinal model:

$$\underset{\sim}{\pi}_i = \underset{\sim}{h}(\underset{\sim}{\eta}_i) = \begin{bmatrix} F(\eta_{i1}) - 0 \\ F(\eta_{i2}) - F(\eta_{i1}) \\ \vdots \\ 1 - F(\eta_{ic}) \end{bmatrix}$$

$$\underset{\sim}{\eta}_i = \begin{bmatrix} \eta_{i1} \\ \eta_{i2} \\ \vdots \\ \eta_{ic} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 & \underset{\sim}{x}_i^T \\ 0 & 1 & & \vdots & \underset{\sim}{x}_i^T \\ \vdots & & \ddots & \vdots & \vdots \\ 0 & \cdots & & 1 & \underset{\sim}{x}_i^T \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_c \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

$c \times (c+k)$ $\quad$ $(c+k) \times 1$

The log-likelihood contribution from obs. i becomes

$$\ell_i(\underline{\beta}) = \sum_{s=1}^{c} y_{is} \ln \pi_{is} + (n_i - y_{i1} - \dots - y_{ic}) \ln (1 - \pi_{i2} - \dots - \pi_{ic}) + C$$

This leads to the score contribution

$$\underline{s}_i(\underline{\beta}) = \frac{\partial}{\partial \underline{\beta}} \ell_i(\underline{\beta}) = \dots = \underset{p \times c}{\underline{X}_i^T} \underset{c \times c}{D_i} \underset{c \times c}{\Sigma_i^{-1}} \underset{c \times 1}{(\underline{y}_i - n_i \underline{\pi}_i)}$$

where

$$\Sigma_i = \underset{c \times c}{\mathrm{Var}(\underline{y}_i)} = n_i \underbrace{\left(\mathrm{diag}(\underline{\pi}_i) - \underset{c \times 1}{\underline{\pi}_i} \underset{1 \times c}{\underline{\pi}_i^T} \right)}_{c \times c}$$

Analogous to corresponding expression $\underline{s}_i(\underline{\beta}) = \underline{x}_i \frac{h'(\eta_i)}{\sigma_i^2} (y_i - \mu_i)$ for GLMs

and

$$D_i = \underset{c \times 1}{\frac{\partial \underline{h}^T}{\partial \underline{\eta}_i}} \underset{1 \times c}{\phantom{\frac{\partial \underline{h}^T}{\partial \underline{\eta}_i}}}$$

Contribution to Fisher information

$$F_i(\underline{\beta}) = \mathrm{Var}(\underline{s}_i(\underline{\beta})) = \dots = \underline{X}_i^T \underbrace{D_i \Sigma_i^{-1} D_i^T}_{W_i} \underline{X}_i$$

Letting

$$\underline{y} = \begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \\ \underline{y}_n \end{bmatrix}, \quad \underline{\pi} = \begin{bmatrix} \underline{\pi}_1 \\ \underline{\pi}_2 \\ \vdots \\ \underline{\pi}_n \end{bmatrix}, \quad N = \text{blockdiag}\left(n_1 \underline{1}_c, n_2 \underline{1}_c, \dots, \right)$$

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \quad \Sigma = \text{blockdiag}\left(\Sigma_1, \Sigma_2, \dots, \Sigma_n\right)$$

$$D = \text{blockdiag}\left(D_1, \dots \right)$$

$$W = \text{blockdiag}\left(W_1, \dots \right)$$

we can write

$$\underline{s}(\underline{\beta}) = \sum_{i=1}^{n} \underline{s}_i(\underline{\beta}) = X^T D \Sigma^{-1} \left(\underline{y} - N\underline{\pi}\right),$$

and

$$F(\underline{\beta}) = \sum_{i=1}^{n} F_i(\underline{\beta}) = X^T W X.$$

Fisher Scoring algorithm:

$$\underline{\beta}^{(t+1)} = \underline{\beta}^{(t)} + F\left(\underline{\beta}^{(t)}\right)^{-1} \underline{s}\left(\underline{\beta}^{(t)}\right)$$

Ordinal regression of chess game outcomes (project 2)

Example data:

Tournament with $k = 3$ players, $Y_i = \begin{cases} 1 & \text{white win} \\ 2 & \text{draw} \\ 3 & \text{black win} \end{cases}$, $n = \frac{k(k+1)}{2}$ games

Tentative model.

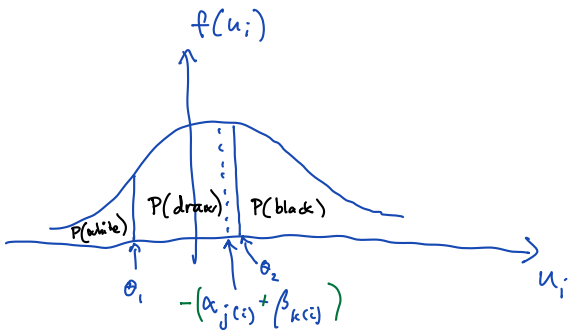
$$u_i = -\underset{\sim}{X}_i \underset{\sim}{\beta} + \varepsilon_i$$

$$-\left(\alpha_{j(i)} + \beta_{k(i)}\right) + \varepsilon_i \quad (*)$$

$\alpha_{j(i)}$: effect of player $j(i)$ having white pieces

$\beta_{k(i)}$: effect of player $k(i)$ playing black

$\varepsilon_i \sim N(0,1)$



vglm(y ~ white + black, cumulative( ))

Simpler model:

$$\alpha_j = -\beta_j \quad \text{for } j = 1, 2, \ldots, k$$

$$u_i = \sim \left(\alpha_{j(i)} - \alpha_{k(i)}\right) + \varepsilon_i \quad (**)$$

vglm(y ~ x2 + x3, ...)

Note also that (**) is a linear hypothesis

$$C\underset{\sim}{\beta} = \underset{\sim}{d}$$

about the parameters of (*).

Game i	Outcome Y_i	player with white pieces $j(i)$	player with black pieces $k(i)$
1	1	1	2
2	1	1	3
3	3	2	1
4	2	2	3
5	1	3	1
6	1	3	2

Model matrix for (*):

$$\begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \vdots \\ \alpha_2 \\ \alpha_3 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Model matrix for (**)

$$\begin{bmatrix} \cdots & -1 & 0 \\ & 0 & -1 \\ & 1 & 0 \\ & 1 & -1 \\ & 0 & 1 \\ & -1 & 1 \end{bmatrix} \begin{bmatrix} \vdots \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$$

x2 x3

model.matrix(~ white + black)[,-1]

Does this matrix have full rank?

7.10

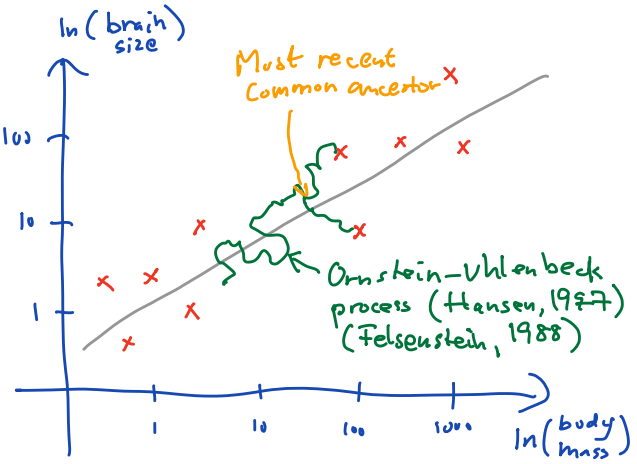
Mixed models: LMMs and GLMMs (Ch. 7)

19.10

General aim: Model and account for non-independent data, estimate dependency structure.

Examples:

- Repeated measurements on the same individuals / from the same group ("clustered data").
- Multiple grouping variables (crossed random effects)
- Temporal/spatial autocorrelation (time series analysis)
- Related individuals  (quantitative genetics)
- Related species



- Smoothing: Effect z_t of age t given by a second order random walk

$$(1-B)^2 z_t = w_t$$

Notation: For clusters $i = 1, 2, \ldots, m$ we measure the response Y_{ij} for different covariate vectors $\underline{x}_{ij}$ for $j = 1, 2, \ldots, n_i$

Random intercept model:

For $i = 1, 2, \ldots, m$, and $j = 1, 2, \ldots, n_i$

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_{0,i} + \varepsilon_{ij}$$

$\gamma_{0,i}$: random effect

where $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

and $\gamma_{0,i} \overset{iid}{\sim} N(0, \tau_0^2)$ (*)

Alternatively: Model $\gamma_{0,i}$ as a fixed effect?

Advantages of mixed model:

- We "borrow strength" between clusters. Higher statistical power (project 3) at the cost of making one more assumption (*)

R: library(lme4)

lmer(y ~ 1 + x + (1 | group))

grouping factor

Conditional model: Within cluster i observations are indep.

$$y_{ij} \mid \gamma_{0,i} \sim N(\beta_0 + \beta_1 x_{ij} + \gamma_{0,i}, \sigma^2)$$

Marginal model: Within clusters,

$$Var(y_{ij}) = Var(\beta_0 + \beta_1 x_{ij} + \gamma_{0,i} + \varepsilon_{ij})$$

$$= \tau_0^2 + \sigma^2$$

$$Cov(y_{ij}, y_{il}) = Cov(\beta_0 + \beta_1 x_{ij} + \gamma_{0,i} + \varepsilon_{ij}, \beta_0 + \beta_1 x_{ij} + \gamma_{0,i} + \varepsilon_{il})$$

$$= Cov(\gamma_{0,i}, \gamma_{0,i}) = Var(\gamma_{0,i}) = \tau_0^2$$

Intraclass correlation

$$\text{Corr}(y_{ij}, y_{ik}) = \frac{\tau_0^2}{\sigma^2 + \tau_0^2}$$

$$\underline{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix} \sim N\left(\underbrace{X_i}_{n_i \times 2} \underline{\beta}, \; \sigma^2 I_{n_i} + \tau_0^2 \underbrace{J_{n_i}}_{\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix}} \right)$$

If incorrectly using a LM (no random effect) we still have

$$E(\hat{\underline{\beta}}_{LM}) = E\left((X^TX)^{-1}X^T\underline{y}\right) = \ldots = \underline{\beta}, \quad (\text{why?})$$

but

$$\text{Var}(\hat{\underline{\beta}}_{LM}) > \sigma^2(X^TX)^{-1}$$

with non-independent data. Also $\hat{\beta}_{LM}$ not efficient.

Random intercept and slope models (7.1.2)

For $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_{0,i} + \gamma_{1,i} x_{ij} + \varepsilon_{ij}$$

$$= \left(\beta_0 + \gamma_{0,i}\right) + \left(\beta_1 + \gamma_{1,i}\right) x_{ij} + \varepsilon_{ij}$$

pop. mean intercept and slope

random deviations

where

$$\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

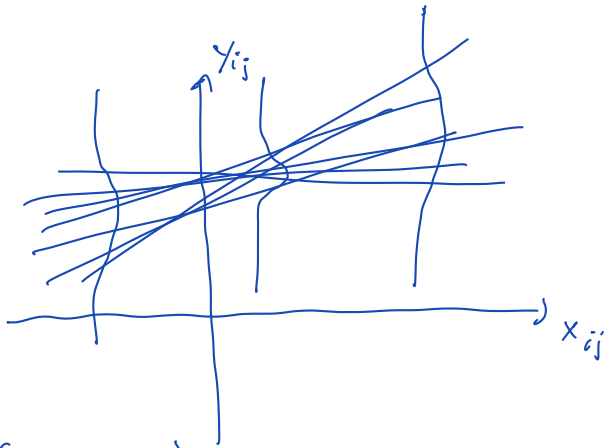
and

$$\underline{\gamma}_i = \begin{bmatrix} \gamma_{0,i} \\ \gamma_{0,i} \end{bmatrix} \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \underbrace{\begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{bmatrix}}_{Q} \right)$$

Marginal

$$Var(y_{ij}) = \tau_0^2 + \tau_1^2 x_{ij}^2 + 2 x_{ij} \tau_{01} + \sigma^2$$

(quadratic in x_{ij})



$$Cov(y_{ij}, y_{il}) = Cov\left(\gamma_{0,i} + \gamma_{1,i} x_{ij} + \varepsilon_{ij},\ \gamma_{0,i} + \gamma_{1,i} x_{il} + \varepsilon_{il} \right)$$

$$= \tau_0^2 + \tau_1^2 x_{ij} x_{il} + \tau_{01}(x_{ij} + x_{il})$$

R: `lmer(y ~ 1 + x + (1 + x | group))`

`1 + x`: fixed effect part

`(1 + x | group)`: random intercept and slope

More general notation (7.2.3–7.1.4, 7.2)

Data $(y_{ij} \mid \underline{x}_{ij}^T)$ for $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n_i$

$$y_{ij} = \underline{x}_{ij}^T \underline{\beta} + \underline{u}_{ij}^T \underline{\gamma}_i + \varepsilon_{ij}$$

Example: Random intercept slope model

$$y_{ij} = \begin{bmatrix} 1 & x_{ij} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} 1 & x_{ij} \end{bmatrix} \begin{bmatrix} \gamma_{0,i} \\ \gamma_{1,i} \end{bmatrix} + \varepsilon_{ij}$$

Letting

$$\underline{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix}, \quad X_i = \begin{bmatrix} \underline{x}_{i1}^T \\ \underline{x}_{i2}^T \\ \vdots \end{bmatrix}, \quad U_i = \begin{bmatrix} \underline{u}_{i1}^T \\ \underline{u}_{i2}^T \\ \vdots \end{bmatrix}, \quad \underline{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \end{bmatrix}$$

then the model is (for cluster $i = 1, 2, \ldots, m$)

$$\underline{y}_i = X_i \underline{\beta} + U_i \underline{\gamma}_i + \underline{\varepsilon}_i, \quad \underline{\gamma}_i \sim N(0, Q)$$

The marginal model is then

$$\underline{y}_i \sim N\left(X_i \underline{\beta}, \; U_i Q U_i^T + \sigma^2 I_{n_i} \right)$$

Conditional model

$$\underline{y}_i \mid \underline{\gamma}_i \sim N\left(X_i \underline{\beta} + U_i \underline{\gamma}_i, \; \sigma^2 I_{n_i} \right)$$

Letting

$$\underline{y} = \begin{bmatrix} \underline{y}_1 \\ \underline{y}_2 \\ \vdots \\ \underline{y}_m \end{bmatrix}, \quad \underline{\varepsilon} = \begin{bmatrix} \underline{\varepsilon}_1 \\ \underline{\varepsilon}_2 \\ \vdots \\ \underline{\varepsilon}_m \end{bmatrix}, \quad \underline{\gamma} = \begin{bmatrix} \underline{\gamma}_1 \\ \underline{\gamma}_2 \\ \vdots \\ \underline{\gamma}_m \end{bmatrix}, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \end{bmatrix},$$

$$U = \text{blockdiag}\left(U_1, U_2, \ldots, U_m\right),$$

the the whole model is

$$\underline{y} = X\underline{\beta} + U\underline{\gamma} + \underline{\varepsilon}, \quad \underline{\varepsilon} \sim N\left(0, \overbrace{\sigma^2 \underline{I}_{\Sigma n_i}}^{=R}\right)$$

where

$$\underline{\gamma} \sim N(0, G), \quad G = \text{blockdiag}\left(Q_1, Q_2, \ldots, Q_m\right).$$

Conditional model

$$\underline{y} \mid \underline{\gamma} \sim N\left(X\underline{\beta} + U\underline{\gamma}, R\right)$$

Marginal

$$\underline{y} \sim N\left(X\underline{\beta}, \underbrace{UGU^T + R}_{V = V(\underline{\theta})}\right)$$

Random intercept and slope model, general notation

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{m1} \\ y_{m2} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{12} \\ & x_{13} \\ & x_{21} \\ & x_{22} \\ & \\ 1 & \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & x_{11} & & & & \\ 1 & x_{12} & & & & \\ 1 & x_{13} & & & & \\ & & 1 & x_{21} & & \\ & & 1 & x_{22} & & \\ & & & & \ddots & \\ & & & & 1 & x_{m1} \\ & & & & 1 & x_{m2} \end{bmatrix} \begin{bmatrix} \gamma_{0,1} \\ \gamma_{1,1} \\ \gamma_{0,2} \\ \gamma_{1,2} \\ \vdots \\ \gamma_{0,m} \\ \gamma_{1,m} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \vdots \\ \end{bmatrix}$$

$n \times 1$ $\quad$ $n \times 2$ $\quad$ 2×1 $\quad$ $n \times 2m$ $\quad$ $2m \times 1$ $\quad$ $n \times 1$

$$n = \sum_{i=1}^{m} n_i$$

$$\underline{\gamma} \sim N(0, G), \quad G = \begin{bmatrix} \tau_0^2 & \tau_{01} & & & \\ \tau_{01} & \tau_1^2 & & & \\ & & \tau_0^2 & \tau_{01} & \\ & & \tau_{01} & \tau_1^2 & \\ & & & & \ddots \end{bmatrix}, \quad \underline{\varepsilon} \sim N(0, R), \; R = \sigma^2 I_n$$

$$\underline{\theta} = (\sigma^2, \tau_0^2, \tau_1^2, \tau_{01}^2)^T$$

19.10

General linear model (Ch. 4.1)

21.10

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}, \quad \underline{\varepsilon} \sim N(\underline{0}, \sigma^2 W^{-1})$$

Multiplying by $W^{1/2}$

$$W^{1/2}\underline{y} = W^{1/2}X\underline{\beta} + W^{1/2}\underline{\varepsilon}$$

$$\underline{y}^* = X^*\underline{\beta} + \underline{\varepsilon}^*$$

where $\mathrm{Var}(\varepsilon^*) = \mathrm{Var}(W^{1/2}\sigma W^{-1}W^{1/2}) = \sigma^2 I_n$

Hence the MLE of $\underline{\beta}$ is

$$\hat{\underline{\beta}}_{GLS} = (X^{*T}X^*)^{-1}X^{*T}y^*$$
$$= (X^TWX)^{-1}X^TWy$$

$\mathrm{Var}(\hat{\beta}_{OLS})$ is less efficient than $\mathrm{Var}(\hat{\beta}_{GLS})$ in the sense that

$\mathrm{Var}(\hat{\beta}_{OLS}) - \mathrm{Var}(\hat{\beta}_{GLS})$

is positive semi-definite (exercise 11).

7.3. ML and REML inference

For $V = V(\theta)$ and θ known the MLE of $\underline{\beta}$ is

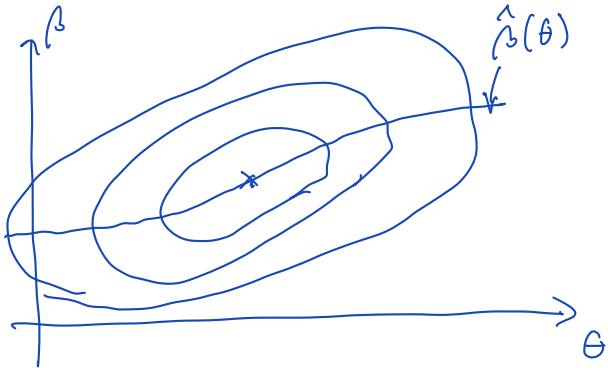
$$\hat{\underline{\beta}} = (X^TV^{-1}(\underline{\theta})X)^{-1}X^TV^{-1}(\underline{\theta})\underline{y}$$

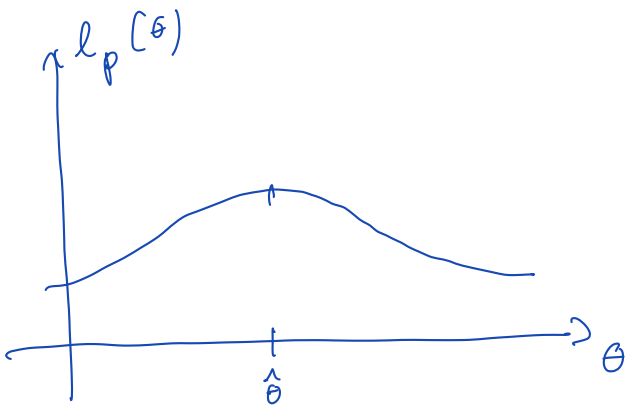
The profile log-likelihood for $\underline{\theta}$ is

$$l_p(\underline{\theta}) = \max_{\underline{\beta}} l(\underline{\theta}, \underline{\beta}) = l(\underline{\theta}, \hat{\underline{\beta}}(\underline{\theta}))$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|V(\underline{\theta})| + \frac{1}{2}(\underline{y} - X\hat{\underline{\beta}}(\underline{\theta}))^TV(\underline{\theta})^{-1}(\underline{y} - X\hat{\underline{\beta}}(\underline{\theta}))$$

Maximising this numerically (e.g. using optim in R) we obtain the joint MLE of $\underline{\beta}, \underline{\theta}$, $\hat{\underline{\beta}}(\hat{\underline{\theta}}), \hat{\underline{\theta}}$





But $\hat{\theta}_{MLE}$ is typically not unbiased.

Restricted / residual / reduced maximum likelihood (REML)
(Patterson & Thompson 1971)

Motivating example: Suppose $y_1, y_2, \dots, y_n \overset{iid}{\sim} N(\mu, \sigma^2)$

MLE of σ^2 if μ is known, $\hat{\sigma}^2 = \frac{1}{n}\sum(y_i - \mu)^2$,
is unbiased. But if estimating μ by $\hat{\mu} = \bar{y}$,
the joint MLE of σ^2, $\hat{\sigma}^2 = \frac{1}{n}\sum(y_i - \bar{y})^2$
is not unbiased. (squared deviation from $\bar{y}$ smaller than from μ)

Idea: Instead of the likelihood of $\underline{y} = (y_1, y_2, \dots, y_n)^T$
we will work with the likelihood of
$n-p$ "error contrasts" $\underline{w} = A^T\underline{y}$ defined such
that A is $n \times (n-p)$, $A^TX = 0$, $A^TA = I_{n-p}$

$S = AA^T = I - H = I - X(X^TX)^{-1}X^T$

The reduced data

$$\underline{w} = A^T \underline{y} = A^T(X\underline{\beta} + U\underline{\gamma} + \underline{\varepsilon})$$

$$= \cancel{A^T X \underline{\beta}} + A^T(U\underline{\gamma} + \underline{\varepsilon})$$

$A^T X = 0$

becomes dense and expensive to invert in general?

$$\sim N(\underline{0}, A^T V(\underline{\theta}) A)$$

i.e. the distribution of $\underline{w}$ doesn't depend on $\underline{\beta}$.

Example cont.: $\underline{y} \sim N(\mu, \sigma^2 I_n)$. We then have

$$\mathrm{Var}(\underline{w}) = \mathrm{Var}(A^T \underline{y}) = \sigma^2 A^T A = \sigma^2 I_{n-p}$$

$$E(\underline{w}) = E(A^T \underline{y}) = E(A^T(X\underline{\beta} + U\underline{\gamma} + \underline{\varepsilon})) = 0$$

Restricted log likelihood

$$l_R(\sigma^2) = -\frac{n-1}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\sigma^2 I_{n-p}\right| - \frac{1}{2\sigma^2}(A^T\underline{y})^T A^T \underline{y}$$

$$= -\frac{n-1}{2}\ln(2\pi) - \frac{n-1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\underline{y}^T(AA^T)\underline{y}$$

$$= \text{---}\,''\,\text{---} - \frac{1}{2\sigma^2}\underline{y}^T(I - H)\underline{y}$$

$$\left((I-H)\underline{y}\right)^T (I-H)\underline{y}$$

$$\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}}$$

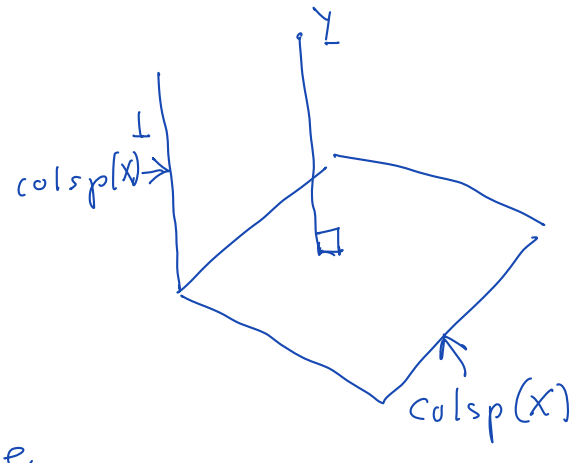
$$= SSE$$

$$\frac{\partial l_R}{\partial \sigma^2} = 0 \Rightarrow \hat{\sigma}^2_{REML} = \ldots = \frac{1}{n-1} SSE = \frac{1}{n-1}\hat{\underline{\varepsilon}}^T\hat{\underline{\varepsilon}}$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

How to find A?

Let $S = I - H = I - X(X^TX)^{-1}X^T$. Eigenvalues and vectors of H satisfies



$$\lambda \underline{v} = S\underline{v}$$

$$\lambda^2 \underline{v} = S^2\underline{v} = S\underline{v}$$

since $S^2 = S$ (why?). Hence

$$S\underline{v} = \lambda\underline{v} = \lambda^2\underline{v}$$

and all eigenvalues satisfies $\lambda = \lambda^2$ and are thus either 0 or 1.

For eigenvalues $\lambda_i = 1$ $(i = 1, \ldots, n-p)$ the eigenvector satisfies

$$S\underline{v}_i = 1 \cdot \underline{v}_i = \underline{v}_i,$$

i.e. $\underline{v}_i$ is in $\text{colsp}(X)^\perp$

For eigenvalues $\lambda = 0$ $(i = n-p+1, \ldots, n)$

$$S\underline{v}_i = \underline{0}$$

i.e. $\underline{v}_i$ is in $\text{colsp}(X)$.

Thus, letting

$$\underset{\text{new coordinates } (n-p)\times 1}{\underline{w}} = A^T \underset{\text{old coordinates } n\times 1}{\underline{y}} = \underbrace{\left[\underline{v}_1\ \underline{v}_2\ \cdots\ \underline{v}_{n-p}\right]^T}_{\text{new basis vectors}} \underline{y}$$

we have $A^T X = 0$, $A^T A = I_{n-p}$ (S is symmetric) and

$$AA^T = PDP^T$$

$$= P \underbrace{\begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} P^T}_{\text{eigendecomposition of } S} = S = I - H$$

R:

```
A <- eigen(I - X(X'X)^-1 X')$vectors[, 1:(n-p)]
```

"Bayesian" interpretation / definition of REML (Harville 1974)

We shall see that the restricted likelihood can be expressed as

$$L_R(\underline{\theta}) = \int L(\underline{\beta}, \underline{\theta})\, d\underline{\beta}$$

(def. in Fahrmeir)

$$\underset{\sim}{\hat{\beta}} = G^T \underset{\sim}{y} \qquad \begin{bmatrix} \underset{\sim}{w} \\ \underset{\sim}{\hat{\beta}} \end{bmatrix} = [A \; G]^T \underset{\sim}{y}$$

$\underset{\sim}{\hat{\beta}}$: $p \times 1$

$$\underset{\sim}{w} = A^T \underset{\sim}{y} \qquad G^T = \left(X^T V^{-1}(\theta) X\right)^{-1} X^T V^{-1}(\theta)$$

$\underset{\sim}{w}$: $(n-p) \times 1$

$\underset{\sim}{\hat{\beta}}$ and $\underset{\sim}{w}$ are independent because

$$\begin{aligned} \mathrm{Cov}(\underset{\sim}{\hat{\beta}}, \underset{\sim}{w}) &= \mathrm{Cov}(G^T \underset{\sim}{y}, A^T \underset{\sim}{y}) \\ &= G^T V(\underset{\sim}{\theta}) A \\ &= \left(X^T V^{-1}(\theta) X\right)^{-1} \underbrace{X^T V^{-1}(\underset{\sim}{\theta}) V(\underset{\sim}{\theta}) A}_{A^T X = 0} = 0 \end{aligned}$$

Now

$$L_R(\underset{\sim}{\theta}) = f_{\underset{\sim}{w}}\left(A^T \underset{\sim}{y} \mid \underset{\sim}{\theta}\right)$$

$$= f_{\underset{\sim}{w}}\left(A^T \underset{\sim}{y} \mid \underset{\sim}{\theta}\right) \int f_{\underset{\sim}{\hat{\beta}}}\left(\underset{\sim}{\hat{\beta}} \mid \underset{\sim}{\beta}, \underset{\sim}{\theta}\right) d\underset{\sim}{\beta}$$

$= 1$ regardless of whether we integrate w.r.t. β or $\hat{\beta}$, for example, note that $\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$

$$= \int f_{\underset{\sim}{w}, \underset{\sim}{\hat{\beta}}}\left(A^T \underset{\sim}{y}, G^T \underset{\sim}{y}\right) d\underset{\sim}{\beta} \qquad \left(\begin{array}{c}\text{since} \\ \underset{\sim}{w} \text{ and } \underset{\sim}{\hat{\beta}} \text{ are} \\ \text{indep.}\end{array}\right)$$

$$= |A \; G| \int f_{\underset{\sim}{y}}\left(\underset{\sim}{y} \mid \underset{\sim}{\beta}, \underset{\sim}{\theta}\right) d\underset{\sim}{\beta} \qquad \left(\begin{array}{c}\underset{\sim}{w}, \underset{\sim}{\hat{\beta}} \text{ linear transf.} \\ \text{of } \underset{\sim}{y}\end{array}\right)$$

$$= |A\,G| \int L(\beta, \theta)\, d\beta$$

$|A\,G|$: does not depend on θ

$\int L(\beta, \theta)\, d\beta$: = alternative def of $L_R(\theta)$

"Bayesian" interpretation:

$L(\beta, \theta) \propto$ joint posterior of β and θ if using uniform improper priors on β and θ

$\int L(\beta, \theta)\, d\beta \propto$ marginal posterior distribution of θ

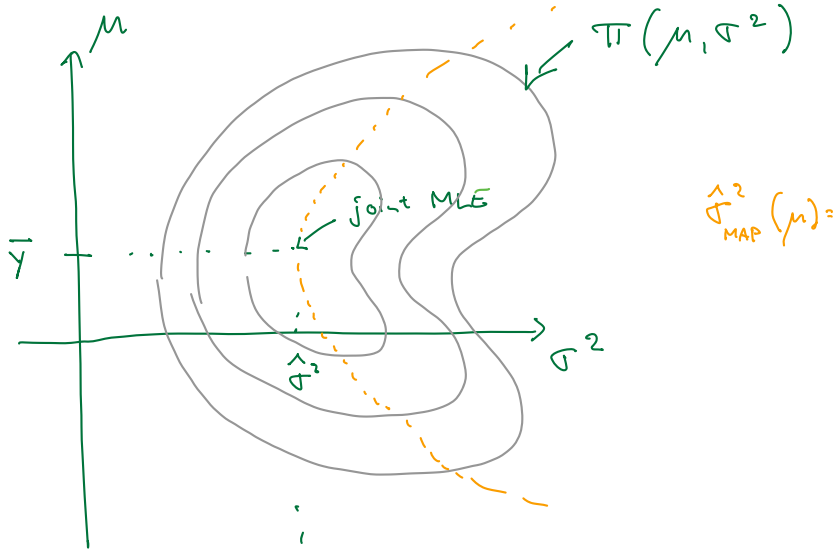
$\hat{\theta}_{REML}$ = Maximum a posteriori (MAP) estimate of θ.

Example: $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$.

Joint posterior of μ, σ^2:

$$\pi(\mu, \sigma^2 \mid \underline{Y}) \propto f(\underline{Y} \mid \mu, \sigma)\, \underbrace{\pi(\mu, \sigma)}_{=1}$$

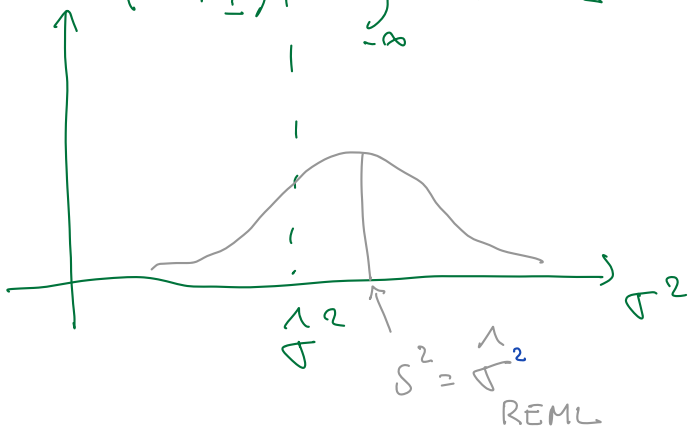
$$= \ldots \propto \frac{1}{\sigma^2} e^{-\frac{(\mu - \bar{Y})^2}{2\sigma^2/n}}$$



$$\hat{\sigma}^2_{MAP}(\mu) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu)^2$$

Marginal posterior of σ^2

$$\pi(\sigma^2|\underline{y}) = \int_{-\infty}^{\infty} \pi(\mu, \sigma^2|\underline{y})\, d\mu$$



21.10 ↓

↑ 26.10

Why doesn't $|A\ G|$ depend on θ?

$$|A\ G| = \left| \begin{bmatrix} A^T \\ G^T \end{bmatrix} [A\ G] \right|^{1/2}$$

wikipedia

$$\left| \begin{matrix} A & B \\ C & D \end{matrix} \right| = |A| \times |D - CA^{-1}B|$$

$A^TA = I_{n-p}$

$AA^T = I - H$

$= I - X(X^TX)^{-1}X^T$

$$= \left| \begin{matrix} A^TA & A^TG \\ G^TA & G^TG \end{matrix} \right|^{1/2}$$

$$= |A^TA|^{1/2} \times |G^TG - G^TA(A^TA)^{-1}A^TG|^{1/2}$$

($A^TA = I$)

$$= |I|^{1/2} \times |G^T(I - AA^T)G|^{1/2}$$

$$= |G^T(I - I + X(X^TX)^{-1}X^T)G|^{1/2}$$

$$= |(X^TV^{-1}X)^{-1}X^TV^{-1}X(X^TX)^{-1}X^TX V^{-1}(X^TV^{-1}X)^{-1}|^{1/2}$$

$$= |(X^TX)^{-1}|^{1/2} = |X^TX|^{-1/2}$$

i.e. doesn't depend on θ (although G depend on θ)

Connection between profile and restricted likelihoods

$L_p(\underline{\theta})$ and $L_R(\underline{\theta})$ (Eq. 7.29 in Fahr)

Recall the profile likelihood

$$L_p(\underline{\theta}) = (2\pi)^{-\frac{n}{2}} |V(\theta)|^{-\frac{1}{2}} e^{-\frac{1}{2}(\underline{y} - X\hat{\beta}(\theta))^T V^{-1}(\theta)(\underline{y} - X\hat{\beta}(\underline{\theta}))} \quad (1)$$

The restricted likelihood

$$L_R(\underline{\theta}) = \int f(\underline{y} | \underline{\beta}, \underline{\theta}) \, d\underline{\beta}$$

$$= (2\pi)^{-\frac{n}{2}} |V(\theta)|^{-\frac{1}{2}} \int e^{-\frac{1}{2}(\underline{y} - X\underline{\beta})^T V^{-1}(\theta)(\underline{y} - X\underline{\beta})} \, d\underline{\beta} \quad (2)$$

$\underbrace{\qquad}_{p \text{ dimensional}}$

Consider the decomposition

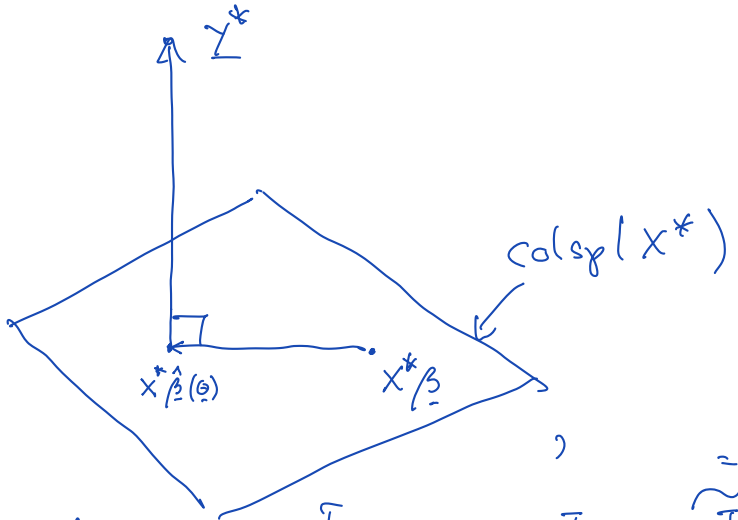
$$V(\theta)^{-1/2}(\underline{y} - X\underline{\beta}) = \underline{y}^* - X^*\underline{\beta}$$

$$= \underline{y}^* - X^*\hat{\beta}(\theta) + X^*\hat{\beta}(\theta) - X^*\underline{\beta}$$

$$= V(\theta)^{-1/2}(\underline{y} - X\hat{\beta}(\theta)) + V(\theta)^{-1/2} X(\hat{\beta}(\theta) - \underline{\beta}) \quad (3)$$

$$= \underline{a} + \underline{b}$$

The vectors on the right hand side are are orthogonal:



such that, $(\underline{a}+\underline{b})^T(\underline{a}+\underline{b}) = \underline{a}^T\underline{a} + \underbrace{\underline{a}^T\underline{b}}_{=0} + \underbrace{\underline{b}^T\underline{a}}_{=0} + \underline{b}^T\underline{b}$.

Hence, inserting (3) into (2) we obtain

$$L_R(\underline{\theta}) = (2\pi)^{-\frac{n}{2}} |V(\theta)|^{-\frac{1}{2}} e^{-\frac{1}{2}(y-X\hat{\beta}(\theta))^T V^{-1}(\theta)(y-X\hat{\beta}(\theta))} \times \int e^{-\frac{1}{2}(\hat{\beta}(\theta)-\beta)^T X^T V^{-1}(\theta) X (\hat{\beta}(\theta)-\beta)} d\underline{\beta}$$

$$= L_p(\underline{\theta}) \underbrace{(2\pi)^{\frac{p}{2}} \underbrace{|X^T V^{-1}(\underline{\theta}) X|}_{p\times p}}_{\text{normalising constant}}$$

(1)

Computing this only involves computing the inverse and determinant of a block diagonal $n \times n$ matrix which is cheap.

Hypothesis testing for LMMs

$\mathrm{Var}(\underline{y}) = V(\hat{\theta})$

We have

$$\hat{\underline{\beta}} = \left(X^T V^{-1}(\hat{\theta}) X\right)^{-1} X^T V^{-1}(\hat{\theta})\, \underline{y}$$

Ignoring uncertainty in $\hat{\theta}$,

$$\mathrm{Var}(\hat{\beta}) = \ldots = X^T V^{-1}(\hat{\theta}) X$$

is an estimate of $\mathrm{Var}(\hat{\beta})$.

The Wald statistic

$$(C\hat{\underline{\beta}} - \underline{d})^T \left(C X^T V^{-1}(\hat{\theta}) X C^T\right)^{-1} (C\hat{\beta} - \underline{d})$$

is approximately chisquare. No exact F-test

LRTs

Fixed effect testing: LRTs must be based on the full likelihood because the restricted maximum likelihoods are likelihoods of different subsets of the data.

Random effects: LRT and RLRT based on $L_R(\hat{\theta})$ can both be used as the likelihoods are based on the same $\underline{w} = A^T \underline{y}$

$\uparrow$

X

R: lmer

H_0 on the boundary

Example 1 (random intercept model)

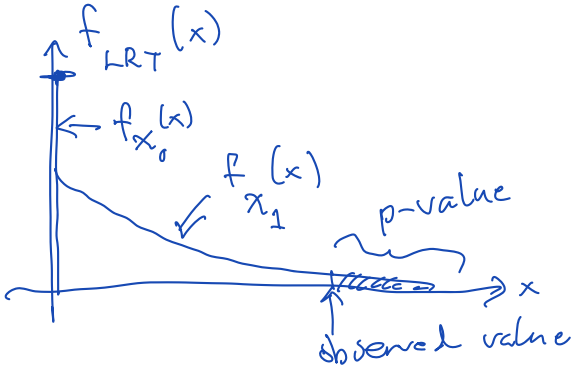
$H_0: \tau_0^2 = 0$ (no random intercept in model)

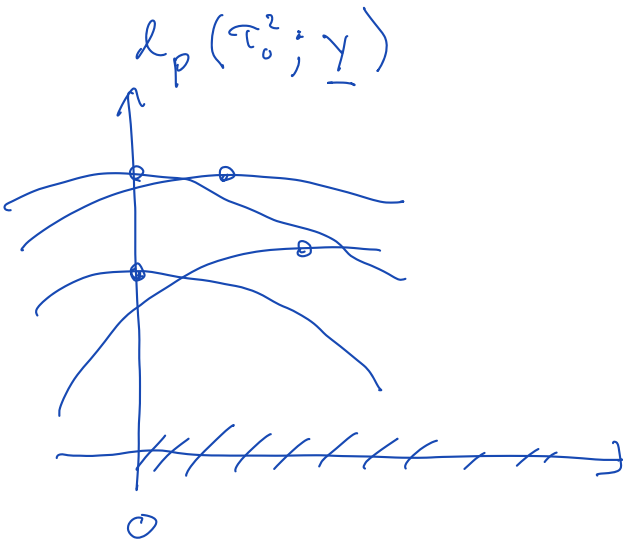
$H_1: \tau_0^2 > 0$ (random effect present)

$$LRT = 2\left(l_p(\hat{\tau}_0^2) - l_p(0)\right) \overset{asymp}{\sim} 0.5\,\chi_0^2 : 0.5\,\chi_1^2$$

i.e.

$$f_{LRT}(x) = 0.5\,\underbrace{\delta(x)}_{\text{point mass at zero}} + 0.5\, f_{\chi_1^2}(x)$$





Under H_0

$$P\left(\hat{\tau}_0 > 0\right) = \frac{1}{2}$$

$\hat{\tau}_0$: MLE under H_1

since

$$E\left(\frac{\partial l}{\partial \tau_0^2}\Big|_{\tau_0^2 = 0}\right) = 0$$

Example 2: Random intercept and slope model

$H_0: \tau_0^2 \geq 0, \ \tau_1^2 = 0, \ \tau_{01}^2 = 0$ (only random intercept)

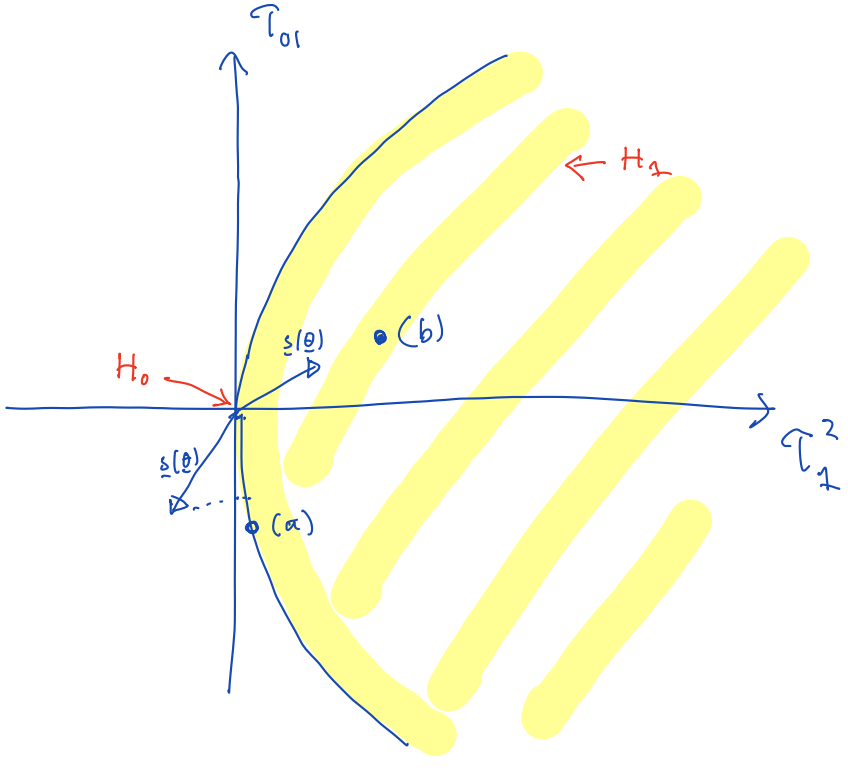
$H_1: \tau_0^2 \geq 0, \ \tau_1^2 > 0, \ |\tau_{01}| < \sqrt{\tau_0^2 \tau_1^2}$ (random intercept and slope)

$$LRT = 2\left(l\left(\hat{\tau}_0^2, \hat{\tau}_1, \hat{\tau}_{01}\right) - l\left(\hat{\tau}_{0,0}^2, 0, 0\right)\right)$$

asymp.

$\sim 0.5\,\chi_1^2 : 0.5\,\chi_2^2$ (Fahrmeir)

Heuristic explanation?



Asymptotically, under H_0, there is a 50% chance that the MLE under H_1 ends up on the boundary (a) and 50% chance that it ends up at the interior (b)

26.10

Estimating the random effects $\underline{\gamma}$ (7.3.1, 7.3.3)

28.10

Recall conditional model

$$\underline{y} \mid \underline{\gamma} \sim N(X\underline{\beta} + U\underline{\gamma}, R)$$

and

$$\underline{\gamma} \sim N(\underline{0}, G)$$

Given $\hat{\theta}$ estimated by REML or ML, $\underline{\alpha} = \begin{bmatrix} \underline{\beta} \\ \underline{\gamma} \end{bmatrix}$ can be estimated by maximising the "joint penalized likelihood"

$$L(\underline{\alpha}) = L(\underline{\beta}, \underline{\gamma})$$

$$= f_{\underline{y}|\underline{\gamma}}(\underline{y}) f_{\underline{\gamma}}(\underline{\gamma})$$

$$\propto e^{-\frac{1}{2}(y - X\beta - U\gamma)^T R^{-1}(y - X\beta - U\gamma)} \; e^{-\frac{1}{2}\underline{\gamma}^T G^{-1}\underline{\gamma}} \Big\} \text{ Gaussian in } \underline{\alpha}$$

or minimising

$$(y - X\beta - U\gamma)^T R^{-1}(y - X\beta - U\gamma) + \gamma^T G^{-1}\gamma$$

$$= (y - C\alpha)^T R^{-1}(y - C\alpha) + \alpha^T B \alpha \quad = (*)$$

where

$$C = \underbrace{[X \; U]}_{n \times (p+\cdot}$$ and $$B = \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} \end{bmatrix}$$

Recall

$$\frac{\partial}{\partial \underline{x}}(\underline{x}^T A \underline{x}) = 2A\underline{x}$$

$$\frac{\partial}{\partial \underline{x}}(A\underline{x})^T = A^T$$

$$\frac{\partial}{\partial \underline{x}} f(\underline{g}(\underline{x})) = \frac{\partial \underline{g}^T}{\partial \underline{x}} \cdot \frac{\partial f}{\partial \underline{g}}$$

Thus

$$\frac{\partial *}{\partial \underline{\alpha}} = -C^T 2R^{-1}(\underline{y} - C\underline{\alpha}) + 2B\underline{\alpha}$$

$$= -2C^T R^{-1}\underline{y} + 2\left(C^T R^{-1} C + B\right)\underline{\alpha}$$

Equating to $\underline{0}$ and solving

$$\hat{\underline{\alpha}} = \left(C^T R^{-1} C + B\right)^{-1} C^T R^{-1} \underline{y}$$

Since $L(\underline{\alpha}) \propto$ Gaussian function of $\underline{\alpha}$ with variance matrix $(C^T R^{-1} C + B)^{-1}$. Thus, the "posterior variance" of $\underline{\alpha}$

$$\mathrm{Var}(\underline{\alpha} \mid \underline{y}) = \left(C^T R^{-1} C + B\right)^{-1}$$

If instead fixing $\underline{\gamma}$ and thinking of $\underline{y}$ as random (frequentist view)

$$\mathrm{Var}(\hat{\underline{\alpha}}) = \mathrm{Var}\left(\left(C^T R^{-1} C + B\right)^{-1} C^T R^{-1} \underline{y}\right)$$

$$= \left(C^T R^{-1} C + B\right)^{-1} C^T R^{-1} R R^{-1} C \left(C^T R^{-1} C + B\right)^{-1}$$

$$= \mathrm{Var}(\underline{\alpha} \mid \underline{y})\, C^T R^{-1} C \left(C^T R^{-1} C + B\right)^{-1}$$

$$\leq \mathrm{Var}(\underline{\alpha} \mid \underline{y}), \quad \text{i.e. } \mathrm{Var}(\underline{\alpha}(\underline{y}) - \mathrm{Var}(\underline{\alpha}) \text{ is p.s.d.}$$

Alternative formulation: Recall that if

$$\begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right)$$

then

$$E(\underline{x}_1 \mid \underline{x}_2) = \underline{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\underline{x}_2 - \underline{\mu}_2)$$

We have

$$\mathrm{Cov}(A\underline{x}, B\underline{y})$$

$$\begin{bmatrix} \underline{\gamma} \\ \underline{y} \end{bmatrix} \sim N\left(\begin{bmatrix} \underline{0} \\ X\underline{\beta} \end{bmatrix}, \begin{bmatrix} G & GU^T \\ UG & \underbrace{UGU^T + R}_{=V} \end{bmatrix}\right)$$

$$y = X\beta + U\gamma + \varepsilon$$

Hence

$$\hat{\underline{\gamma}} = E(\underline{\gamma} \mid \underline{y}) = \underline{0} + G U^T \hat{V}^{-1} (\underline{y} - X\hat{\underline{\beta}}) \qquad (BLUP)$$

Example: Random intercept model: Observation $j = 1, \dots, n_i$ in cluster i

$$y_{ij} = \underline{x}_{ij}^T \underline{\beta} + \gamma_{0,i} + \varepsilon_{ij}$$

$$\underline{y}_i = X_i^T \underline{\beta} + \underbrace{U_i}_{\underline{1}_{n_i}} \gamma_{0,i} + \underline{\varepsilon}_i, \quad \gamma_{0,i} \sim N(0, \tau_0^2), \quad \underline{\varepsilon}_i \sim N(\underline{0}, \sigma^2 \underline{I}_{n_i})$$

$$V_i = \mathrm{Var}(\underline{y}_i) = U_i \tau_0^2 U_i^T + \sigma^2 \underline{I}_{n_i}$$

$$= \tau_0^2 \underbrace{\underline{J}}_{n_i \times n_i} + \sigma^2 \underline{I}_{n_i}$$

We need

$$V_i^{-1} = a \underline{I}_{n_i} + b \underline{J}_{n_i}$$

satisfying

$$(\sigma^2 \underline{I} + \tau_0^2 \underline{J})(a \underline{I} + b \underline{J}) = \underline{I}$$

$$a\sigma^2 \underline{I} + \sigma^2 b \underline{J} + \tau_0^2 a \underline{J} + \tau_0^2 b n_i \underline{J} = \underline{I}$$

$$a\sigma^2 = 1 \Rightarrow a = \frac{1}{\sigma^2}$$

$$\Rightarrow b = -\frac{\tau_0^2}{(\sigma^2 + n_i \tau_0^2)\sigma^2}$$

$$V_i^{-1} = \frac{1}{\sigma^2}\left(\underline{I}_{n_i} - \frac{\tau_0^2}{\sigma^2 + n_i \tau_0^2} \underline{J}_{n_i}\right)$$

$A^{-1}A = \mathbb{1}$

$J = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$

$J^2 = JJ = ?$

We have

$$\hat{\underline{\gamma}} = G Z^T \hat{V}^{-1} (\underline{y} - X\hat{\underline{\beta}})$$

and within each cluster $i = 1, \dots, m$

$$\hat{\underline{\gamma}}_i = \tau_0^2 \underline{1}_{n_i}^T \frac{1}{\sigma^2} \left(I_{n_i} - \frac{\tau_0^2}{\sigma^2 + n_i \tau_0^2} J_{n_i} \right) (\underline{y}_i - X_i \hat{\underline{\beta}})$$

$$= \frac{\tau_0}{\sigma^2} \left(\underline{1}^T - \underline{1}^T n_i \frac{\tau_0^2}{\sigma^2 + n_i \tau_0^2} \right) (\underline{y}_i - X_i \hat{\underline{\beta}})$$

$$= \frac{\tau_0}{\sigma^2} \left(1 - \frac{\tau_0^2 n_i}{\sigma^2 + n_i \tau_0^2} \right) n_i \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - \underline{x}_{ij}^T \hat{\underline{\beta}})$$

$$= \underbrace{\frac{n_i \tau_0^2}{\sigma^2 + n_i \tau_0^2}}_{\text{shrinking factor} < 1} \cdot \underbrace{\frac{1}{n_i} \sum (y_{ij} - \underline{x}_{ij}^T \hat{\underline{\beta}})}_{\text{mean residual within cluster } i}$$

Small n_i or τ_0^2
$\Rightarrow$ more shrinkage

R: `mod <- lmer(Reaction ~ 1 + Days + (1 + Days | subject))`

`gammahat <- ranef(mod, condVar = TRUE)`

`gammahat$Subject` # $\hat{\gamma}$

`attributes(gammahat$Subject)` # $\mathrm{Var}(\hat{\gamma} \mid \underline{y})$ as a $m \times 2 \times 2$ array

Uncertainty in $\underline{\theta} = (\sigma^2, \tau_0^2, \tau_1^2, \tau_{01}^2)$ ignored.

Generalized linear mixed model (GLMMs) (Ch. 7.5)

Distribution of the response y_{ij} conditional on random effect $\underline{\gamma}_i$, $f(y_{ij} | \underline{\gamma}_i)$ belongs to the exponential family with

$$\mu_{ij} = E(y_{ij} | \underline{\gamma}_i)$$

related to

$$\eta_{ij} = \underline{x}_{ij}^T \underline{\beta} + \underline{u}_{ij}^T \underline{\gamma}_i \quad , \quad \underline{\gamma}_i \sim N(\underline{0}, Q)$$

for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n_i$ through

$$\mu_{ij} = h(\eta_{ij}) \quad \text{or} \quad g(\mu_{ij}) = \eta_{ij}.$$

All y_{ij} are conditionally independent.

Conditional versus marginal models (7.5.2)

Marginal distribution of y_{ij} becomes

$$f(y_{ij}) \overset{LTP}{=} \int f(y_{ij}, \underline{\gamma}_i) \, d\underline{\gamma}_i$$

$$= \int f(y_{ij} | \underline{\gamma}_i) f(\underline{\gamma}_i) \, d\underline{\gamma}_i$$

that is, a compound distribution or uncountable mixture of e.g. binomial or poisson distributions

logitnormal binomial

lognormal Poisson

Marginal joint distribution of $\underline{y}_i = (y_{i1}, y_{i2}, \dots, y_{in_i})$

$$f(\underline{y}_i) = \int f(\underline{y}_i \mid \underline{\gamma}_i) f(\underline{\gamma}_i) \, d\underline{\gamma}_i$$

$$= \int \prod_{j=1}^{n_i} f(y_{ij} \mid \underline{\gamma}_i) \, f(\underline{\gamma}_i) \, d\underline{\gamma}_i$$

must be approximated / computed numerically

Poisson random intercept model

$$E(y_{ij} \mid \gamma_{0,i}) = e^{\eta_{ij}} = e^{\underline{x}_{ij}^T \underline{\beta} + \underline{u}_{ij}^T \underline{\gamma}_i} \qquad \gamma_{0,i} \sim N(0, \tau_0^2)$$

$$= e^{\underline{x}_{ij}^T \underline{\beta} + \gamma_{0,i}} = e^{\underline{x}_{ij}^T \underline{\beta}} e^{\gamma_{0,i}}$$

$Z \sim \text{lognormal}(\mu, \sigma^2)$

$\ln Z = X \sim N(\mu, \sigma^2)$, $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

$M_X(t) = E e^{tX}$

$$EZ = E(e^X) = M_X(1) = e^{\mu + \frac{1}{2}\sigma^2}$$

$$E(Z^2) = E(e^{2X}) = M_X(2) = e^{2\mu + 2\sigma^2}$$

$$\text{Var } Z = E(Z^2) - (EZ)^2 = \dots = (EZ)^2 \left(e^{\sigma^2} - 1\right)$$

Law of total expectation (LTE)

$$E(EX|Y) = EX$$

Meaning of $EX|Y$?

Conditional expectation

$$E(X|Y=y) = g(y)$$

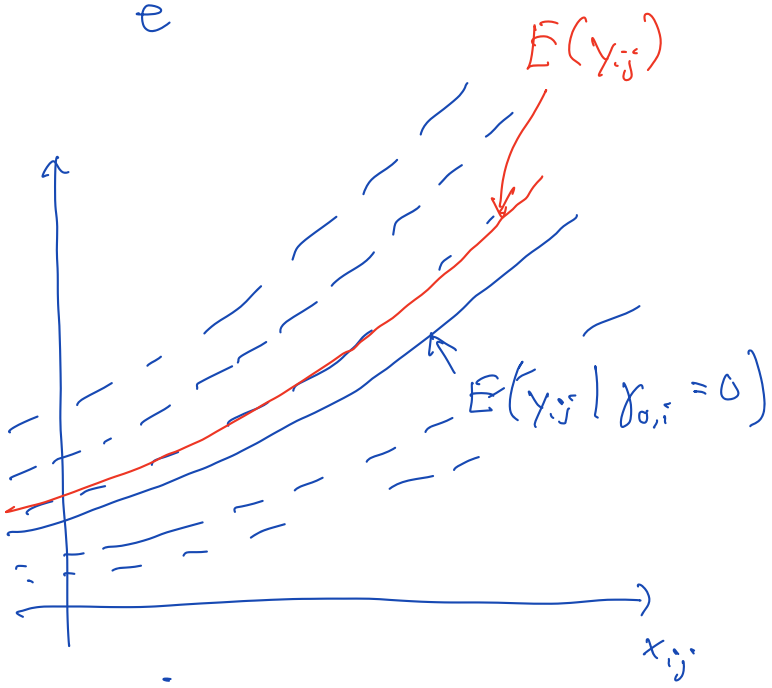
By $E(X|Y)$ we mean the random $g(Y)$

Hence, using LTE,

$$Ey_{ij} = E(E y_{ij} | \gamma_{0,i}) = E\, e^{x_{ij}^T \beta} e^{\gamma_{0,i}}$$

$$= e^{x_{ij}^T \beta} E(e^{\gamma_{0,i}})$$

$$= e^{x_{ij}^T \beta + \frac{1}{2}\tau_0^2}$$



Effect of unit change in x_{ijk} is that both $E(y_{ij})$ and $E(y_{ij}|\gamma_{0,i})$ increase by factor e^{β_k}

Using the law of total variance,

$$\text{Var}\, y_{ij} = E\,\text{Var}(y_{ij}|\gamma_{0,i}) + \text{Var}(E\, y_{ij}|\gamma_{0,i}))$$

$$= E\, e^{x_{ij}^T \beta + \gamma_{0,i}} + \text{Var}\, e^{x_{ij}^T \beta} e^{\gamma_{0,i}}$$

$$= e^{\underline{x}_{ij}^T\underline{\beta} + \frac{1}{2}\tau_0^2} + e^{2\underline{x}_{ij}^T\underline{\beta}} e^{\tau_0^2}\left(e^{\tau_0^2} - 1\right)$$

Using law of total covariance, the intraclass covariance

$$\mathrm{Cov}(y_{ij}, y_{ik}) = E\,\mathrm{Cov}(y_{ij}, y_{ik} \mid \gamma_{0,i}) + \mathrm{Cov}\left[E(y_{ij} \mid \gamma_{0,i}),\ E(y_{ik} \mid \gamma_{0,i})\right]$$

$$= E\ 0 + \mathrm{Cov}\left(e^{\underline{x}_{ij}^T\underline{\beta} + \gamma_{0,i}},\ e^{\underline{x}_{ik}^T\underline{\beta} + \gamma_{0,i}}\right)$$

$$= e^{(\underline{x}_{ij}^T + \underline{x}_{ik}^T)\underline{\beta}}\,\mathrm{Var}\left(e^{\gamma_{0,i}}\right)$$

$$= e^{(\underline{x}_{ij}^T + \underline{x}_{ik}^T)\underline{\beta}}\, e^{\tau_0^2}\left(e^{\tau_0^2} - 1\right).$$

Intraclass correlation:

$\mathrm{corr}(y_{ij}, y_{ik}) = \ldots$

Binary random intercept probit model

Conditional model

$$\text{probit}\ P(y_{ij} = 1 \mid \gamma_{0,i}) = \underline{x}_{ij}^T\underline{\beta} + \gamma_{0,i}$$

or

$$P(y_{ij} = 1 \mid \gamma_{0,i}) = \Phi\left(\underline{x}_{ij}^T\underline{\beta} + \gamma_{0,i}\right), \quad \gamma_{0,i} \sim N(0, \tau_0^2)$$

Marginal model

$$P(y_{ij} = 1) \overset{\text{LTP}}{=} \int P(y_{ij} = 1 \mid \gamma_{0,i})\, f(\gamma_{0,i})\, d\gamma_{0,i}$$

$$= \int \Phi\left(\underline{x}_{ij}^T\underline{\beta} + \gamma_{0,i}\right) f(\gamma_{0,i})\, d\gamma_{0,i}$$

$$= \int P\left(Z \leq \underline{x}_{ij}^{\tau}\underline{\beta} + \gamma_{0,i} \mid \gamma_{0,i}\right) f\left(\gamma_{0,i}\right) d\gamma_{0,i}$$

LTP

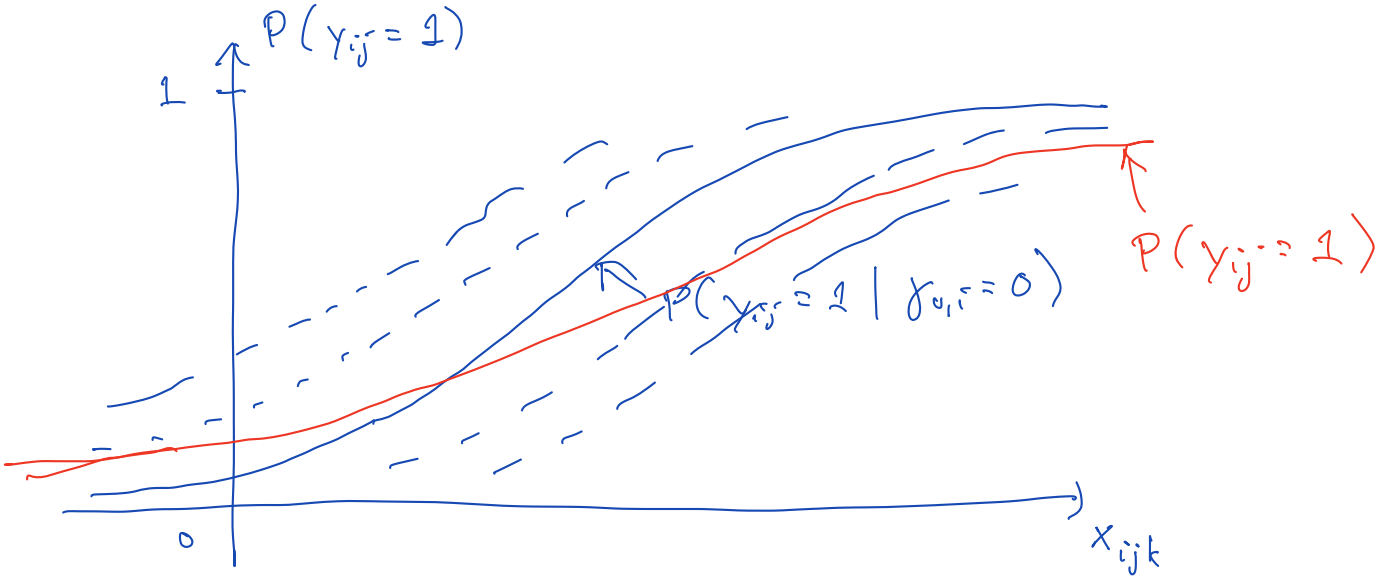
$$= P\left(Z \leq \underline{x}_{ij}^{\tau}\underline{\beta} + \gamma_{0,i}\right)$$

$$= P\left(\frac{Z - \gamma_{0,i}}{\sqrt{1+\tau_0^2}} \leq \frac{\underline{x}_{ij}^{\tau}\underline{\beta}}{\sqrt{1+\tau_0^2}}\right)$$

$\sim N(0,1)$

$$= \Phi\left(\frac{\underline{x}_{ij}^{\tau}\underline{\beta}}{\sqrt{1+\tau_0^2}}\right)$$

The effect is that $\underline{\beta}$ is deflated by a factor $\frac{1}{\sqrt{1+\tau_0^2}}$



Binary random intercept logit model

Conditionally

$$P(y_{ij} = 1 \mid \gamma_{0,i}) = \frac{1}{1 + e^{-(\underline{x}_{ij}^{T}\underline{\beta} + \gamma_{0,i})}}$$

Marginally

$$P(y_{ij} = 1) \approx \frac{1}{1 + e^{-\underline{x}_{ij}^{T}\underline{\beta} / \sqrt{1 + 0.6\,\tau_0^2}}}$$

(Agresti 2002, p. 499)

Inference for GLMMs

1) Numerical integration lme4::glmer(, nAGQ = 10)

$$L(\underline{\beta}, \underline{\theta}) = f(\underline{y} \mid \underline{\beta}, \underline{\theta})$$

$$= \prod_{i=1}^{m} f(\underline{y}_i \mid \underline{\beta}, \underline{\theta}) \qquad \text{(clustered data)}$$

$$= \prod_{i=1}^{m} \int f(\underline{y}_i \mid \underline{\gamma}_i, \underline{\beta}) f(\underline{\gamma}_i \mid \underline{\theta}) \, d\underline{\gamma}_i \qquad \text{(LTP)}$$

$$= \prod_{i=1}^{m} \underbrace{\left(\int \prod_{j=1}^{n_i} f(y_{ij} \mid \underline{\gamma}_i, \underline{\beta}) f(\underline{\gamma}_i \mid \underline{\theta}) \, d\underline{\gamma}_i \right)}_{\int_{-\infty}^{\infty} g(t)\, dt}$$

Numerical integration of $g(t)$ via Gauss quadrature:

Laplace approximation example (Stirling's formula)

$$n! = \Gamma(n+1)$$

$$= \int_0^\infty x^n e^{-x}\, dx$$

$$= \int_0^\infty e^{-(x - n\ln x)}\, dx$$

$$= \int_0^\infty e^{-g(x)}\, dx$$

$$= \int_0^\infty e^{-g(n) - \frac{1}{2} g''(n)(x-n)^2}$$

$$= e^{-n + n\ln n} \int_0^\infty e^{-\frac{1}{2n}(x-n)^2}\, dx$$

$$\overset{n \text{ large}}{=} e^{-n + n\ln n} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-n}{\sqrt{n}}\right)^2}\, dx$$

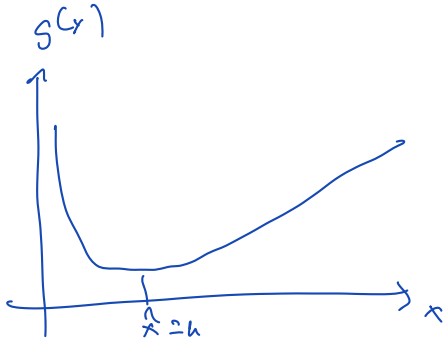
$$= e^{-n + n\ln n} \sqrt{2\pi n}$$

$$= \sqrt{2\pi n}\, e^{-n}\, n^n$$

$g(x) = x - n\ln x$

$g'(x) = 1 - \frac{n}{x} = 0$

Minimum at $\hat{x} = n$



$x = \hat{x}$

$g''(x) = \frac{n}{x^2} = \frac{1}{n}$

2) Laplace approximation of marginal likelihood ↑ 4.11

$$L(\underline{\beta}, \underline{\theta}) = \int \underbrace{f(\underline{y} \mid \underline{\hat{\gamma}}, \underline{\beta}) f(\underline{\gamma} \mid \underline{\theta})}_{f(\underline{y}, \underline{\gamma} \mid \underline{\beta}, \underline{\theta})} d\underline{\gamma}$$

$$= \int e^{-g(\underline{\gamma}, \underline{\beta}, \underline{\theta})} d\underline{\gamma}$$

$$\approx \int e^{-g(\hat{\gamma}(\underline{\beta}, \underline{\theta}), \underline{\beta}, \underline{\theta}) - \frac{1}{2}(\underline{\gamma} - \hat{\gamma}(\underline{\beta}, \underline{\theta}))^T H(\underline{\beta}, \underline{\theta})(\underline{\gamma} - \hat{\gamma}(\underline{\beta}, \underline{\theta}))} d\underline{\gamma}$$

where $\hat{\gamma}(\underline{\beta}, \underline{\theta}) = \underset{\gamma}{\operatorname{argmin}}\, g(\underline{\gamma}, \underline{\beta}, \underline{\theta})$ (inner optimization)

and

$$H(\underline{\beta}, \underline{\theta}) = \frac{\partial^2}{\partial \underline{\gamma} \partial \underline{\gamma}^T} g(\underline{\gamma}, \underline{\beta}, \underline{\theta}) \Big|_{\underline{\gamma} = \hat{\underline{\gamma}}(\underline{\beta}, \underline{\theta})}$$

Thus,

$$L(\underline{\beta}, \underline{\theta}) \approx e^{-g(\hat{\gamma}(\underline{\beta}, \underline{\theta}), \underline{\beta}, \underline{\theta})} (2\pi)^{k/2} |H(\underline{\beta}, \underline{\theta})|^{-\frac{1}{2}}$$

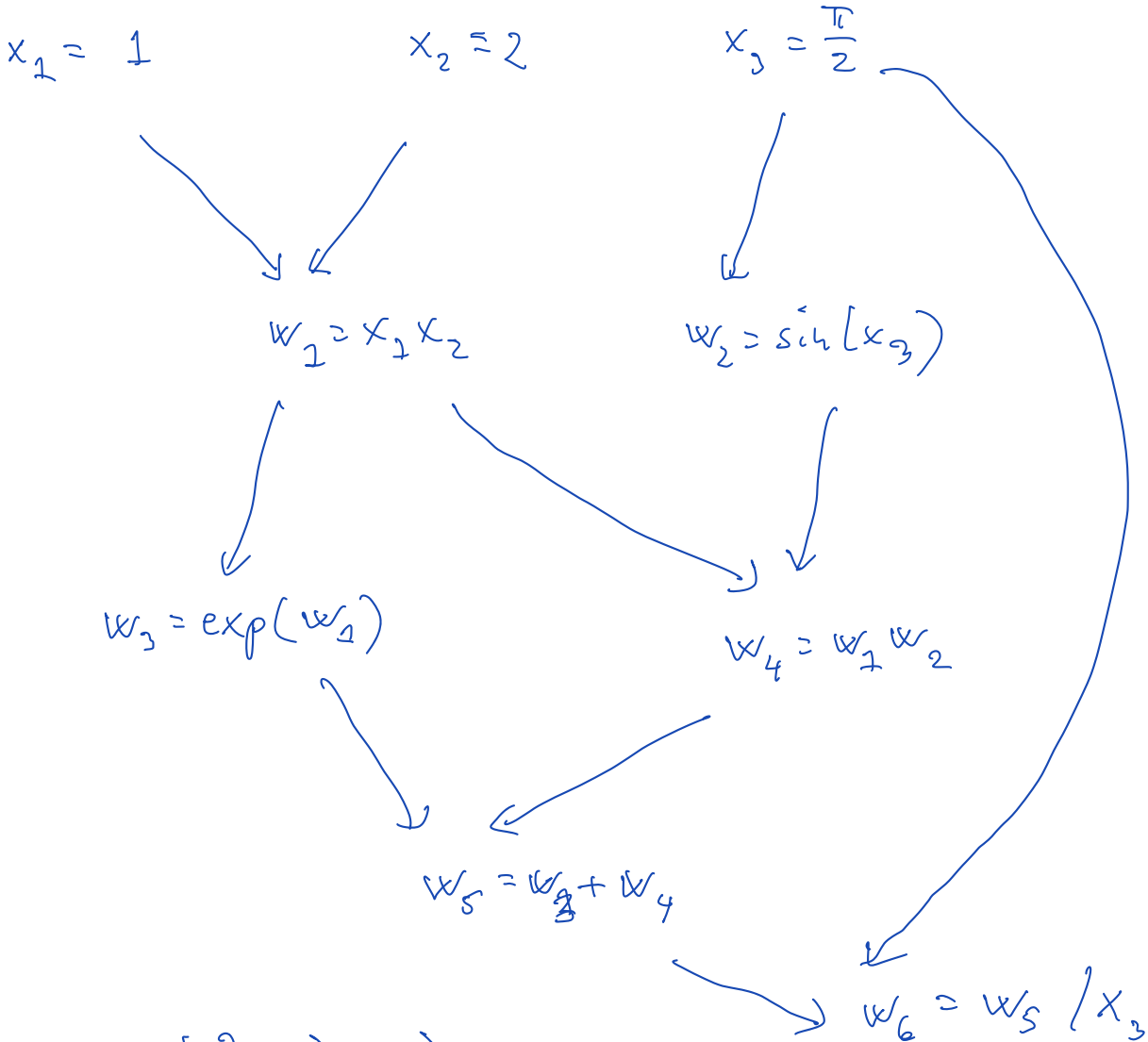
where k is the number of random effects

Automatic differentiation

Used by R-packages TMB and glmmTMB, Stan (different from symbolic and numerical differentiation)

Example: Find $\frac{\partial f}{\partial \underline{x}}$ for $f(\underline{x}) = f(x_1, x_2, x_3)$

$$= \left(x_1 x_2 \sin(x_3) + e^{x_1 x_2} \right) / x_3$$



		$\frac{\partial}{\partial x_1}$	$\frac{\partial}{\partial x_2}$	$\frac{\partial}{\partial x_3}$
x_1	1	1	0	0
x_2	2	0	1	0
x_3	$\pi/2$	0	0	1
w_1	2	2	1	0
w_2	1	0	0	0
w_3	-	-	-	-

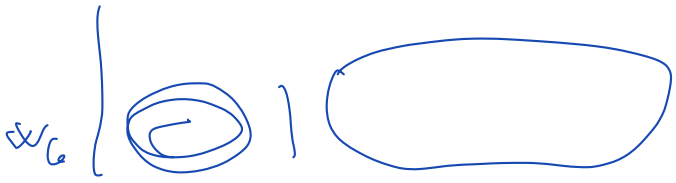
dual numbers

$w_1 = x_1 x_2 = 2$

$$\frac{\partial w_1}{\partial \underline{x}} = \frac{\partial}{\partial \underline{x}}(x_1 x_2)$$

$$= \frac{\partial x_1}{\partial \underline{x}} x_2 + x_1 \frac{\partial x_2}{\partial \underline{x}}$$

$$= \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \cdot 2 + 1 \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$



$$w_3 = e^{w_1}$$

$$\frac{\partial w_3}{\partial \underline{x}} = e^{w_1} \frac{\partial w_1}{\partial \underline{x}} = e^2 \cdot \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

etc...

REML for GLMMs (Millar 2011)

For LMMs:

$$L_R(\underline{\theta}) = f_{\underline{w}}(\underline{w}), \quad \underline{w} = A^T \underline{y}$$

$\underline{w}$ in $n-p \times 1$, $\underline{y}$ in $n \times 1$

but this doesn't generalize to GLMMs.

But the equivalent definition

$$L_R(\underline{\theta}) = \int L(\underline{\theta}, \underline{\beta}) \, d\underline{\beta}$$

$$= \int \int f(\underline{y} | \underline{\gamma}, \underline{\beta}) f(\underline{\gamma} | \underline{\theta}) \, d\underline{\gamma} \, d\underline{\beta}$$

does and can be evaluated by Laplace approximation w.r.t. both $\underline{\gamma}$ and $\underline{\beta}$

glmmTMB(, REML = TRUE)

Gauss Legendre / Hermite quadrature

Warm up: If $f(x)$ is a polynomial of degree $n-1$, then

$$\int_{-\infty}^{\infty} f(x)\, \underbrace{w(x)}_{\text{weight function}}\, dx \simeq \sum_{i=1}^{n} w_i f(x_i) \qquad (1a)$$

Legendre: $w(x) = \begin{cases} 1 & \text{for } -1 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$

Hermite: $w(x) = e^{-\frac{x^2}{2}}$

where $w_1, \ldots, w_n$ satisfies

$$\left.\begin{aligned} f(x) = 1 &\Rightarrow \int w(x)\,dx = \sum w_i \\ f(x) = x &\Rightarrow \int x\, w(x)\,dx = \sum x_i w_i \\ &\vdots \\ f(x) = x^{n-1} &\Rightarrow \int x^{n-1} w(x)\,dx = \sum x_i^n w_i \end{aligned}\right\} \; n \text{ eqs.}$$

or

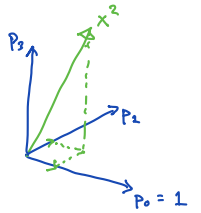
$$\begin{bmatrix} 1 & 1 & 1 & \\ x_1 & x_2 & x_3 & \\ x_1^2 & x_2^2 & \cdots & \\ & & & \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \simeq \begin{bmatrix} \int w(x)\,dx \\ \int x\, w(x)\,dx \\ \vdots \\ \int x^{n-1} w(x)\,dx \end{bmatrix} \qquad (1b)$$

How to best choose evaluation points $x_1, \ldots, x_n$?

Choose $\langle p, q \rangle = \int p(x) q(x) w(x)\,dx$ as inner product and construct orthogonal polynomials (basis vectors) from the polynomials $1, x, x^2, \ldots$ via Gram-Schmidt:

$$p_0(x) = 1$$

$$p_1(x) = x - \frac{\langle x, p_0 \rangle}{\langle p_0, p_0 \rangle} p_0(x) = \ldots = x$$



$$p_2(x) = x^2 - \frac{\langle x^2, p_0 \rangle}{\langle p_0, p_0 \rangle} p_0(x) - \frac{\langle x^2, p_1 \rangle}{\langle p_1, p_1 \rangle} p_1(x) = \ldots = x^2 - 1$$

$$p_3(x) = \ldots = x^3 - 3x$$

for $w(x) = e^{-\frac{x^2}{2}}$, $p_0(x), p_1(x), \ldots$ are known as Hermite polynomials.

Suppose that $f(x)$ instead is of (at most) degree $2n-1$. Via polynomial long division by $p_n(x)$ we have

$$f(x) = \underbrace{q(x)}_{\text{quotient polynomial of order} \leq n-1} p_n(x) + \underbrace{r(x)}_{\text{remainder polynomial of order} \leq n-1} \qquad (2)$$

E.g.
$x^3 - 2x^2 - 4 = (x^2 + x + 3)(x-3) + 5$
quotient $q(x)$, divisor, $r(x)$

and hence

$$\int f(x)\, w(x)\, dx = \underbrace{\int q(x)\, p_n(x)\, w(x)\, dx}_{= \langle q, p_n \rangle = 0^*} + \int r(x)\, w(x)\, dx \qquad (3)$$

since p_n is othogonal to all polynomials spanned by $p_{n-1}, p_{n-2}, \ldots, p_0$ including q (of order $n-1$).

Choosing the n real roots $x_1, \ldots, x_n$ of $p_n(x)$ as evaluation (quadrature) points, (2) implies that

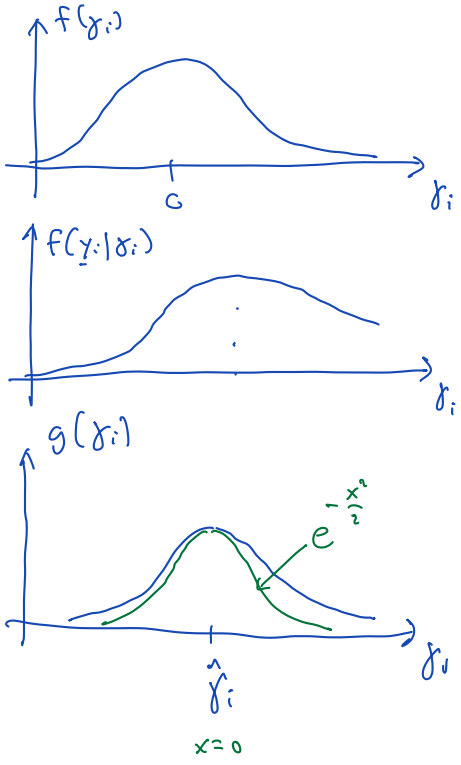
$$r(x_i) = f(x_i) \text{ for } i = 1, \ldots, n. \qquad (4)$$

Eqs. (1) – (4) then leads to

$$\int_{-\infty}^{\infty} f(x)\, w(x)\, dx \underset{(3)}{=} \int r(x)\, w(x)\, dx \underset{(1)}{=} \sum_{i=1}^{n} w_i\, r(x_i) \underset{(4)}{=} \sum_{i=1}^{n} w_i\, f(x_i)$$

Note: Exact as long as $f(x)$ is a polynomial of at most order $2n-1$.

Adaptive Gauss-Hermite quadrature (Liu & Pierce, 1994)



Marginal likelihood, clustered data

$$L(\beta,\theta) = f(\underline{y}_1, \underline{y}_2, \ldots, \underline{y}_m)$$

$$= \prod_{i=1}^{m} f(\underline{y}_i)$$

$$= \prod_{i=1}^{m} \left(\int_{R^1} \underbrace{\left(\prod_{j=1}^{n_i} f(y_{ij}|\gamma_i) \right)}_{\text{approx. Gaussian if } n_i \text{ is large}} \underbrace{f(\gamma_i)}_{\text{Gaussian}} d\gamma_i \right)$$

$$\underbrace{\qquad\qquad}_{\approx g(\gamma) \text{ approx. Gaussian}}$$

Letting $\hat{\gamma} = \underset{\gamma}{\operatorname{argmax}}\, g(\gamma)$, $h = -\frac{\partial^2}{\partial \gamma^2} \ln g(\gamma)\Big|_{\gamma=\hat{\gamma}}$ and $x = \frac{\gamma - \hat{\gamma}}{1/\sqrt{h}}$,

such that $d\gamma = \sqrt{h}\, dx$ and $\gamma = \hat{\gamma} + \frac{x}{\sqrt{h}}$, each integral

$$\int g(\gamma)\, d\gamma = \sqrt{h} \int g\left(\hat{\gamma} + \frac{x}{\sqrt{h}}\right) dx$$

$$= \int f(x)\, e^{-\frac{x^2}{2}}\, dx \quad \text{where} \quad f(x) = \frac{\sqrt{h}\, g\left(\hat{\gamma} + \frac{x}{\sqrt{h}}\right)}{e^{-x^2/2}} \left\{ \begin{array}{l} \text{deviation} \\ \text{of } g \text{ from} \\ \text{Gaussianity} \end{array} \right.$$

$$\approx \sum_{i=1}^{nAQP} w_i f(x_i) \qquad (5)$$

Used by lme4::glmer(, nAQP =)

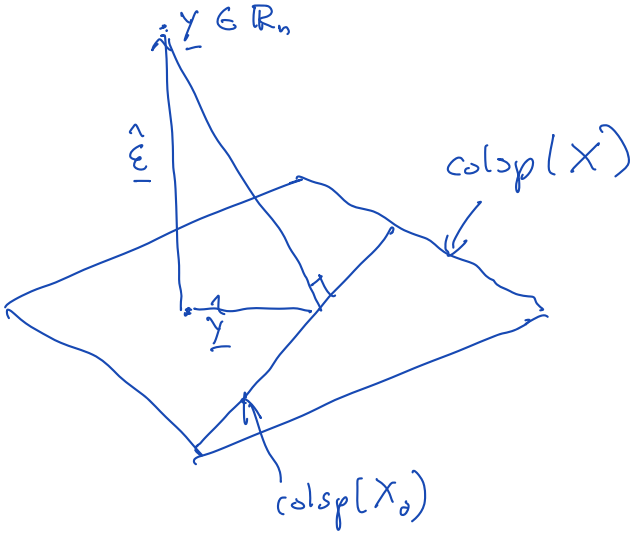
Note that for nAQP = 1, $w_1 = \int e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}$, $x_1 = 0$ (1b) and (5) simplifies to $\sqrt{2\pi h}\, g(\hat{\gamma})$, the Laplace approximation of $\int g(\gamma) d\gamma$

Summary (november 16)

LMs

$$\underset{\sim}{y} = X\underline{\beta} + \underline{\varepsilon}, \quad \underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I_n)$$

$$= \underset{\sim}{x}_1\beta_1 + \underset{\sim}{x}_2\beta_2 + \ldots + \underset{\sim}{x}_p\beta_p + \underline{\varepsilon}$$



$$\hat{\underline{y}} = \underset{\hat{y} \in \text{colsp}(X)}{\text{argmin}} \|\hat{\underline{y}} - \underline{y}\|$$

$$= \text{proj}_{\text{colsp}(X)} \underline{y}$$

$$= H\underline{y} = \underbrace{X(X^TX)^{-1}X^T}_{\hat{\beta}_{OLS}}\underline{y}$$

General linear model

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}, \quad \underline{\varepsilon} \sim N(0, V)$$

$$\Updownarrow$$

$$\underline{y}^* = \underline{x}^*\beta + \underline{\varepsilon}^*, \quad \underline{\varepsilon}^* \sim N(0, I) \qquad y^* = V^{-1/2}y$$

$$\hat{\underline{\beta}}_{GLS} = (X^TV^{-1}X)^{-1}X^TV\underline{y}$$

LMM

$$\underline{y} = X\underline{\beta} + U\underline{\gamma} + \underline{\varepsilon}, \quad \underline{\gamma} \sim N(0, G), \quad \underline{\varepsilon} \sim N(0, R)$$

For example, random slope, intercept model

$$y_{ij} = (\beta_0 + \gamma_{0,i}) + (\beta_1 + \gamma_{1,i}) x_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

for $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n_i$

$$\begin{bmatrix} \gamma_{0,i} \\ \gamma_{1,i} \end{bmatrix} \overset{iid}{\sim} N\left(\underline{0}, \begin{bmatrix} \tau_0^2 & \tau_{01} \\ \tau_{01} & \tau_1^2 \end{bmatrix} \right)$$

ML:

$$\ell_p(\theta) = \max_{\beta} \ell(\beta, \theta) = \ell(\hat{\beta}(\theta), \theta)$$

$$\hat{\beta}(\theta) = \left(X^T V^{-1}(\theta) X \right)^{-1} X^T V^{-1}(\theta) \underline{y} \quad \text{where} \quad V(\theta) = U G_1(\theta) U^T + R(\theta)$$

REML:

$$L_R(\theta) \overset{\text{def}}{=} \int L(\beta, \theta) \, d\beta$$

$$= |X^T X|^{-\frac{1}{2}} \; f_{\tilde{w}}\left(\underset{(n-p)\times n}{A^T} \underline{y} \right)$$

$$= L_p(\theta) \left| X^T V(\theta)^{-1} X \right|^{-\frac{1}{2}}$$

GLMs. For $i = 1, 2, \ldots, n$

$y_i \sim \text{bin}(n_i, \pi_i)$

Link function $\eta_i = g(\pi_i)$ Response function

$\text{logit}\ \pi_i = \ln \frac{\pi_i}{1-\pi_i}$ $\qquad \pi_i = \frac{1}{1+e^{-\eta_i}}$

$\text{probit}\ \pi_i = \Phi^{-1}(\pi_i)$ $\qquad \pi_i = \Phi(\eta_i)$

$\text{cloglog} = \ln(-\ln(1-\pi_i))$ $\qquad \pi_i = 1 - e^{-e^{\eta_i}}$

$y_i \sim \text{Poisson}(\lambda_i)$

$\log(\lambda_i) = \eta_i$ $\qquad \lambda_i = e^{\eta_i}$

cloglog examples

~ Presence - absense data

Poisson process: $z_i \sim \text{Poisson}(\lambda_i)$, $\ln \lambda_i = \eta_i$

Suppose we only observe

$$\pi_i = P(y_i = 1) = P(z_i \geq 1)$$

$$= 1 - P(z_i = 0)$$

$$= 1 - \frac{e^{-\lambda_i} \lambda_i^0}{0!} = 1 - e^{-\lambda_i}$$

$$= 1 - e^{-e^{\eta_i}}$$

- Survival data : $T_i \sim$ Weibull

$$\pi_i = P(\text{Dead after a time interval of length } t_i)$$

$$= P(T_i < t_i)$$

$$= 1 - e^{-a t_i^b}$$

$$\eta_i = \beta_0 + \beta_1 \ln t_i$$

$$\ln a + b \ln t_i$$

$$= 1 - e^{-e^{\eta_i}}$$

Probit link

Age of menarche $T_i \sim N(\mu, \sigma^2)$

$$\pi_i = P(y_i = 1) = P(T_i \le t_i)$$

$$= \Phi\left(\frac{t_i - \mu}{\sigma}\right)$$

$$= \Phi\left(-\frac{\mu}{\sigma} + \frac{1}{\sigma} t_i\right)$$

β_0 β_1

Multinomial models

$$\underline{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ic} \end{bmatrix} \sim \text{Multinomial}\left(n_i, \underline{\pi}_i\right) \qquad \underline{\pi}_i = \begin{bmatrix} \pi_{i1} \\ \pi_{i2} \\ \vdots \\ \pi_{ic} \end{bmatrix}$$

Nominal models

non-parallell slopes

$$\pi_{ir} = \frac{e^{\underline{x}_i^T \underline{\beta}_r}}{1 + \sum_{s=1}^{c} e^{\underline{x}_i^T \underline{\beta}_s}} = P\left(u_{ir} - \max_{s \neq r} u_{is} < 0\right)$$

$$u_{ir} = \underline{x}_i^T \underline{\beta}_r + \varepsilon_{ir}$$

$\varepsilon_{ir} \sim$ Gumbel

Ordinal models (example tomorrow)

$$\text{logit}\left(Y \leq r\right) = \theta_r + \underline{x}_i^T \underline{\beta}$$

parallell slopes

GLMMs

Clustered data (unlike problem 2 in project 3)

$$y_{ij} \mid \underline{\gamma}_i \sim \text{bin}(n_{ij}, \pi_{ij})$$

or

$$y_{ij} \mid \underline{\gamma}_i \sim \text{Poisson}(\lambda_{ij})$$

and

$$\underline{\gamma}_i \sim N(\underline{0}, Q)$$

Marginal likelihood

$$L(\beta, \theta) = f(\underline{y})$$

LTP

$$= \int f(\underline{y}, \underline{\gamma}) \, d\underline{\gamma}$$

$$= \int f(\underline{y} \mid \underline{\gamma}) f(\underline{\gamma}) \, d\underline{\gamma}$$

non-Gaussian · Gaussian

$\approx$ Laplace approximation

= Gauss-Hermite quadrature

Difference between marginal and conditional model

Possible forms of mathematical notation

Symbolic (Wilkinson & Rodgers) notation

$$\sim 1 + x + j + x : i$$

or

$$\sim age + sex + age : sex$$

- age: ↑ numeric
- sex: ↑ factor with k levels
- age:sex: interaction

Option 1: For $i = 1, 2, \ldots, n$

$$\eta_i = \mu + \beta \cdot x_i + \alpha_{j(i)} + \gamma_{j(i)} x_i$$

(see e.g. Wood pp. 194 – 195)

Option 1b:

$$\eta_i = \mu + \beta \cdot age_i + \alpha_{sex(i)} + \gamma_{sex(i)} \cdot age_i$$

Option 2: For $j = 1, 2, \ldots, k$ and $i = 1, 2, \ldots, n_j$

$$\eta_{ij} = \mu + \beta \cdot x_{ij} + \alpha_j + \gamma_j \cdot x_{ij}$$

Option 3:

\boldsymbol{\beta}

$$\underline{\eta} = X\underline{\beta}$$

If we have two levels $k = 2$

$$= \begin{bmatrix} 1 & & \\ 1 & & \\ 1 & & \\ \vdots & & \\ 1 & & \\ 1 & & \end{bmatrix} \quad 2. \quad \begin{bmatrix} \mu \\ \beta \\ \alpha_2 \\ \gamma_2 \end{bmatrix}$$

Project 3

y_{ijk} : Number of goals of team i against team j when playing away ($k=2$) or at home ($k=1$)

Model : For $i = 1, \dots, 16$, $j = 1, \dots, 16$, $i \neq j$ and $k = 1, 2$

$$y_{ijk} \mid \beta_i, \gamma_j \sim \text{Poisson}(\lambda_{ijk})$$

where (fixed effects; random effects (crossed))

$$\ln \lambda_{ijk} = \mu + \alpha_k + \beta_i + \gamma_j$$

- α_k: away vs. home
- β_i: attacking team
- γ_j: defending team

where

$$\beta_i \overset{iid}{\sim} N(0, \tau_\beta^2) \quad \text{and} \quad \gamma_j \overset{iid}{\sim} N(0, \tau_\gamma^2)$$

Alternatively; For $i = 1, 2, \dots, 480$

$$y_i \sim \text{Poisson}(\lambda_i)$$

where

$$\ln \lambda_i = \mu + \alpha_{k(i)} + \beta_{j(i)} + \gamma_{l(i)}$$

Alternative 2:

$$\underline{y} \mid \underline{\gamma} \sim \text{Poisson}(\underline{\eta})$$

$$\underline{\eta} = X\underline{\beta} + U\underline{\gamma}, \quad \underline{\gamma} \sim N(\underline{0}, G)$$

TMA4315, 2006, problem 1

19 psychotic male ($x_{i2} = 1$) and female ($x_{i2} = 0$) patients with initial CGP = x_{i3} given injection every second week ($x_{i1} = 0$) or every third week ($x_{i1} = 1$). $\underline{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \end{bmatrix}$ number of patients with final CGP = 0, 1, 2.

a) Formulate prop. odds. model with no interactions

Response for $i = 1, 2, \ldots, 12$ is

$$\underline{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3})^T \sim \text{Multinom}(n_i, \underline{\pi}_i)$$

where

$$\underline{\pi}_i = (\pi_{i1}, \pi_{i2}, \pi_{i3})^T$$

$c + 1 = 3$ categories

Covariates

$$\underline{x}_i = (x_{i1}, x_{i2}, x_{i3})^T$$

- x_{i1}: freq. of medical treatment
- x_{i2}: sex
- x_{i3}: initial CGI

a) Proportional odds model

$$P(CGI_i \leq r) = P(-\underline{x}_i^T \underline{\beta} + u_i \leq \theta_r)$$

$u_i \sim \text{logistic}(0, 1)$

$$= P(u_i \leq \theta_r + \underline{x}_i^T \underline{\beta})$$

$$= \text{logit}^{-1}(\theta_r + \underline{x}_i^T \underline{\beta})$$

$$= \frac{1}{1 + e^{-(\theta_r + \underline{x}_i^T \underline{\beta})}}$$

$$\underset{-\infty}{\theta_0} < \theta_1 < \theta_2 < \underset{\infty}{\theta_3}$$

Unknown parameters: $\theta_1, \theta_2, \beta_1, \beta_2, \beta_3$

$$\pi_{ir} = P(CGI_i = r)$$

$$= P(CGI_i \leq r) - P(CGI_i \leq r-1)$$

$$= \frac{1}{1 + e^{-\theta_r - \underline{x}_i^T \underline{\beta}}} - \frac{1}{1 + e^{-\theta_{r-1} - \underline{x}_i^T \underline{\beta}}}$$

b) The model is that

$$P(CGI \leq r) = \text{logit}^{-1}(\theta_r + \underline{x}_i^T \underline{\beta})$$

$$\text{logit}\, P(CGT \le r) = \theta_r + \underline{x}_i^T \underline{\beta}$$

$$\ln \frac{P(CGT \le r)}{P(CGT > r)} = \theta_r + \underline{x}_i^T \underline{\beta}$$

$$\frac{P(CGT \le r)}{P(CGT > r)} = e^{\theta_r + \underline{x}_i^T \underline{\beta}}$$

If x_{i3} increases by one unit the cumulative odds changes by a factor (an odds ratio) of e^{β_3}.

c) Saturated model

(π_{i1}, π_{i2}), $i = 1, 2, \ldots, 12$

free parameters ($p_1 = 24$)

with MLEs $\hat{\pi}_{ij}^{(sat)} = \frac{y_{ij}}{n_i}$

The deviance is given by

$$D = 2\left[l\left(\hat{\pi}_{sat}\right) - l\left(\hat{\pi}_{fitted}\right) \right]$$

$$= 2 \sum_{i=1}^{12} \sum_{j=1}^{3} Y_{ij} \ln\left(\frac{Y_{ij}/n_i}{\hat{\pi}_{ij}} \right)$$

The deviance (under H_0) is chi-square with

$$p_1 - p_0 = 24 - 5 = 19$$

according to asymtotic theory.

d) Model in symbolic notation:

B * K + B * I

B + K + B:K + I + B:I

In mathematical notation the model is that

$$\text{logit}\left(P(CGI \le r)\right) = \theta_r + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

(2 ↑ B, K, I)

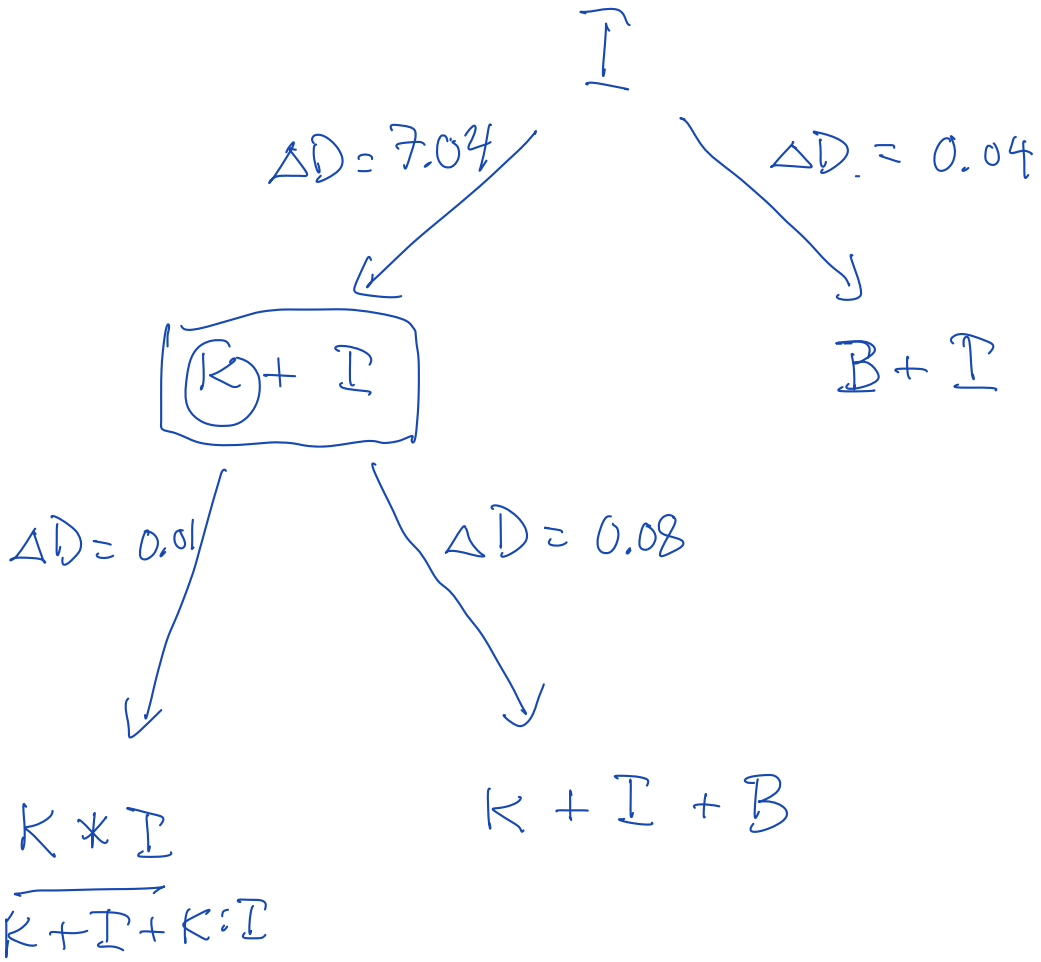
$$+ \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$

This model has $2 + 5 = 7$

The deviance for this model $24 - 7 = 17$

Critical value

$$\chi^2_{0.05,1} = 3.84$$



The best model (based on hypothesis testing) is $K+I$

Could have done model selection via AIC.

e) Estimates of e^{β_k}, $k = 1, 2, 3$.

$$\widehat{e^{\beta_1}} = e^{\hat{\beta}_1} = e^{-0.21} = 0.8026$$

Confidence interval for β_1 based on

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} \overset{approx}{\sim} N(0,1)$$

$$\hat{\beta}_1 \pm 1.96 \cdot \widehat{SE(\beta_1)} = -0.21 \pm 1.96 \cdot 0.75$$

$z_{\alpha/2}$

$= (-0.18, 3.53)$

Confidence interval for e^{β_1}

Recall that if (A, B) is a conf. int for θ, then

$$P(A < \theta < B) = 1 - \alpha$$

and

$$P(g(A) < g(\theta) < g(B)) = 1 - \alpha$$

such that

$(g(A), g(B))$ is a $(1-\alpha)$ − conf. int for $g(\theta)$

Hence, $(e^{0.18}, e^{3.53}) = (0.18, 3.53)$

$$\hat{\pi}_{ij} = P(CGI_i = 1)$$

$$= P(CGI_i \leq 1)$$

$$= \frac{1}{1 - e^{-\underbrace{\left(\hat{\theta}_1 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3\right)}_{\hat{\eta}_{ir}}}}$$

$$= \frac{1}{1 - e^{-\underbrace{\left(8.47 + -0.21 \cdot 0 + -2.45 \cdot 0 - 1.85 \cdot 5\right)}_{r}}}$$

$$= \dots$$

$$\mathrm{Var}\left(\hat{\theta}_1 + \underline{x}_i^T \hat{\underline{\beta}}\right)$$

$$= \mathrm{Var}\left(\underbrace{\begin{bmatrix} 1 & 0 & \underline{x}_i^T \end{bmatrix}}_{A} \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}\right)$$

$$= A \, \mathrm{Var}\left(\begin{bmatrix} \hat{\underline{\theta}} \\ \hat{\underline{\beta}} \end{bmatrix}\right) A^T$$

$$\| \quad \leftarrow \mathrm{vcov}(\mathrm{mod})$$

$$F^{-1}\left(\begin{bmatrix} \hat{\underline{\theta}} \\ \hat{\underline{\beta}} \end{bmatrix}\right)$$

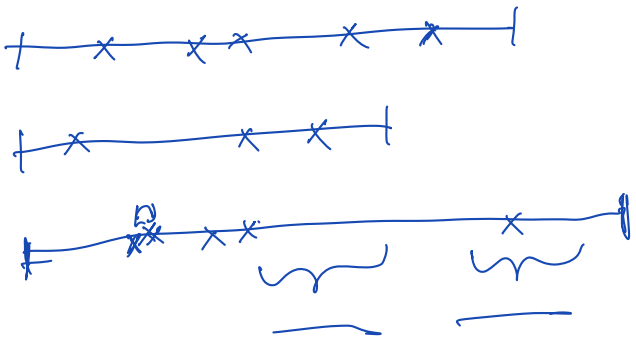
Delta method: Since $\hat{\pi}_{it} = h(\hat{\eta}_{it})$

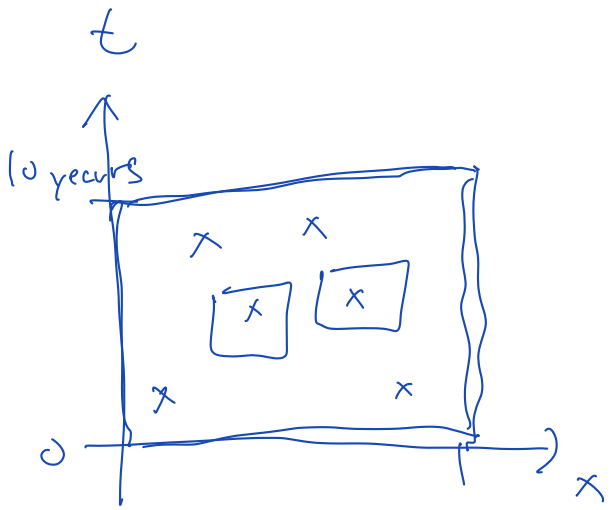
$$\mathrm{Var}(\hat{\pi}_{it}) = \left(\left. \frac{dh}{d\eta} \right|_{\eta = \hat{\eta}_{it}} \right)^2 \mathrm{Var}(\hat{\theta} + \underline{x}_i^T \underline{\beta})$$

ST2304 2016, problem 3

a)

SFSSFFFSSS (X)





log-link: ensures that the poisson parameter > 0

The model is

$$\ln \lambda_i = \underline{x}_i^T \underline{\beta} + \log(\text{length}_i)$$

$$\lambda_i = e^{\underline{x}_i^T \underline{\beta} + \log(\text{length}_i)}$$

$$= e^{\underline{x}_i^T \underline{\beta}} \cdot \text{length}_i$$

The offset term means that we assume direct proportionality between λ_i and length_i.

b) Changing the speedlimit from 80 to 70 changes $Ey_i = \lambda_i$ by a factor of

$$e^{\hat{\beta}_1(\ln 70 - \ln 80)} = e^{\hat{\beta}_1 \ln \frac{70}{80}}$$

$$= \left(\frac{7}{8}\right)^{\hat{\beta}_1} = \left(\frac{7}{8}\right)^{4.92} = 0.51$$

i.e. 49% reduction in number of

accidents.

Absolute change greatest for roads surrounded by woodland.

c) Overdispersion

- Positive covariance between subintervals

- Missing covariates

(Unexplained variation)

- Incorrect link function

Since $D < \chi^2_{0.05, 295}$ we don't reject H_0: no overdispersion

qchisq(0.95, df = 295) ≈ 350

TMA4315 2019, Problem 2

a) The model assumes that

$$Y_i \sim \text{Poisson}(\lambda_i)$$

and independent where

$$\ln \lambda_i = \beta_0 + \beta_1 x_i + \log(\text{area}_i)$$

where β_0 and β_1 are unknown parameters

If altitude x_i increases by 1000m, the expected density (plants $\cdot m^{-2}$)

$$E\left(\frac{Y_i}{\text{area}_i}\right) = \frac{\lambda_i}{\text{area}_i} = e^{\beta_0 + \beta_1 x_i}$$

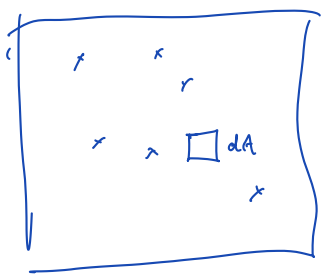
changes by a factor of $e^{\beta_1 \cdot 1000} = e^{-0.58} = 0.56$,

i.e., plant density is reduced by 44%.

b) Poisson assumption follows if

Individual plants occurances

in spatial Poisson process



Indep. between disjoint subareas

Constant intensity

$P(\text{one occurance in } dA) = \lambda\, dA$

Offset term:

A priori $\lambda_i = E Y_i \propto \text{area}_i$

Model $\ln \lambda_i = \beta_0 + \beta_1 x_i + \ln(\text{area}_i)$

$$\lambda_i = \text{area}_i \cdot e^{\beta_0 + \beta_1 x_i}$$

c) Overdispersion? Under H_0 (no overdisp.)

$$D \sim \chi^2_{100-2}$$

$$\chi^2_{0.95,98} = 122.1$$

qchisq

Obs. val

$$D = 235.$$

Reject H_0 in favor of H_1: overdisp.

Mechanisms: Non-indep.
Missing covariates
Wrong link function

Estimated disp. par. φ is

$$\hat{\varphi} = \frac{D}{n-p} = \frac{235}{98} = 2.4$$

This changes Wald 2-test to Wald T-test with

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\varphi}}\, SE(\hat{\beta}_1)} = \frac{-0.00058}{\sqrt{2.4} \cdot 0.00045} = -0.909$$

Since $|T| < t_{0.025,98}$ we keep $H_0: \beta_1 = 0$

d) • Including year as a random effect, the model is that

$$Y_i \mid \gamma_{year(i)} \sim Poisson(\lambda_i)$$

where

$$\ln \lambda_i = \beta_0 + \beta_i x_i + \gamma_{year(i)} + \ln(area_i)$$

and

$$\gamma_{year} \overset{iid}{\sim} N(0, \tau_0^2) \quad \text{for year} = 2001, 2002, \ldots, 2020.$$

Alternative notation (Fahrmeir)

$$Y_{ij} \mid \gamma_i \sim \text{Poisson}(\lambda_{ij})$$

where

$$\lambda_{ij} = \beta_0 + \beta_1 x_{ij} + \ln(\text{area}_{ij}) + \gamma_i$$

for years (clusters) $i = 1, \ldots, m$ and obs. $j = 1, 2, \ldots, n_i$

- MLE of $\tau_0^2 = \text{Var}(\gamma_{year})$ is $\tau_0^2 = 0.44$
- Given $\gamma_{year} = 0$, for $x_i = 0$ (at sea level) and $\text{area}_i = 50$

$$E(Y_i \mid \gamma_{year(i)} = 0) = e^{\beta_0 + \ln(50)} = e^{-2.77} \cdot 50 = 3.13$$

Choosing a year at random

$$EY_i = E\,E(Y_i \mid \gamma)$$

$$= E\, e^{\beta_0 + \ln 50 + \gamma}$$

$$= 3.13\, e^{\gamma} = 3.13\, e^{0 + \tau_0^2/2} = 3.90$$

e) Does the random year effect sign. improve model?

- $H_0: \tau_0^2 = 0$
  $H_1: \tau_0^2 > 1$ (H_0 at boundary)

- Under H_0

$$E\left(\underline{s}(\beta_1, \tau_0^2)\right) = E\left(\left(\frac{\partial}{\partial \beta_0}, \frac{\partial}{\partial \beta_1}, \frac{\partial}{\partial \tau_0^2}\right)^T \ell\right) = \underline{0} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

- $LRT = 2\left(\ell(H_1) - \ell(H_0)\right) \sim 0.5\chi_0^2 : 0.5\chi_1^2$

- Rejection region. We want

$$P(\text{Type I error}) = \alpha$$

$$P(LRT > c \mid H_0) = \alpha$$

$$\frac{1}{2}\underbrace{P(\chi_0^2 > c)}_{=0} + \frac{1}{2}P(\chi_1^2 > c) = \alpha$$

$$P(\chi_1^2 > c) = 2\alpha$$

$$c = \chi^2_{2\alpha,1} = \chi^2_{0.10,1} = 6.63$$

↑ tabell

Observed value

$$LRT = 2\left(-177.5 - (-225.1)\right) = 96$$

Concl.: Reject H_0 in favour of H_1

ST2304, 2013, problem 3

a) count ~ weekend + precip
= yes, no (2 levels)
} H_0, $p_0 = 3$

b) count ~ weekday + precip
= mon, tue, ..., sun (7 levels)
} H_1, $p_1 = 8$

H_0 is nested in H_1 since H_0 is equivalent to H_1 when $\beta_{mon} = \beta_{tue} = \ldots = \beta_{fri}$ and $\beta_{sat} = \beta_{sun}$

Alternative view: Design matrices

$$X_0 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix} \quad X_1 = \begin{bmatrix} 1 & & & & & & \\ 1 & 1 & & & & & \\ 1 & & 1 & & & & \\ 1 & & & 1 & & & \\ 1 & & & & 1 & & \\ 1 & & & & & 1 & \\ 1 & & & & & & 1 \end{bmatrix}$$

$\Rightarrow \text{Colsp}(X_0) \subset \text{Colsp}(X_1)$

$\Rightarrow$ H_0 is nested in H_1.

Saturated model has $X_s = I_n$ with $\text{colsp}(X_s) = \mathbb{R}^n$.

All models nested in satur. mod.

ST2304, august 2015, problem 2 e, f, g

hemoglobin explained by