Introduction to General and Generalized Linear Models

Course Summary (plus integrated models)

Henrik Madsen
Jan Kloppenborg Møller
Anders Nielsen

May 5, 2012

This lecture

- Course Summary
- .....
- Integrated models

What have we been doing?

- Likelihood principle
- General linear models
- Generalized linear models
- General mixed effects models
- Repeated measurements
- Random effects models
- Hierarchical models
- Crossed and nested models
- Heteroscedasticity and correlation structures
- Points on using R

The book covers a lot more than its title, and we went beyond that.

Likelihood inference

- Likelihood function $L(\theta) = P_\theta(Y = y)$
- Log likelihood function $\ell(\theta) = \log(L(\theta))$
- Score function $\ell'(\theta)$
- Maximum likelihood estimate $\widehat{\theta} = \underset{\theta \in \Theta}{\text{argmax}}\ \ell(\theta)$
- Observed information matrix $-\ell''(\widehat{\theta})$
- Distribution of the ML estimator $\widehat{\theta} \sim \mathrm{N}(\theta, (-\ell''(\widehat{\theta}))^{-1})$
- Likelihood ratio test $2(\ell_A(\widehat{\theta_A}, Y) - \ell_B(\widehat{\theta_B}, Y)) \sim \chi^2_{\dim(A) - \dim(B)}$
- Invariance property
- Dealing with nuisance parameters

Likelihood inference - When we use it

- Indirectly all the time
- Directly when no prepackaged tool is available

Likelihood inference - How we do it

- State the model
- Write the (negative log) likelihood contribution
- Optimize the likelihood for data w.r.t. model parameters
- Optimum gives the parameter estimate
- Curvature quantifies uncertainty
- Likelihood value can be used to compare models
- Example (from last time):

$$Y_i \sim NB(\alpha, 1/(1+\beta))$$

General Linear Model

- A general linear model is:

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

Consider the well known two way ANOVA:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{ i.i.d. } N(0, \sigma^2), \quad i = 1, 2, \quad j = 1, 2, 3.$$

An expanded view of this model is:

$$\begin{array}{lllllllll}
y_{11} = \mu & + \alpha_1 & & + \beta_1 & & & + \varepsilon_{11} \\
y_{21} = \mu & & + \alpha_2 & + \beta_1 & & & + \varepsilon_{21} \\
y_{12} = \mu & + \alpha_1 & & & + \beta_2 & & + \varepsilon_{12} \\
y_{22} = \mu & & + \alpha_2 & & + \beta_2 & & + \varepsilon_{22} \\
y_{13} = \mu & + \alpha_1 & & & & + \beta_3 & + \varepsilon_{13} \\
y_{23} = \mu & & + \alpha_2 & & & + \beta_3 & + \varepsilon_{23}
\end{array} \tag{1}$$

The exact same in matrix notation:

$$\underbrace{\begin{pmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \\ y_{13} \\ y_{23} \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{pmatrix}}_{\boldsymbol{\varepsilon}} \tag{2}$$

$$\underbrace{\begin{pmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \\ y_{13} \\ y_{23} \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

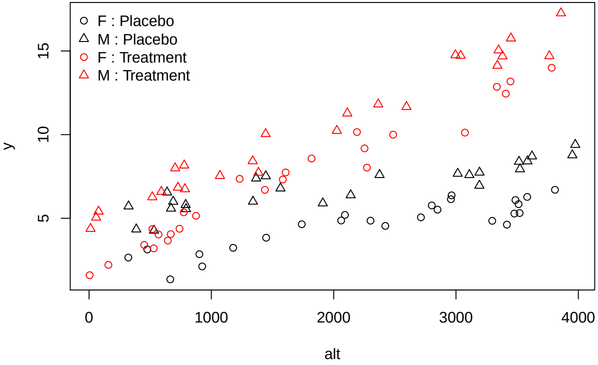
- $\mathbf{y}$ is the vector of all observations
- $\mathbf{X}$ is known as the *design matrix*
- $\boldsymbol{\beta}$ is the vector of parameters
- $\boldsymbol{\varepsilon}$ is a vector of independent $N(0, \sigma^2)$ “measurement noise”
    - The vector $\boldsymbol{\varepsilon}$ is said to follow a *multivariate normal distribution*
    - Mean vector $\mathbf{0}$
    - Covariance matrix $\sigma^2\mathbf{I}$
    - Written as: $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$
- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ specifies the model, and everything can be calculated from $\mathbf{y}$ and $\mathbf{X}$.

General Linear Model - when we use it

- When our observations are normally distributed
- When a simple transformation (e.g. logarithm) can make our observations normally distributed
- When our model prediction is a linear function of our model parameters

General Linear Model - how we use it

Consider this dataset:



Remember our talks about model formulation

- How a statement like this

  *> fit0<-lm(y~sex*tmt+sex*tmt*alt)*

- Is really the model

$$y_i = \mu+\alpha(\mathsf{S}_i)+\beta(\mathsf{T}_i)+\gamma(\mathsf{S}_i,\mathsf{T}_i)+\delta(\mathsf{S}_i)\cdot\mathsf{a}_i+\phi(\mathsf{T}_i)\cdot\mathsf{a}_i+\psi(\mathsf{S}_i,\mathsf{T}_i)\cdot\mathsf{a}_i+\varepsilon_i$$

- Which is over-parametrized, and really the same as:

$$y_i = \gamma(\mathsf{S}_i,\mathsf{T}_i) + \psi(\mathsf{S}_i,\mathsf{T}_i)\cdot\mathsf{a}_i + \varepsilon_i$$

- But we use the long form to be able to test for model reductions

R example

```
> fit0<-lm(y~sex*tmt+sex*tmt*alt)
> drop1(fit0,test='F')

Single term deletions

Model:
y ~ sex * tmt + sex * tmt * alt
            Df Sum of Sq    RSS     AIC F value  Pr(F)
<none>                   42.983 -68.437
sex:tmt:alt  1  0.077585 43.060 -70.257  0.1661 0.6846
```

```
> fit1<-lm(y~sex*tmt+(sex+tmt)*alt)
> drop1(fit1,test='F')
```

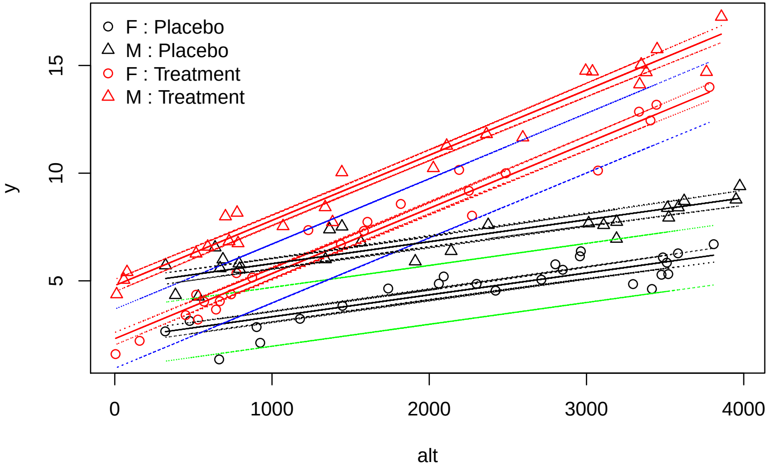
```
Single term deletions

Model:
y ~ sex * tmt + (sex + tmt) * alt
        Df Sum of Sq     RSS     AIC  F value  Pr(F)
<none>                43.060 -70.257
sex:tmt  1     0.245  43.305 -71.690   0.5287 0.4690
sex:alt  1     0.848  43.909 -70.306   1.8324 0.1791
tmt:alt  1   143.386 186.446  74.297 309.6786 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> fit2<-lm(y~sex+tmt+(sex+tmt)*alt)
> drop1(fit2,test='F')

Single term deletions

Model:
y ~ sex + tmt + (sex + tmt) * alt
        Df Sum of Sq     RSS     AIC  F value  Pr(F)
<none>                43.305 -71.690
sex:alt  1     0.694  43.999 -72.101   1.5054 0.2229
tmt:alt  1   143.628 186.933  72.558 311.7645 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> fit3<-lm(y~sex+tmt*alt)
> fit4<-lm(y~sex+tmt:alt)
> drop1(fit3,test='F')

Single term deletions

Model:
y ~ sex + tmt * alt
        Df Sum of Sq     RSS     AIC F value     Pr(F)
<none>                43.999 -72.101
sex      1    150.34 194.338  74.443  324.61 < 2.2e-16 ***
tmt:alt  1    143.95 187.946  71.099  310.80 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(fit4,fit3)

Analysis of Variance Table

Model 1: y ~ sex + tmt:alt
Model 2: y ~ sex + tmt * alt
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     96 44.005
2     95 43.999  1 0.0061976 0.0134 0.9082
```

```
> fit4<-lm(y~sex+tmt:alt)
> drop1(fit4,test='F')

Single term deletions

Model:
y ~ sex + tmt:alt
        Df Sum of Sq    RSS     AIC F value     Pr(F)
<none>                44.00 -74.087
sex      1    151.90 195.90  73.245  331.38 < 2.2e-16 ***
tmt:alt  2    950.58 994.59 233.716 1036.88 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Results



F : Placebo
M : Placebo
F : Treatment
M : Treatment
y
alt
15
10
5
0
1000
2000
3000
4000

Exponential families of distributions

Consider a univariate random variable Y with a distribution described by a family of densities $f_Y(y;\theta),\ \ \theta \in \Omega$.

Definition (A natural exponential family)

A family of probability densities which can be written on the form

$$f_Y(y;\theta) = c(y)\exp(\theta y - \kappa(\theta)),\ \ \theta \in \Omega$$

is called a *natural exponential family* of distributions. The function $\kappa(\theta)$ is called the *cumulant generator*. This representation is called the *canonical parametrization* of the family, and the parameter θ is called the *canonical parameter*.

Exponential families of distributions

Definition (An exponential dispersion family)

A family of probability densities which can be written on the form

$$f_Y(y;\theta) = c(y,\lambda)\exp(\lambda\{\theta y - \kappa(\theta)\})$$

is called an *exponential dispersion family* of distributions. The parameter $\lambda > 0$ is called the *precision parameter*.

- Basic idea: separate the mean value related distributional properties described by the *cumulant generator* $\kappa(\theta)$ from features as sample size, common variance, or common over-dispersion.
- In some cases the precision parameter represents a known number of observations as for the binomial distribution, or a known shape parameter as for the gamma (or χ^2-) distribution.
- In other cases the precision parameter represents an unknown dispersion like for the normal distribution, or an over-dispersion that is not related to the mean.

Example: Poisson distribution

Consider $Y \sim \text{Pois}(\mu)$. The probability function for Y is:

$$\begin{aligned} f_Y(y;\mu) &= \frac{\mu^y e^{-\mu}}{y!} \\ &= \frac{1}{y!}\exp\{y\log(\mu)-\mu\} \end{aligned}$$

Comparing with the equation for the natural exponential family it is seen that $\theta = \log(\mu)$ which means that $\mu = \exp(\theta)$.

Thus the Poisson distribution is a special case of a natural exponential family with canonical parameter $\theta = \log(\mu)$, cumulant generator $\kappa(\theta) = \exp(\theta)$ and $c(y) = 1/y!$.

The natural exponential family: $f_Y(y;\theta) = c(y)\exp(\theta y - \kappa(\theta))$

The Generalized Linear Model

Definition (The generalized linear model)

Assume that $Y_1, Y_2, \ldots, Y_n$ are mutually independent, and the density can be described by an exponential dispersion model with the same variance function $V(\mu)$.

A *generalized linear model* for $Y_1, Y_2, \ldots, Y_n$ describes an affine hypothesis for $\eta_1, \eta_2, \ldots, \eta_n$, where

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

is a transformation of the mean values $\mu_1, \mu_2, \ldots, \mu_n$.

The hypothesis is of the form

$$\mathcal{H}_0 : \boldsymbol{\eta} - \boldsymbol{\eta}_0 \in L,$$

where L is a linear subspace $\mathbb{R}^n$ of dimension k, and where $\boldsymbol{\eta}_0$ denotes a vector of *known off-set values*.

GLM vs GLM

General linear models	**Generalized linear models**
Normal distribution	Exponential dispersion family
Mean value linear	Function of mean value linear
Independent observations	Independent observations
Same variance	Variance function of mean
Easy to apply	Almost as easy to apply
Exact results	Approximate results

Generalized Linear Model - when we use it

- When observations are not following a normal distribution, but an exponential (dispersion) family
- When a link function of then mean can be expressed as a linear function of the model parameters

Specification of a generalized linear model in R

```
> mice.glm <- glm(formula = resp ~ conc,
+                 family = binomial(link = logit),
+                 weights = NULL,
+                 data = mice
+                )
```

- formula; as in general linear models
- family
  - binomial(link = logit | probit | cauchit | log | cloglog)
  - gaussian(link = identity | log | inverse)
  - Gamma(link = inverse | identity | log)
  - inverse.gaussian(link = 1/mu^2 | inverse | identity | log)
  - poisson(link = log | identity | sqrt)
  - quasi(link = ... , variance = ...))
  - quasibinomial(link = logit | probit | cauchit | log | cloglog)
  - quasipoisson(link = log | identity | sqrt)

Overdispersion

- It may happen that even if one has tried to fit a rather comprehensive model (i.e. a model with many parameters), the fit is not satisfactory, and the residual deviance $\mathrm{D}\left(\boldsymbol{y}; \boldsymbol{\mu}(\widehat{\boldsymbol{\beta}})\right)$ is larger than what can be explained by the χ^2-distribution.
- An explanation for such a poor model fit could be an improper choice of linear predictor, or of link or response distribution.
- If the residuals exhibit a random pattern, and there are no other indications of misfit, then the explanation could be that the variance is larger than indicated by $V(\mu)$.
- We say that the data are *overdispersed*.

Overdispersion

- When data are *overdispersed*, a more appropriate model might be obtained by including a *dispersion parameter*, σ^2, in the model, i.e. a distribution model of the form with $\lambda_i = w_i/\sigma^2$, and σ^2 denoting the overdispersion, $\mathrm{Var}[Y_i] = \sigma^2 V(\mu_i)/w_i$.
- As the dispersion parameter only would enter in the score function as a constant factor, this does not affect the estimation of the mean value parameters $\boldsymbol{\beta}$.
- However, because of the larger error variance, the distribution of the test statistics will be influenced.
- If, for some reasons, the parameter σ^2 had been known beforehand, one would include this known value in the weights, w_i.
- Most often, when it is found necessary to choose a model with overdispersion, σ^2 shall be estimated from the data.

The mixed linear model

Consider now the one way ANOVA with random block effect:

$$Y_{ij} = \mu + \alpha_i + B_j + \varepsilon_{ij}, \quad B_j \sim N(0, \sigma_B^2),\ \varepsilon_{ij} \sim N(0, \sigma^2),\ i = 1, 2,\ j = 1, 2, 3$$

The matrix notation is:

$$\underbrace{\begin{pmatrix} Y_{11} \\ Y_{21} \\ Y_{12} \\ Y_{22} \\ Y_{13} \\ Y_{23} \end{pmatrix}}_{\mathbf{Y}} = \underbrace{\begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}}_{\mathbf{X}} \underbrace{\begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}}_{\mathbf{Z}} \underbrace{\begin{pmatrix} B_1 \\ B_2 \\ B_3 \end{pmatrix}}_{\mathbf{U}} + \underbrace{\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \varepsilon_{13} \\ \varepsilon_{23} \end{pmatrix}}_{\boldsymbol{\varepsilon}}$$

Notice how this matrix representation is constructed in exactly the same way as for the fixed effects model — but separately for fixed and random effects.

A general linear mixed effects model

A general linear mixed model can be presented in matrix notation by:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{U} + \varepsilon, \quad \text{where } \mathbf{U} \sim N(0, \mathbf{G}) \text{ and } \varepsilon \sim N(0, \mathbf{R}).$$

- $\mathbf{Y}$ is the observation vector
- $\mathbf{X}$ is the design matrix for the fixed effects
- $\boldsymbol{\beta}$ is the vector containing the fixed effect parameters
- $\mathbf{Z}$ is the design matrix for the random effects
- $\mathbf{U}$ is the vector of random effects
  - It is assumed that $\mathbf{U} \sim N(\mathbf{0}, \mathbf{G})$
  - cov$(U_i, U_j) = G_{i,j}$ (typically $\mathbf{G}$ has a very simple structure (for instance diagonal))
- $\boldsymbol{\varepsilon}$ is the vector of residual errors
  - It is assumed that $\varepsilon \sim N(\mathbf{0}, \mathbf{R})$
  - cov$(\varepsilon_i, \varepsilon_j) = R_{i,j}$ (typically $\mathbf{R}$ is diagonal, but we shall later see some useful exceptions for repeated measurements)

Motivating example: Paired observations

- Two methods A and B to measure blood cell count (to check for the use of doping).
- Paired study.

Person ID	**Method A**	**Method B**
1	5.5	5.4
2	4.4	4.9
3	4.6	4.5
4	5.4	4.9
5	7.6	7.2
6	5.9	5.5
7	6.1	6.1
8	7.8	7.5
9	6.7	6.3
10	4.7	4.2

- It must be expected that two measurements from the same person are correlated, so a paired t-test is the correct analysis
- The t-test gives a p-value of 5.1%, which is a borderline result...
- But more data is available

- In addition to the planned study 10 persons were measured with only one method
- Want to use all data, which is possible with random effects
- Assume these 20 are ramdomly selected from a population where the blod cell count is normally distributed
- Consider the following model:
  $C_i = \alpha(M_i) + B(P_i) + \varepsilon_i, \;\; i = 1 \dots 30$
  $\alpha(M_i)$ the 2 fixed method effects
  $B(P_i) \sim \mathcal{N}(0, \sigma_P^2)$ the 20 rand. eff.
  $\varepsilon_i \sim \mathcal{N}(0, \sigma_R^2)$ measurement noise
  All $B(P_i)$ and ε_i are independent
- This model uses all data
- Allows us to test method difference

ID	Meth. A	Meth. B
1	5.5	5.4
2	4.4	4.9
3	4.6	4.5
4	5.4	4.9
5	7.6	7.2
6	5.9	5.5
7	6.1	6.1
8	7.8	7.5
9	6.7	6.3
10	4.7	4.2
11		3.4
12		4.7
13		3.9
14		2.5
15		4.1
16	4.0	
17	6.3	
18	6.0	
19	6.4	
20	3.5	

General Linear Mixed Model - when we use it

- When our observations are normally distributed
- When a simple transformation (e.g. logarithm) can make our observations normally distributed
- When our model prediction is a linear function of our model parameters
- When observational units are themselves sampled from a larger population (where normal assumption is OK)
- When it is helpful in expressing a needed covariance structure
- When we have repeated measurements

General (non-linear and/or non-normal) Mixed Models

The general mixed effects model can be represented by its likelihood function:

$$L_M(\boldsymbol{\theta}; \boldsymbol{y}) = \int_{\mathbb{R}^q} L(\boldsymbol{\theta}; \boldsymbol{u}, \boldsymbol{y}) \, d\boldsymbol{u}$$

- $\boldsymbol{y}$ is the observed random variables
- $\boldsymbol{u}$ is the q unobserved random variables
- $\boldsymbol{\theta}$ is the model parameters to be estimated

The likelihood function L is the joint likelihood of both the observed and the unobserved random variables.

The likelihood function for estimating $\boldsymbol{\theta}$ is the marginal likelihood L_M obtained by integrating out the unobserved random variables.

The Laplace approximation

$$\ell_M(\boldsymbol{\theta}, \boldsymbol{y}) \approx \ell(\boldsymbol{\theta}, \hat{\boldsymbol{u}}_{\boldsymbol{\theta}}, \boldsymbol{y}) - \frac{1}{2}\log(|\left(-\ell''_{uu}(\boldsymbol{\theta}, \boldsymbol{u}, \boldsymbol{y})|_{\boldsymbol{u}=\hat{\boldsymbol{u}}_{\boldsymbol{\theta}}}\right)|) + \frac{q}{2}\log(2\pi)$$

Formulation of hierarchical model

Theorem (Compound Poisson Gamma model)

Consider a hierarchical model for Y specified by

$$\begin{aligned} Y|\mu &\sim Pois(\mu), \\ \mu &\sim G(\alpha, \beta), \end{aligned}$$

i.e. a two stage model.

In the first stage a random mean value μ is selected according to a Gamma distribution. The Y is generated according to a Poisson distribution with that value as mean value. Then the the marginal distribution of Y is a negative binomial distribution, $Y \sim \mathrm{NB}(\alpha, 1/(1+\beta))$

Hierarchical Binomial-Beta distribution model

The natural conjugate distribution to the binomial is a Beta-distribution.

Theorem

Consider the generalized one-way random effects model for $Z_1, Z_2, \ldots, Z_k$ *given by*

$$\begin{aligned} Z_i | p_i &\sim B(n, p_i) \\ p_i &\sim \mathrm{Beta}(\alpha, \beta) \end{aligned}$$

i.e. the conditional distribution of Z_i *given* p_i *is a Binomial distribution, and the distribution of the mean value* p_i *is a Beta distribution. Then the marginal distribution of* Z_i *is a Polya distribution*
with probability function

$$P[Z = z] = g_Z(z) = \binom{n}{z} \frac{\Gamma(\alpha + x)}{\Gamma(\alpha)} \frac{\Gamma(\beta + n - z)}{\Gamma(\beta)} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha + \beta + n)}$$

for $z = 0, 1, 2, \ldots, n$.

Density for Y_i	Sufficient statistic $T(Y_1,\ldots,Y_n)$	Density for T	$\mathsf{E}[T\vert\theta]$	$\mathsf{V}[T\vert\theta]$
$\mathsf{Bern}(\theta)$	$\sum Y_i$	$\mathsf{B}(n,\theta)$	$n\theta$	$n\theta(1-\theta)$
$\mathsf{B}(r,\theta)$	$\sum Y_i$	$\mathsf{B}(rn,\theta)$	$rn\theta$	$rn\theta(1-\theta)$
$\mathsf{Geo}(\theta)$	$\sum Y_i$	$\mathsf{NB}(n,\theta)$	$n\frac{1-\theta}{\theta}$	$n\frac{1-\theta}{\theta}^2$
$\mathsf{NB}(r,\theta)$	$\sum Y_i$	$\mathsf{NB}(rn,\theta)$	$rn\frac{1-\theta}{\theta}$	$rn\frac{1-\theta}{\theta}^2$
$\mathsf{P}(\theta)$	$\sum Y_i$	$\mathsf{P}(n\theta)$	$n\theta$	$n\theta$
$\mathsf{P}(r\theta)$	$\sum Y_i$	$\mathsf{P}(rn\theta)$	$rn\theta$	$rn\theta$
$\mathsf{Ex}(\theta)$	$\sum Y_i$	$\mathsf{G}(n,\theta)$	$n\theta$	$n\theta^2$
$\mathsf{G}(\alpha,\theta)$	$\sum Y_i$	$\mathsf{G}(n\alpha,\theta)$	$\alpha n\theta$	$\alpha n\theta^2$
$\mathsf{U}(0,\theta)$	$\max Y_i$	$\mathsf{Inv\text{-}Par}(\theta,n)$	$\frac{n\theta}{n+1}$	$\frac{n\theta^2}{(n+1)^2(n+2)}$
$\mathsf{N}(\theta,\sigma^2)$	$\sum Y_i$	$\mathsf{N}(n\theta,n\sigma^2)$	$n\theta$	$n\sigma^2$
$\mathsf{N}(\mu,\theta)$	$\sum(Y_i-\mu)^2$	$\mathsf{G}(n/2,2\theta)$	$n\theta$	$2n\sigma^2$
$\mathsf{N}_k(\boldsymbol{\theta},\boldsymbol{\Sigma})$	$\sum \boldsymbol{Y}_i$	$\mathsf{N}_k(n\boldsymbol{\theta},n\boldsymbol{\Sigma})$	$n\boldsymbol{\theta}$	$n\boldsymbol{\Sigma}$
$\mathsf{N}_k(\boldsymbol{\mu},\theta\boldsymbol{\Sigma})$	$\sum(\boldsymbol{Y}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^-1(\boldsymbol{Y}_i-\boldsymbol{\mu})$	$\mathsf{G}(n/2,2\theta)$	$n\theta$	$2n\sigma^2$
$\mathsf{N}_k(\boldsymbol{\mu},\boldsymbol{\theta})$	$\sum(\boldsymbol{Y}_i-\boldsymbol{\mu})(\boldsymbol{Y}_i-\boldsymbol{\mu})^T$	$\mathsf{Wis}(k,n,\boldsymbol{\theta})$	$n\boldsymbol{\theta}$	

Table: Sufficient statistic $T(Y_1,\ldots,Y_n)$ (see p. 16 in the book) given a sample of n iid random variables $Y_1, Y_2,\ldots,Y_n$. Notice that in some cases the observation is a k dimensional random vector, and here a bold notation $\boldsymbol{Y}_i$ is used.

Conditional density of T given θ	Conjugate prior for θ	Posterior density for θ after the obs. $T = t(y_1, \dots, y_n)$	Marginal density of $T = t(Y_1, \dots, Y_n)$
$\mathsf{B}(n, \theta)$	$\mathsf{Beta}(\alpha, \beta)$	$\mathsf{Beta}(t + \alpha, n + \beta - t)$	$\mathsf{PI}(n, \alpha, \alpha + \beta)$
$\mathsf{NB}(n, \theta)$	$\mathsf{Beta}(\alpha, \beta)$	$\mathsf{Beta}(n + \alpha, \beta + t)$	$\mathsf{NPI}(n, \beta, \alpha + \beta)$
$\mathsf{P}(n\theta)$	$\mathsf{G}(\alpha, 1/\beta)$	$\mathsf{G}(t + \alpha, 1/(\beta + n)$	$\mathsf{NB}(\alpha, \beta/(\beta + n))$
$\mathsf{G}(n, \theta)$	$\mathsf{Inv\text{-}G}(\alpha, \beta)$	$\mathsf{Inv\text{-}G}(n + \alpha, \beta + t)$	$\mathsf{Inv\text{-}Beta}(\alpha, n, \beta)$
$\mathsf{Inv\text{-}Par}(\theta, n)$	$\mathsf{Par}(\beta, \mu)$	$\mathsf{Par}(\max(t, \beta), n + \mu)$	$\mathsf{BPar}\beta, \mu, n)$
$\mathsf{N}(n\theta, n\sigma^2)$	$\mathsf{N}(\mu, \sigma_0^2)$	$\mathsf{N}(\mu_1, \sigma_1^2)$ $\mu_1 = (\mu/\sigma_0^2 + t/\sigma^2)$ $1/\sigma_1^2 = 1/\sigma_0^2 + n/\sigma^2$	$\mathsf{N}(n\mu, n\sigma^2 + n^2\sigma_0^2)$
$\mathsf{N}_k(n\boldsymbol{\theta}, n\boldsymbol{\Sigma})$	$\mathsf{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma_0})$	$\mathsf{N}_k(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$ $\boldsymbol{\mu_1} = \boldsymbol{\Sigma_1}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu} + \boldsymbol{\Sigma}^{-1}\boldsymbol{t})$ $\boldsymbol{\Sigma_1}^{-1} = \boldsymbol{\Sigma_0}^{-1} + n\boldsymbol{\Sigma}^{-1}$	$\mathsf{N}_k(n\boldsymbol{\mu}, n\boldsymbol{\Sigma} + \boldsymbol{\Sigma_0})$

Table: Conditional densities of the statistic T given the parameter θ, conjugate prior densities for θ, posterior densities for θ after having observed the statistic $T = t(y_1, \dots, y_n)$, and the marginal densities for $T = t(Y_1, \dots, Y_n)$ – cf. also the discussion on page 16 and 17 in the book.(Notice that in some cases the observation is a random vector)

What else is out there

- Time series
- Multivariate analysis
- Non-parametric models
- Integrated analysis
- ...

But you are now well prepared to tackle those also.

Integrated analysis

- One nice thing about being able to write your own likelihood is flexibility
- Remember how we set up the log likelihood as the sum of the contributions from each independent observation:

$$\ell(\boldsymbol{\theta}, \boldsymbol{X}) = \ell(\boldsymbol{\theta}, x_1) + \ell(\boldsymbol{\theta}, x_2) + \cdots + \ell(\boldsymbol{\theta}, x_n)$$

- We did not say that our observations should come from the same distribution
- It is no problem to have some that are say normally distributed and others that are Poisson distributed inform us about the same model parameters
- That is only problematic when we are confined to a formula interface.