Research Organization of Information and Systems
The Institute of Statistical Mathematics

Effect Displays for Generalized Linear Models, Polytomous-Response Models, and Mixed-Effects Models

John Fox

McMaster University
Canada

November 2012
Tokyo, Japan

1. Topics

- Effect displays for generalized linear models
- Extension of effect displays to:
  - Multinomial logit model
  - Proportional-odds model
  - Generalized linear mixed-effects models
- Examples using the **effects** package in R
  - The **effects** package was written with Sanford Weisberg and Jangman Hong.

2. References

- J. Fox (1987), "Effect Displays for Generalized Linear Models,"*Sociological Methodology*, 17: 347–361.
- J. Fox (2003), "Effect Displays in R for Generalised Linear Models," *Journal of Statistical Software*, 8(15), 1–27.
- J. Fox and R. Andersen (2006), "Effect Displays for Multinomial and Proportional-Odds Logit Models," *Sociology Methodology*, 36: 225–255.
- J. Fox and J. Hong (2009), "Effect Displays in R for Multinomial and Proportional-Odds Logit Models: Extensions to the effects Package," *Journal of Statistical Software*, 32(1): 1–24.
- J. Fox and S.Weisberg (2011), *An R Companion to Applied Regression, Second Edition*, Sage Publications.
- J. Nelder (1977), "A Reformulation of Linear Models [with commentary]," *Journal of the Royal Statistical Society, A*, 140: 48–76.

3. Effect Displays for Generalized Linear Models

- Effect displays, in the sense of Fox (1987, 2003), are tabular or graphical summaries of statistical models.
  - They are typically simpler to interpret than the regression coefficients — often much simpler — for models with complex structure, such as interactions, polynomial regressors, or regression splines, and for models that are nonlinear on the scale of the response variable.
- The general idea underlying effect displays is to represent a statistical model by showing portions of its response surface, allowing some predictors to vary over their observed ranges while others are held constant at typical values.
- A general principle of interpretation for statistical models containing terms that are marginal to others (Nelder, 1977) is that high-order terms should be combined with their lower-order relatives.

- For example, an interaction between two factors should be combined with the main effects marginal to the interaction.
- Fox (1987) suggests identifying the high-order terms in a generalized linear model.
  - Fitted values under the model are computed for each such term.
  - The lower-order "relatives" of a high-order term (e.g., main effects marginal to an interaction, or a linear and quadratic term in a third-order polynomial) are absorbed into the term, allowing the predictors appearing in the term to range over their values.
  - The values of other predictors are fixed at typical values:
    - A covariate could be fixed at its mean or median.
    - A factor could be fixed at its proportional distribution in the data, or to equal proportions in its several levels.

- Some models have high-order terms that "overlap" — that is, that share a lower-order relative (other than the constant).
  - For example, a generalized linear model may include interactions AB, AC, and BC among the three factors A, B, and C.
  - Although the three-way interaction ABC is not in the model, it can be illuminating to combine the three high-order terms and their lower-order relatives (Fox, 2003).
- Consider a generalized linear model with linear predictor $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ and link function $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$, where $\boldsymbol{\mu}$ is the expectation of the response vector $\mathbf{y}$.
  - We have an estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, along with the estimated covariance matrix $\widehat{V}(\widehat{\boldsymbol{\beta}})$ of $\widehat{\boldsymbol{\beta}}$.

- Let the rows of $\mathbf{X}^*$ include all combinations of values of predictors appearing in a high-order term, along with typical values of the remaining predictors.
  - The column structure of $\mathbf{X}^*$ with respect, for example, to interactions, is the same as that of the model matrix $\mathbf{X}$.
- Then the fitted values $\widehat{\boldsymbol{\eta}}^* = \mathbf{X}^*\widehat{\boldsymbol{\beta}}$ represent the *effect* in question.
  - A table or graph of these values — or of the fitted values transformed to the scale of the response, $g^{-1}(\widehat{\boldsymbol{\eta}}^*)$ — is an *effect display*.
- The standard errors of $\widehat{\boldsymbol{\eta}}^*$ are the square-root diagonal entries of $\mathbf{X}^*\widehat{V}(\widehat{\boldsymbol{\beta}})\mathbf{X}^{*\prime}$.
  - These may be used to compute point-wise confidence intervals for the effects, the end-points of which may then also be transformed to the scale of the response.

- We prefer plotting on the scale of the linear predictor (where the structure of the model — e.g., linearity — is preserved) but labelling the response axis on the scale of the response.
  - This approach makes the display invariant with respect to the values at which the omitted predictors are held constant, in that only the labelling of the response axis changes with a different selection of these values.

3.1 Preliminary Example: A Binary Logit Model for Toronto Arrests for Marijuana Possession

- We construct effect displays for a binary logit model fit to data on police treatment of individuals arrested in Toronto for simple possession of small quantities of marijuana, where the police have the option of releasing an arrestee with a summons.*
  - The principal question of interest is whether and how the probability of release is influenced by the subject's sex, race ("color"), age, employment status, and citizenship, the year in which the arrest took place, and the subject's previous police record ("checks").
- Preliminary analysis of the data suggested a logit model including interactions between color and year and between color and age, and main effects of employment status, citizenship, and checks.

* I am grateful to Michael Friendly of York University, Toronto, for making these data available to me.

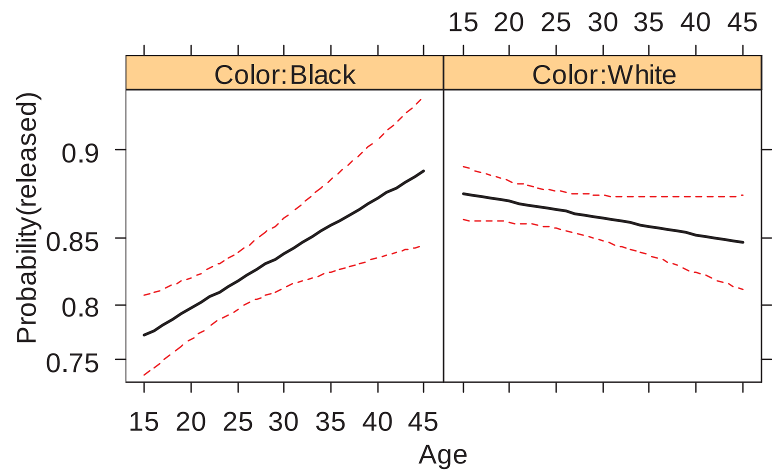
- Estimated coefficients and their standard errors:

Coefficient	*Estimate*	*Standard Error*
Constant	0.344	0.310
Employed (Yes)	0.735	0.085
Citizen (Yes)	0.586	0.114
Checks	−0.367	0.026
Color (White)	1.213	0.350
Year (1998)	−0.431	0.260
Year (1999)	−0.094	0.261
Year (2000)	−0.011	0.259
Year (2001)	0.243	0.263
Year (2002)	0.213	0.353

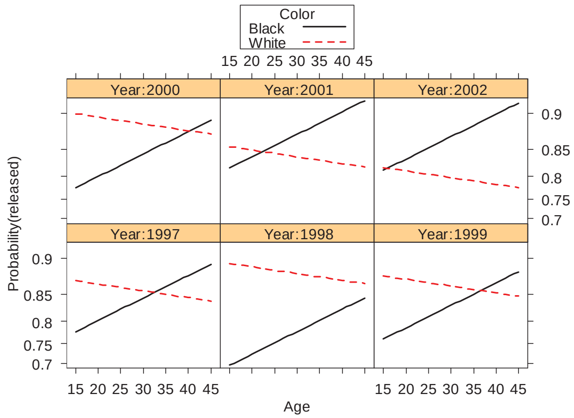
Coefficient	*Estimate*	*Standard Error*
Age	0.029	0.009
Color (White) × Year (1998)	0.652	0.313
Color (White) × Year (1999)	0.156	0.307
Color (White) × Year (2000)	0.296	0.306
Color (White) × Year (2001)	−0.381	0.304
Color (White) × Year (2002)	−0.617	0.419
Color (White) × Age	−0.037	0.010

  - It is difficult to tell from the coefficients how the predictors combine to influence the response.
- Two illustrative effect displays for the Toronto marijuana-arrests data:

Effect display for the color-by-age interaction:



Effect display combining the color-by-age and color-by-year interactions:



▶ For the effect display of the color-by-age interaction, $\mathbf{X}^*$ has the following structure:

(b_1)	(b_2)	(b_3)	(b_4)	(b_5)	(b_6)	(b_7)	(b_8)	(b_9)	
constant	employed	citizen	checks	color	1998	1999	2000	2001	
1	0.79	0.85	1.64	0	0.17	0.21	0.24	0.23	
1	0.79	0.85	1.64	1	0.17	0.21	0.24	0.23	
1	0.79	0.85	1.64	0	0.17	0.21	0.24	0.23	
1	0.79	0.85	1.64	1	0.17	0.21	0.24	0.23	
1	0.79	0.85	1.64	0	0.17	0.21	0.24	0.23	...
1	0.79	0.85	1.64	1	0.17	0.21	0.24	0.23	
1	0.79	0.85	1.64	0	0.17	0.21	0.24	0.23	
1	0.79	0.85	1.64	1	0.17	0.21	0.24	0.23	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
1	0.79	0.85	1.64	1	0.17	0.21	0.24	0.23	

	(b_{10})	(b_{11})	(b_{12})	(b_{13})	(b_{14})	(b_{15})	(b_{16})	(b_{17})
	2002	age	col × 98	col × 99	col × 00	col × 01	col × 02	col × age
	0.05	15	0	0	0	0	0	0
	0.05	15	0.17	0.21	0.24	0.23	0.05	15
	0.05	16	0	0	0	0	0	0
...	0.05	16	0.17	0.21	0.24	0.23	0.05	16
	0.05	17	0	0	0	0	0	0
	0.05	17	0.17	0.21	0.24	0.23	0.05	17
	0.05	18	0	0	0	0	0	0
	0.05	18	0.17	0.21	0.24	0.23	0.05	18
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	0.05	65	0.17	0.21	0.24	0.23	0.05	65

- Column 1 of $\mathbf{X}^*$ of 1s represents the constant regressor.
- Column 2 reflects the 79 percent of arrestees who were at level Yes of employed, and hence had values of 1 on the treatment-coded contrast for this factor.

  – 0.79 is therefore also the mean of the contrast. The display is, however, invariant with respect to contrast coding.
  – This column, along with other constant columns in $\mathbf{X}^*$, is in effect absorbed in the constant term, and therefore influences only the average level of the computed effects.
- Column 3 reflects the 85 percent of arrestees who were in level Yes of citizen.
- Column 4 reflects the average value of checks, 1.64.
- Column 5 repeats the two values 0 and 1 for the contrast for color (to be taken in combination with the values of age in column 11).

- Columns 6 through 10 represent the contrasts for year, and contain the proportions of arrestees in years 1998 through 2002; this reflects the use of the first level of year, 1997, as the baseline level.
- Column 11 contains the twice-repeated integer values of age, from 15 through 65. Because the age effect is linear on the logit scale, we really only need the two extreme values 15 and 65 — as long as we plot on the logit scale.
- Columns 12 through 16 are for the interaction of color with year (which is absorbed in the color term — i.e., these columns are constant within color).
- Column 17 is for the color by age interaction.

4. Extending Effect Displays

4.1 The Multinomial Logit Model

- Letting μ_{ij} denote the probability that observation i belongs to response category j of m categories, *the multinomial logit model* is

$$\mu_{ij} = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_j)}{\sum_{l=1}^{m}\exp(\mathbf{x}_l'\boldsymbol{\beta}_j)} \quad \text{for } j = 1, ..., m$$

  - where $\mathbf{x}_i' = (1, x_{i2}, \ldots, x_{ip})$ is the model vector for observation i;
  - and $\boldsymbol{\beta}_j = \left(\beta_1, \beta_2, \ldots, \beta_p\right)'$ is the parameter vector for response category j.
- The model is over-parametrized because $\sum_{j=1}^{m}\mu_{ij} = 1$.
  - To handle this feature of the model, we set $\boldsymbol{\beta}_m = \mathbf{0}$.

- Manipulating the model,

$$\log\frac{\mu_{ij}}{\mu_{im}} = \mathbf{x}_i'\boldsymbol{\beta}_j \quad \text{for } j = 1, ..., m-1$$

  - For any pair of categories:

$$\log\frac{\mu_{ij}}{\mu_{ij'}} = \mathbf{x}_i'(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}) \text{ for } j, j' \neq m$$

- But this does not produce intuitively easy-to-grasp coefficients, even for models in which the structure of the model vector $\mathbf{x}'$ is simple.
- Our strategy for building effect displays is essentially the same as for generalized linear models: Find fitted values — in this case, fitted probabilities — under the model for selected combinations of the predictors.
- Finding standard errors for fitted values on the probability scale is harder.
  - The fitted probabilities are nonlinear functions of the model parameters.
  - The linear predictor $\eta_{ij} = \mathbf{x}_i'\boldsymbol{\beta}_j$ is for the logit comparing category j to category m, not for the logit comparing category j to its complement, $\log\left[\mu_{ij}/(1-\mu_{ij})\right]$.
- We proceed by the delta method:
  - Suppose that we compute the fitted value at $\mathbf{x}_0'$.

– Differentiating μ_{0j} with respect to the model parameters:

$$\frac{\partial \mu_{0j}}{\partial \boldsymbol{\beta}_j} = \frac{\exp(\mathbf{x}_0'\boldsymbol{\beta}_j)\left[1+\sum_{j'=1,j'\neq j}^{m-1}\exp(\mathbf{x}_0'\boldsymbol{\beta}_{j'})\right]\mathbf{x}_0}{\left[1+\sum_{j'=1}^{m-1}\exp(\mathbf{x}_0'\boldsymbol{\beta}_{j'})\right]^2}$$

$$\frac{\partial \mu_{0j}}{\partial \boldsymbol{\beta}_{j'\neq j}} = -\frac{\left\{\exp\left[\mathbf{x}_0'\left(\boldsymbol{\beta}_{j'}+\boldsymbol{\beta}_j\right)\right]\right\}\mathbf{x}_0}{\left[1+\sum_{j'=1}^{m-1}\exp(\mathbf{x}_0'\boldsymbol{\beta}_{j'})\right]^2}$$

$$\frac{\partial \mu_{0m}}{\partial \boldsymbol{\beta}_j} = -\frac{\exp(\mathbf{x}_0'\boldsymbol{\beta}_j)\mathbf{x}_0}{\left[1+\sum_{j'=1}^{m-1}\exp(\mathbf{x}_0'\boldsymbol{\beta}_{j'})\right]^2}$$

- Let the estimated asymptotic covariance matrix of the (stacked) coefficient vectors be given by

$$\widehat{\mathcal{V}}(\widehat{\boldsymbol{\beta}}) = \widehat{\mathcal{V}}\begin{bmatrix}\widehat{\boldsymbol{\beta}}_1\\ \widehat{\boldsymbol{\beta}}_2\\ \vdots\\ \widehat{\boldsymbol{\beta}}_{m-1}\end{bmatrix} = [v_{st}]\,,\ s,t=1,\ldots,r$$

where $r = p(m-1)$ represents the total number of parameters in the combined parameter vectors.

- Then

$$\widehat{\mathcal{V}}(\widehat{\mu}_{0j}) \approx \sum_{s=1}^{r}\sum_{t=1}^{r} v_{st}\frac{\partial\widehat{\mu}_{0j}}{\partial\widehat{\boldsymbol{\beta}}_s}\frac{\partial\widehat{\mu}_{0j}}{\partial\widehat{\boldsymbol{\beta}}_t}$$

- Because the $\widehat{\mu}_{0j}$ are bounded by 0 and 1, we re-express the category probabilities μ_{0j} as logits,

$$\lambda_{0j} = \log\frac{\mu_{0j}}{1-\mu_{0j}}$$

- Differentiating with respect to μ_{0j}:

$$\frac{d\lambda_{0j}}{d\mu_{0j}} = \frac{1}{\mu_{0j}(1-\mu_{0j})}$$

- By a second application of the delta method,

$$\widehat{\mathcal{V}}(\widehat{\lambda}_{0j}) \approx \frac{1}{\widehat{\mu}_{0j}^2(1-\widehat{\mu}_{0j})^2}\widehat{\mathcal{V}}(\widehat{\mu}_{0j})$$

- Using this result, we can form a confidence interval around $\widehat{\mu}_{0j}$, and then translate the end-points back to the probability scale.

4.2 The Proportional-Odds Logit Model

▶ The *proportional-odds logit model* is a common model for an ordinal response variable.

- Suppose that there is a continuous, but unobservable, response variable, ξ, which is a linear function of a predictor vector $\mathbf{x}'$ plus a random error:

$$\begin{aligned}\xi_i &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i\\ &= \eta_i + \varepsilon_i\end{aligned}$$

- We cannot observe ξ directly, but instead implicitly dissect its range into m class intervals at the (unknown) *thresholds* $\alpha_1 < \alpha_2 < \cdots < \alpha_{m-1}$, producing the observed ordinal response variable y:

$$y_i = \begin{cases} 1 & \text{for } \xi_i \le \alpha_1\\ 2 & \text{for } \alpha_1 < \xi_i \le \alpha_2\\ \vdots & \\ m-1 & \text{for } \alpha_{m-2} < \xi_i \le \alpha_{m-1}\\ m & \text{for } \alpha_{m-1} < \xi_i\end{cases}$$

- The cumulative probability distribution of y_i is given by

$$\begin{aligned}\Pr(y_i \le j) &= \Pr(\xi_i \le \alpha_j)\\ &= \Pr(\eta_i + \varepsilon_i \le \alpha_j)\\ &= \Pr(\varepsilon_i \le \alpha_j - \eta_i)\end{aligned}$$

for $j = 1, 2, ..., m-1$.

- If the errors ε_i are independently distributed according to the standard logistic distribution, with distribution function

$$\Lambda(z) = \frac{1}{1+e^{-z}}$$

then we get the proportional-odds logit model:

$$\begin{aligned}\text{logit}[\Pr(y_i > j)] &= \log_e \frac{\Pr(y_i > j)}{\Pr(y_i \le j)}\\ &= -\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i\end{aligned}$$

for $j = 1, 2, ..., m-1$.

- This model is over-parametrized: Since the $\boldsymbol{\beta}$ vector typically includes a constant, say β_1, we have $m-1$ regression equations, the intercepts of which are expressed in terms of m parameters.
  - A solution is to eliminate the constant from $\boldsymbol{\beta}$ – i.e., setting $\beta_1 = 0$, which establishes the origin of the latent continuum ξ.
  - For convenience, we absorb the negative sign into the intercept:
  $$\text{logit}[\Pr(y_i > j)] = \alpha_j + \boldsymbol{\beta}'\mathbf{x}_i, \text{ for } j = 1, 2, ..., m-1$$
  - Then the thresholds are the negatives of the intercepts α_j.
  - When it adequately represents the data, the proportional-odds model (with $m + p - 2$ independent parameters) is more parsimonious than the multinomial logit model [with $p(m-1)$ independent parameters]. The proportional-odds model isn't, however, nested in the multinomial logit model.

▶ We consider two strategies for constructing effect displays for the proportional-odds model:

(a) Plot on the scale of the latent continuum, using the estimated thresholds, $-\widehat{\alpha}_j$, to show the division of the continuum into ordered categories.
  - A nice characteristic of the standard logistic distribution is that its quartiles are very close to ± 1, making the conditional distribution of the latent variable easy to interpret visually.

(b) Display fitted probabilities of category membership, as for the multinomial logit model.
  - Suppose that we need the fitted probabilities at $\mathbf{x}_0'$
  - Let $\eta_0 = \mathbf{x}_0'\boldsymbol{\beta}$, and let $\mu_{0j} = \Pr(Y_0 = j)$.

- Then

$$\begin{aligned}\mu_{01} &= \frac{1}{1+\exp(\alpha_1 + \boldsymbol{\eta}_0)}\\ \mu_{0j} &= \frac{\exp(\boldsymbol{\eta}_0)\,[\exp(\alpha_{j-1}) - \exp(\alpha_j)]}{[1+\exp(\alpha_{j-1}+\eta_0)]\,[1+\exp(\alpha_j+\eta_0)]},\ j = 2, \ldots, m-1\\ \mu_{0m} &= 1 - \sum_{j=1}^{m-1} \mu_{0j}\end{aligned}$$

– We derive approximate standard errors by the delta method:

$$\frac{\partial\mu_{01}}{\partial\alpha_1} = -\frac{\exp(\alpha_1+\boldsymbol{\eta}_0)}{[1+\exp(\alpha_1+\boldsymbol{\eta}_0)]^2}$$

$$\frac{\partial\mu_{01}}{\partial\alpha_j} = 0,\ j=2,\ldots,m-1$$

$$\frac{\partial\mu_{01}}{\partial\boldsymbol{\beta}} = -\frac{\exp(\alpha_1+\boldsymbol{\eta}_0)\mathbf{x}_0}{[\exp(\alpha_1+\boldsymbol{\eta}_0)]^2}$$

$$\frac{\partial\mu_{0j}}{\partial\alpha_{j-1}} = \frac{\exp(\alpha_{j-1}+\boldsymbol{\eta}_0)}{[1+\exp(\alpha_{j-1}+\boldsymbol{\eta}_0)]^2}$$

$$\frac{\partial\mu_{0j}}{\partial\alpha_j} = -\frac{\exp(\alpha_j+\boldsymbol{\eta}_0)}{[1+\exp(\alpha_j+\boldsymbol{\eta}_0)]^2}$$

$$\frac{\partial\mu_{0j}}{\partial\alpha_{j'}} = 0,\ j'\neq j, j-1$$

$$\frac{\partial\mu_{0j}}{\partial\boldsymbol{\beta}} = \frac{\exp(\boldsymbol{\eta}_0)\left[\exp(\alpha_j)-\exp(\alpha_{j-1})\right]\left[\exp(\alpha_{j-1}+\alpha_j+2\boldsymbol{\eta}_0)-1\right]\mathbf{x}_0}{[1+\exp(\alpha_{j-1}+\boldsymbol{\eta}_0)]^2\,[1+\exp(\alpha_j+\boldsymbol{\eta}_0)]^2}$$

$$\frac{\partial\mu_{0m}}{\partial\alpha_{m-1}} = \frac{\exp(\alpha_{m-1}+\boldsymbol{\eta}_0)}{[1+\exp(\alpha_{m-1}+\boldsymbol{\eta}_0)]^2}$$

$$\frac{\partial\mu_{0m}}{\partial\alpha_j} = 0,\ j=1,\ldots,m-2$$

$$\frac{\partial\mu_{0m}}{\partial\boldsymbol{\beta}} = \frac{\exp(\alpha_{m-1}+\boldsymbol{\eta}_0)\mathbf{x}_0}{[1+\exp(\alpha_{m-1}+\boldsymbol{\eta}_0)]^2}$$

– Stack up all of the parameters in the vector $\boldsymbol{\gamma} = (\alpha_1,\ldots,\alpha_{m-1},\boldsymbol{\beta}')'$, and let

$$\widehat{\mathcal{V}}(\widehat{\boldsymbol{\gamma}}) = [v_{st}],\ s,t=1,\ldots,r$$

where $r=m+p-2$.

– Then, as for the multinomial logit model,

$$\widehat{\mathcal{V}}(\widehat{\mu}_{0j}) \approx \sum_{s=1}^{r}\sum_{t=1}^{r} v_{st}\frac{\partial\widehat{\mu}_{0j}}{\partial\widehat{\gamma}_s}\frac{\partial\widehat{\mu}_{0j}}{\partial\widehat{\gamma}_t}$$

and

$$\widehat{\mathcal{V}}(\widehat{\lambda}_{0j}) \approx \frac{1}{\widehat{\mu}_{0j}^2(1-\widehat{\mu}_{0j})^2}\widehat{\mathcal{V}}(\widehat{\mu}_{0j})$$

where

$$\lambda_{0j} = \log\frac{\mu_{0j}}{1-\mu_{0j}}$$

are the individual-category logits.

4.3 Generalized Linear Mixed-Effects Models

- *Mixed-effects models* are models for dependent data, where, in the simplest cases, level-1 units (such as individuals) are *clustered* into level-2 units (such as institutions).
- Mixed-effects models are also appropriate for longitudinal data, in which level-2 units (say, individuals) are observed on multiple occasions.
- In both cases, it is generally unreasonable to assume that observatiions within a cluster are independent of one-another.
- The *linear mixed-effects model* takes the form

$$\begin{aligned}\mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i\\ \mathbf{b}_i &\sim \mathbf{N}_q(\mathbf{0},\boldsymbol{\Psi})\\ &\quad \mathbf{b}_i, \mathbf{b}_{i'} \text{ are independent for } i\neq i'\\ \boldsymbol{\varepsilon}_i &\sim \mathbf{N}_{n_i}(\mathbf{0},\sigma^2\boldsymbol{\Lambda}_i)\\ &\quad \boldsymbol{\varepsilon}_i, \boldsymbol{\varepsilon}_{i'} \text{ are independent for } i\neq i'\end{aligned}$$

where

- $\mathbf{y}_i$ is the $n_i \times 1$ response vector for observations in the ith cluster;
- $\mathbf{X}_i$ is the $n_i \times p$ model matrix for the fixed effects for observations in cluster i;
- $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed-effect coefficients, which are the same across clusters;
- $\mathbf{Z}_i$ is the $n_i \times q$ model matrix for the random effects for observations in cluster i;
- $\mathbf{b}_i$ is the $q \times 1$ vector of random-effect coefficients for cluster i; expressed as a deviation from the fixed effects;
- $\boldsymbol{\varepsilon}_i$ is the $n_i \times 1$ vector of errors for observations in cluster i;
- $\boldsymbol{\Psi}$ is the $q \times q$ covariance matrix for the random effects;
- $\sigma^2\boldsymbol{\Lambda}_i$ is the $n_i \times n_i$ covariance matrix for the errors in cluster i, and is $\sigma^2\mathbf{I}_{n_i}$ if the within-group errors are independent with constant variance.

▶ The *generalized linear mixed-effects model* (*GLMM*) is a straightforward extension of the generalized linear model, adding random effects to the linear predictor, and expressing the expected value of the response conditional on the random effects:

$$g(\mu_{ij}) = g[E(y_{ij}|b_{1i}, \ldots, b_{qi})] = \eta_{ij}$$
$$\eta_{ij} = \beta_1 + \beta_2 x_{2ij} + \cdots + \beta_p x_{pij} + b_{1i}z_{1ij} + \cdots + b_{qi}z_{qij}$$

- The link function $g()$ is as in generalized linear models.
- The conditional distribution of y_{ij} given the random effects is a member of an exponential family, or — for quasi-likelihood estimation — the variance of $y_{ij}|b_{1i}, \ldots, b_{qi}$ is a function of μ_{ij} and a dispersion parameter ϕ.
- We make the usual assumptions about the random effects: That they are multinormally distributed with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Psi}$.

▶ Adapting effect displays to the fixed effects in GLMMs is straightforward since in the process of fitting the model to data we obtain estimates of the fixed-effects parameters and their asymptotic covariance matrix.

5. Examples

5.1 A Multinomial Logit Model: Political Knowledge and Party Choice in Britain

▶ The data for this example are from the 2001 wave of the British Election Panel Study (BEPS).

- The response variable is party choice, with three categories: Labour, Conservative, and Liberal Democrat.

- Explanatory variables:
  - "Europe" is an 11-point scale that measures respondents' attitudes towards European integration; high scores represent "Eurosceptic" sentiment.
  - "Political knowledge" taps knowledge of party platforms on the European integration issue; the scale ranges from 0 (low knowledge) to 3 (high knowledge).
    - An analysis of deviance suggests that a linear specification for knowledge is acceptable.
  - The model also includes age, gender, perceptions of economic conditions over the past year (both national and household), and evaluations of the leaders of the three major parties.

▶ Estimated coefficients and their standard errors from a final multinomial logit model fit to the data:

	Labour/Liberal Democrat	
Coefficient	*Estimate*	*Standard Error*
Constant	-0.155	0.612
Age	-0.005	0.005
Gender (male)	0.021	0.144
Perceptions of Economy	0.377	0.091
Perceptions of Household Econ. Position	0.171	0.082
Evaluation of Blair (Labour leader)	0.546	0.071
Evaluation of Hague (Cons. leader)	-0.088	0.064
Evaluation of Kennedy (Lib. Dem. leader)	-0.416	0.072
Europe	-0.070	0.040
Political Knowledge	-0.502	0.155
Europe $\times$ Knowledge	0.024	0.021

	Cons./Liberal Democrat	
Coefficient	*Estimate*	*Standard Error*
Constant	0.718	0.734
Age	0.015	0.006
Gender (male)	-0.091	0.178
Perceptions of Economy	-0.145	0.110
Perceptions of Household Econ. Position	-0.008	0.101
Evaluation of Blair (Labour leader)	-0.278	0.079
Evaluation of Hague (Cons. leader)	0.781	0.079
Evaluation of Kennedy (Lib. Dem. leader)	-0.656	0.086
Europe	-0.068	0.049
Political Knowledge	-1.160	0.219
Europe $\times$ Knowledge	0.183	0.028

5.2 A Proportional-Odds Logit Model: Cross-National Differences in Attitudes Towards Government Efforts to Reduce Poverty

▶ Data for this example are taken from the World Values Survey of 1995-97, focusing on four countries: Australia, Norway, Sweden, and the United States.
- The response variable: "Do you think that what the government is doing for people in poverty in this country is about the right amount, too much, or too little?" — ordered: too little $<$ about right $<$ too much.
- Explanatory variables include gender, religion (coded 1 if the respondent belonged to a religion, 0 if the respondent did not), education (coded 1 if the respondent had a university degree, 0 if not), and country (dummy coded, with Sweden as the reference category).

- Preliminary analysis of the data suggested modeling the effect of age as a cubic polynomial (we use an orthogonal cubic polynomial) and including an interaction between age and country.

▶ The coefficients and their standard errors from a final model:

Coefficient	*Estimate*	*Standard Error*
Gender (male)	0.169	0.053
Religion (Yes)	0.168	0.078
University degree (Yes)	0.141	0.067
Age (linear)	10.659	5.404
Age (quadratic)	7.535	6.245
Age (cubic)	8.887	6.663
Norway	0.250	0.087
Australia	0.572	0.082
USA	1.176	0.087

Coefficient	*Estimate*	*Standard Error*
Norway × Age (linear)	−7.905	7.091
Australia × Age (linear)	9.267	6.313
USA × Age (linear)	10.871	6.647
Norway × Age (quadratic)	−0.623	8.028
Australia × Age (quadratic)	−17.719	7.035
USA × Age (quadratic)	−7.689	7.352
Norway × Age (cubic)	0.489	8.568
Australia × Age (cubic)	−2.761	7.385
USA × Age (cubic)	−11.164	7.587
Thresholds		
Too Little | About Right	0.785	0.109
About Right | Too Much	2.598	0.115

5.3 A Linear Mixed-Effects Model: Exercise and Eating Disorders

▶ This example is drawn from work by Blackmoor, Davis, and Fox on the exercise histories of 138 teenaged girls who were hospitalized for eating disorders and of 93 “control” subjects.

▶ There are several observations for each subject, but because the girls were hospitalized at different ages, the number of observations and the age at the last observation vary.

▶ The variables in the data set are:

- `subject`: an identification number, necessary to keep track of which observations belong to each subject.
- `age`: the subject’s age, in years, at the time of observation. All but the last observation for each subject were collected retrospectively at intervals of two years, starting at age eight. The age at the last observation is recorded to the nearest day.
- `exercise`: the amount of exercise in which the subject engaged, expressed as hours per week.
- `group`: a factor indicating whether the subject is a patient or a control.

▶ It is of interest here to determine the typical trajectory of exercise over time, and to establish whether this trajectory differs between eating-disordered and control subjects.

- Preliminary examination of the data suggests a log transformation of exercise.
- The model that I eventually fit to the data is of the form

$$\log_2\left(\text{exercise}_{ij} + \frac{5}{60}\right) = \beta_1 + \beta_2\left(\text{age}_{ij} - 8\right) + \beta_3 \text{ group}_i + \beta_4\left(\text{age}_{ij} - 8\right) \times \text{ group}_i + b_{1i} + \varepsilon_{ij}$$

for individual i measured on occasion j, where the level-1 errors, ε_{ij}, follow a continuous-time first-order autoregressive process, for which

$$\text{corr}(\varepsilon_{it}, \varepsilon_{i,t+s}) = \rho(s) = \phi^{|s|}$$

and s is not necessarily an integer.

  - group is a dummy variable, coded 1 for patients and 0 for controls.
  - This is a so-called random-intercept model.

- The estimates for this model are as follows:

Parameter	*Estimate*	*Standard Errors*
Fixed-Effect Parameters		
β_1	-0.307	0.189
β_2	0.073	0.032
β_3	-0.284	0.245
β_4	0.227	0.040
Random-Effect Parameters		
$\sqrt{\widehat{\psi}(b_1)}$	1.150	
$\widehat{\sigma}$	1.529	
$\widehat{\rho}$	0.631	