Effect Displays
for Complex Regression Models

John Fox

McMaster University
Canada

November 2009

FIOCRUZ, Brazil

1. Introduction

- Effect displays for generalized linear models:
  - Background and preliminary examples
- Extension of effect displays to
  - multinomial logit models
  - proportional-odds logit models

- References (including joint work with Robert Andersen and Jangman Hong):
  - Fox, J. (1987) Effect displays for generalized linear models. *Sociological Methodology* **17**, 347–361.
  - Fox, J. (2003) Effect displays in R for generalised linear models. *Journal of Statistical Software* **8**:15, 1–27
  - Fox, J. and R. Andersen (2006) Effect displays for multinomial and proportional-odds logit models. *Sociological Methodology* **36**, 225–255.
  - Fox, J. 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.
- The methods that I will describe are implemented in the **effects** package for R.

2. Effect displays for generalized linear models

- Effect displays, in the sense of Fox (1987, 2003), are tabular or graphical summaries of statistical models.
- The general idea underlying effect displays is to represent a statistical model by showing portions of its response surface

- 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 that includes 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 is 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 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.

- I 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.

2.1 A Binary Logit Model: Toronto Arrests for Marijuana Possession

- I will 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.

▶ Estimated coefficients and their standard errors:

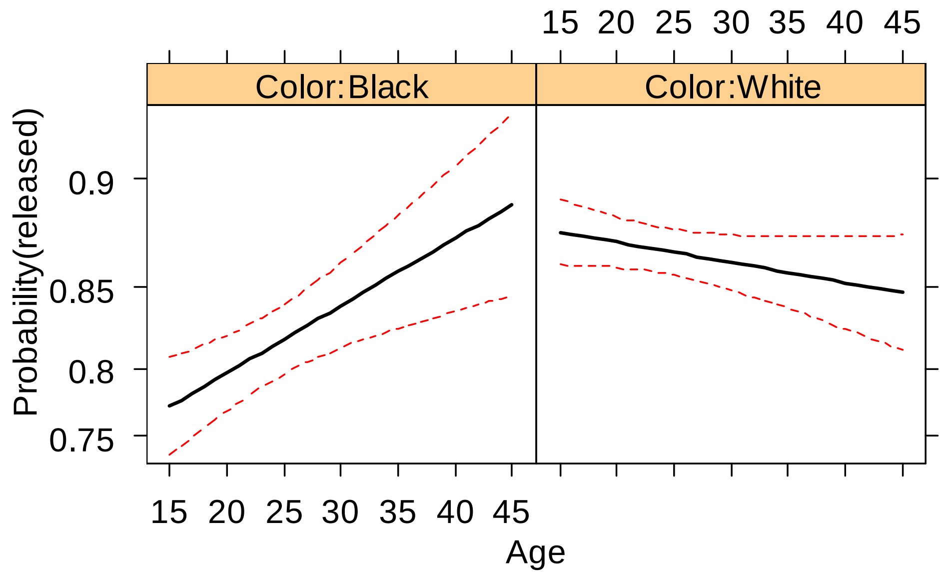
Coefficient	*Estimate*	*Standard Error*
Constant	0.344	0.310
Employed (Yes)	0.735	0.085
Citizen (Yes)	0.586	0.114
Checks	-0.367	0.026
Color (White)	1.213	0.350
Year (1998)	-0.431	0.260
Year (1999)	-0.094	0.261
Year (2000)	-0.011	0.259
Year (2001)	0.243	0.263
Year (2002)	0.213	0.353

Coefficient	*Estimate*	*Standard Error*
Age	0.029	0.009
Color (White) $\times$ Year (1998)	0.652	0.313
Color (White) $\times$ Year (1999)	0.156	0.307
Color (White) $\times$ Year (2000)	0.296	0.306
Color (White) $\times$ Year (2001)	-0.381	0.304
Color (White) $\times$ Year (2002)	-0.617	0.419
Color (White) $\times$ 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 × Age interaction:



- In this case, $\mathbf{X}^*$ has the following structure:

$$
\begin{array}{c}
\begin{array}{ccccccccc}
(b_1) & (b_2) & (b_3) & (b_4) & (b_5) & (b_6) & (b_7) & (b_8) & (b_9) \\
\text{constant} & \text{employed} & \text{citizen} & \text{checks} & \text{colour} & \text{1998} & \text{1999} & \text{2000} & \text{2001}
\end{array} \\
\begin{bmatrix}
1 & 0.79 & 0.85 & 1.64 & 0 & 0.17 & 0.21 & 0.24 & 0.23 & \\
1 & 0.79 & 0.85 & 1.64 & 1 & 0.17 & 0.21 & 0.24 & 0.23 & \\
1 & 0.79 & 0.85 & 1.64 & 0 & 0.17 & 0.21 & 0.24 & 0.23 & \\
1 & 0.79 & 0.85 & 1.64 & 1 & 0.17 & 0.21 & 0.24 & 0.23 & \\
1 & 0.79 & 0.85 & 1.64 & 0 & 0.17 & 0.21 & 0.24 & 0.23 & \cdots \\
1 & 0.79 & 0.85 & 1.64 & 1 & 0.17 & 0.21 & 0.24 & 0.23 & \\
1 & 0.79 & 0.85 & 1.64 & 0 & 0.17 & 0.21 & 0.24 & 0.23 & \\
1 & 0.79 & 0.85 & 1.64 & 1 & 0.17 & 0.21 & 0.24 & 0.23 & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
1 & 0.79 & 0.85 & 1.64 & 1 & 0.17 & 0.21 & 0.24 & 0.23 &
\end{bmatrix}
\end{array}
$$

$$
\begin{array}{cccccccc}
(b_{10}) & (b_{11}) & (b_{12}) & (b_{13}) & (b_{14}) & (b_{15}) & (b_{16}) & (b_{17}) \\
\text{2002} & \text{age} & \text{col} \times 98 & \text{col} \times 99 & \text{col} \times 00 & \text{col} \times 01 & \text{col} \times 02 & \text{col} \times \text{age}
\end{array}
$$

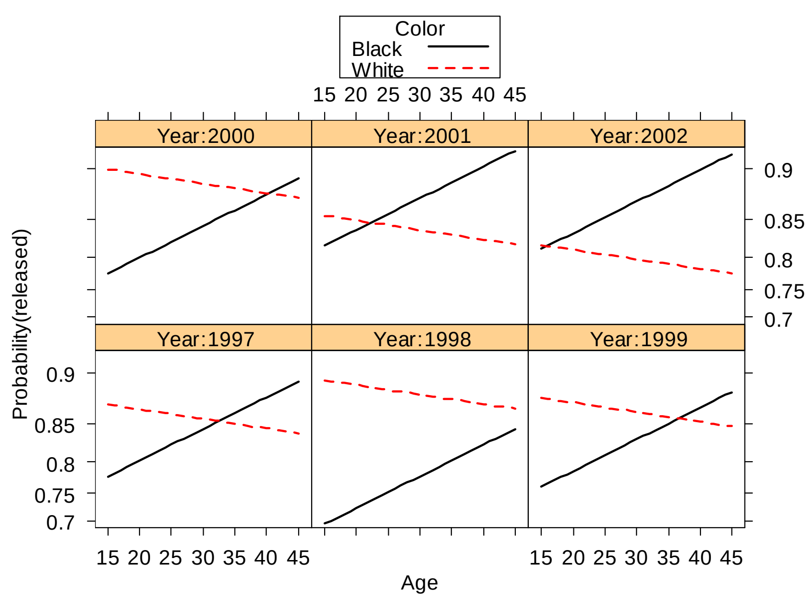
$$
\begin{bmatrix}
 & 0.05 & 15 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 0.05 & 15 & 0.17 & 0.21 & 0.24 & 0.23 & 0.05 & 15 \\
 & 0.05 & 16 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 0.05 & 16 & 0.17 & 0.21 & 0.24 & 0.23 & 0.05 & 16 \\
\cdots & 0.05 & 17 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 0.05 & 17 & 0.17 & 0.21 & 0.24 & 0.23 & 0.05 & 17 \\
 & 0.05 & 18 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & 0.05 & 18 & 0.17 & 0.21 & 0.24 & 0.23 & 0.05 & 18 \\
 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 & 0.05 & 65 & 0.17 & 0.21 & 0.24 & 0.23 & 0.05 & 65
\end{bmatrix}
$$

- Column 1 of $\mathbf{X}^*$ represents the constant.
- 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.
  - 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 `colour` (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.
- Columns 12 through 16 are for the interaction of `colour` with `year` (which is absorbed in the `colour` term).
- Column 17 is for the `colour` by `age` interaction.

▶ Combining the Colour × Age and Colour × Year interactions:



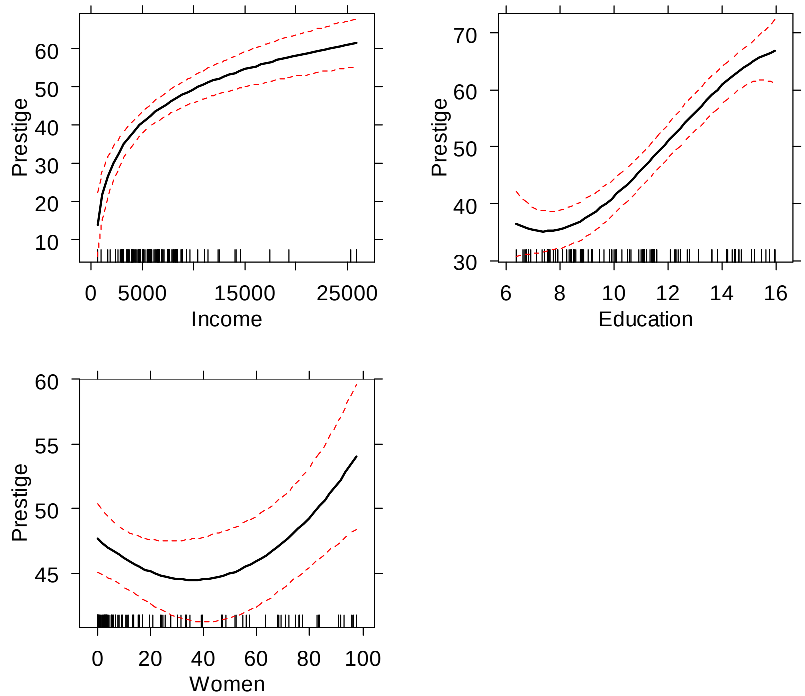
2.2 A Linear Model: Canadian Occupational Prestige

- The data for our second example pertain to the rated prestige of 102 Canadian occupations, regressed on three predictors from the 1971 Census of Canada
  (a) the average income of occupational incumbents, in dollars (represented in the model as the log of income)
  (b) the average education of occupational incumbents, in years (represented by a B-spline with three degrees of freedom)
  (c) the percentage of occupational incumbents who were women (represented by an orthogonal polynomial of degree two).

- Estimated coefficients and standard errors:

Coefficient	*Estimate*	*Standard Error*
Constant	-72.92	15.49
log Income	12.67	1.84
Education (1)	-8.20	7.80
Education (2)	25.66	5.50
Education (3)	30.42	4.59
Women (linear)	11.98	9.38
Women (quadratic)	18.47	6.83

  - This model does a decent job of summarizing the data, but the meaning of its coefficients is relatively obscure — despite the fact that the model includes no interactions.
- Effect displays for the terms in the model (with 95-percent point-wise confidence bands):



3. 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

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

  - where $\mathbf{x}_i' = (1, x_{i2}, \ldots, x_{ip})$ is the model vector for observation i;
  - and $\boldsymbol{\beta}_j = (\beta_1, \beta_2, \ldots, \beta_p)'$ 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, $\log\left[\mu_{ij}/(1-\mu_{ij})\right]$.
  - Fox and Andersen (2006) get approximate standard errors by the delta method.

3.1 Example: 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.

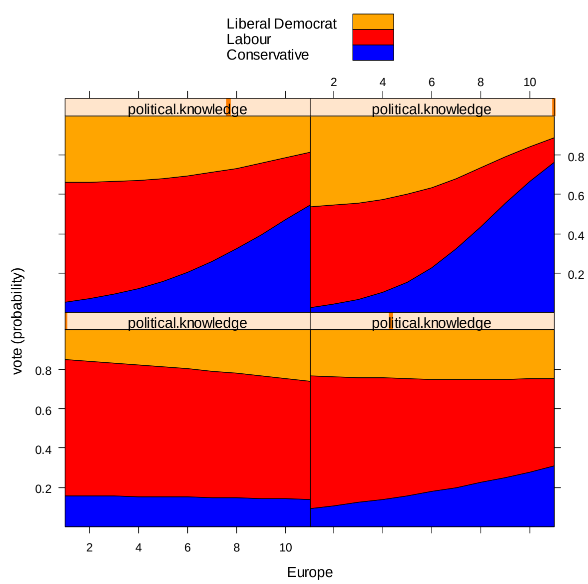
- Explanatory variables:
  - "Europe" is an 11-point scale that measures respondents' attitudes towards European integration; high scores represent Eurosceptic sentiment.
  - "Political knowledge" taps knowledge of party platforms on the European integration issue; the scale ranges from 0 (low knowledge) 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.

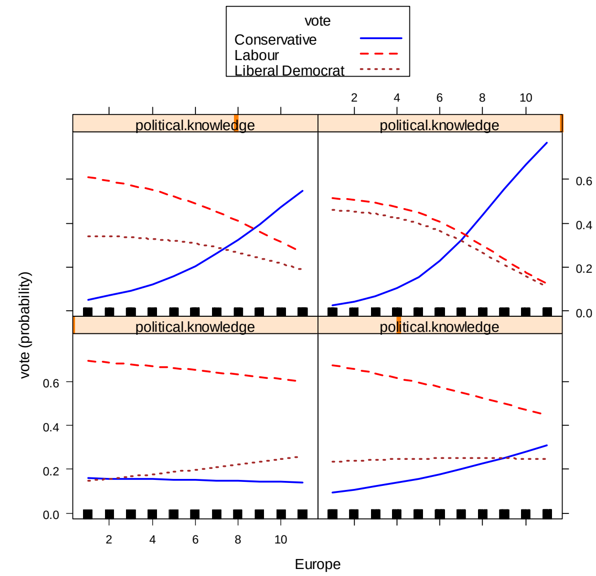
► Estimated coefficients and their standard errors from a final multinomial logit model fit to the data:

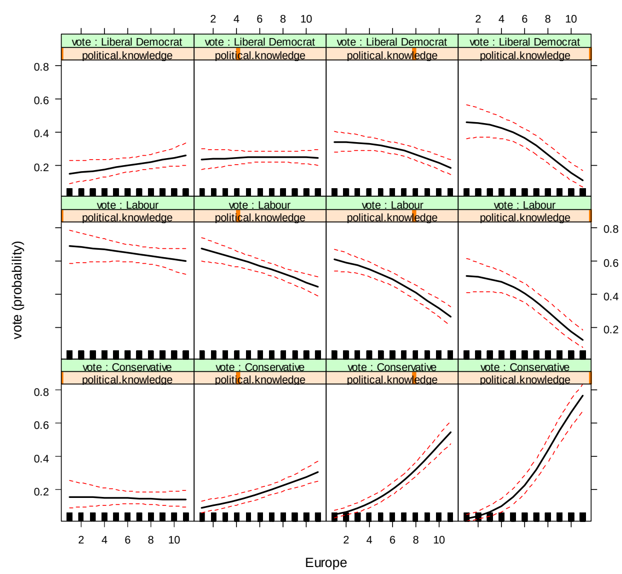
Coefficient	*Labour/Liberal Democrat* *Estimate*	*Standard Error*
Constant	-0.155	0.612
Age	-0.005	0.005
Gender (male)	0.021	0.144
Perceptions of Economy	0.377	0.091
Perceptions of Household Econ. Position	0.171	0.082
Evaluation of Blair (Labour leader)	0.546	0.071
Evaluation of Hague (Cons. leader)	-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

Coefficient	*Cons./Liberal Democrat* *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

- Several different styles of effect displays for the interaction between attitude towards Europe and political knowledge:
  - Plotted fitted probabilities as 'stacked areas.'
  - Plotting fitted probabilities as lines.
  - Showing confidence bands around the fitted effects.




vote
Conservative
Labour
Liberal Democrat
political.knowledge
vote (probability)
Europe
0.0
0.2
0.4
0.6
2
4
6
8
10


vote : Liberal Democrat
vote : Labour
vote : Conservative
political.knowledge
vote (probability)
Europe
0.2
0.4
0.6
0.8
2
4
6
8
10

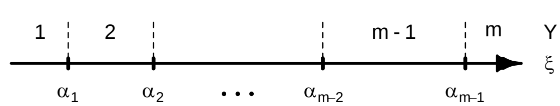
4. The Proportional-Odds Logit Model

- 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:

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

- 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:

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

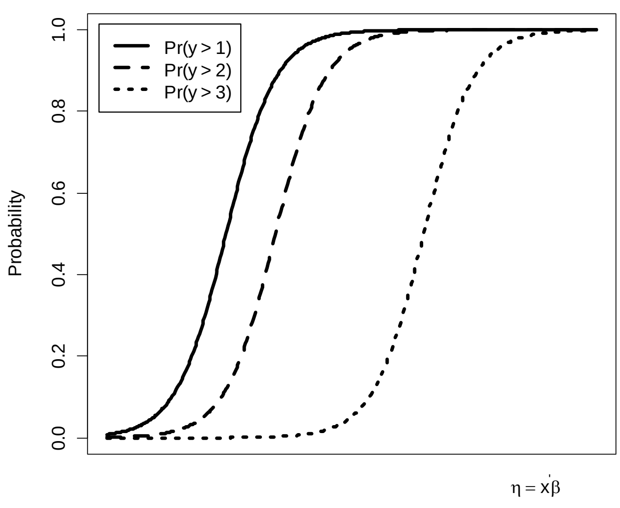


- The cumulative probability distribution of y_i is given by

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

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

▶ The regression surfaces for the proportional-odds model are parallel horizontally:



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

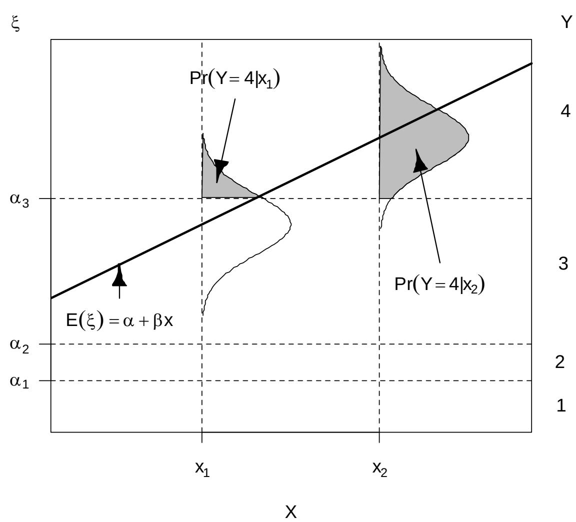
$$\Lambda(z) = \frac{1}{1 + e^{-z}}$$

then we get the proportional-odds logit model:

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

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

▶ The proportional-odds model:



▶ 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 or 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

$$\mu_{01} = \frac{1}{1+\exp(\alpha_1 + \boldsymbol{\eta}_0)}$$

$$\mu_{0j} = \frac{\exp(\boldsymbol{\eta}_0)\left[\exp(\alpha_{j-1}) - \exp(\alpha_j)\right]}{\left[1+\exp(\alpha_{j-1}+\eta_0)\right]\left[1+\exp(\alpha_j+\eta_0)\right]},\ j = 2,\ldots,m-1$$

$$\mu_{0m} = 1 - \sum_{j=1}^{m-1} \mu_{0j}$$

· As for the multinomial logit model, we can get approximate standard errors by the delta method.

4.1 Example: 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).

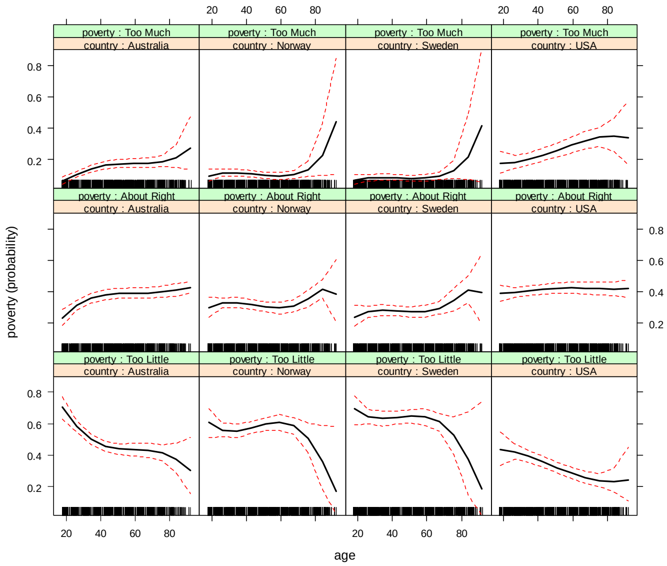
- Preliminary analysis of the data suggested modeling the effect of age as a cubic polynomial (we use an orthogonal cubic polynomial) and including an interaction between age and country.

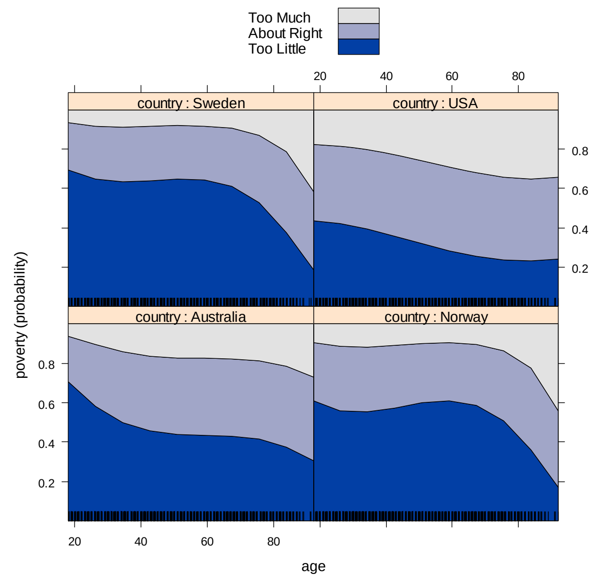
▶ The coefficients and their standard errors from a final model:

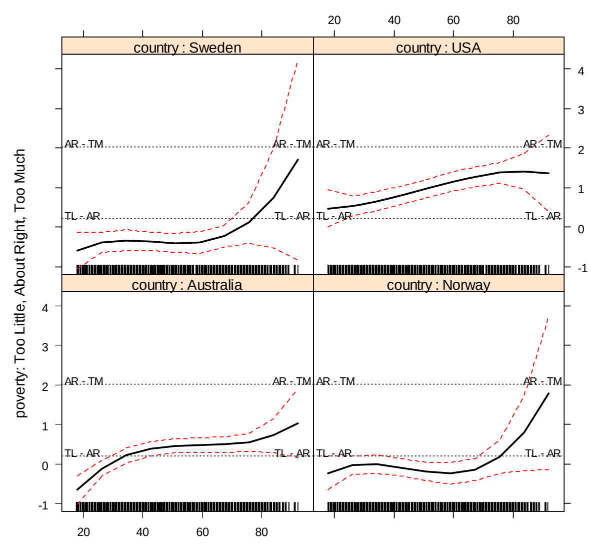
Coefficient	*Estimate*	*Standard Error*
Gender (male)	0.169	0.053
Religion (Yes)	-0.168	0.078
University degree (Yes)	0.141	0.067
Age (linear)	10.659	5.404
Age (quadratic)	7.535	6.245
Age (cubic)	8.887	6.663
Norway	0.250	0.087
Australia	0.572	0.823
USA	1.176	0.087

Coefficient	*Estimate*	*Standard Error*
Norway $\times$ Age (linear)	-7.905	7.091
Australia $\times$ Age (linear)	9.264	6.312
USA $\times$ Age (linear)	10.868	6.647
Norway $\times$ Age (quadratic)	-0.625	8.027
Australia $\times$ Age (quadratic)	-17.716	7.034
USA $\times$ Age (quadratic)	-7.692	7.352
Norway $\times$ Age (cubic)	0.485	8.568
Australia $\times$ Age (cubic)	-2.762	7.385
USA $\times$ Age (cubic)	-11.163	7.587
Thresholds		
Too Little $\mid$ About Right	0.449	0.106
About Right $\mid$ Too Much	2.262	0.111

- Alternative effect displays:
  - Plotting fitted category-membership probabilities (with and without 95-percent confidence bands).
  - Plotting fitted values on the scale of the latent response continuum (with thresholds between categories of the observed response).




Too Much
About Right
Too Little
country : Sweden
country : USA
country : Australia
country : Norway
poverty (probability)
age


country : Sweden
country : USA
country : Australia
country : Norway
AR - TM
TL - AR
poverty: Too Little, About Right, Too Much
age