

Effect Displays for Multinomial and Proportional-Odds Logit Models

John Fox and Robert Andersen
Department of Sociology
McMaster University
1280 Main Street West
Hamilton, Ontario
Canada L8S 4M4

Paper to be read at the ASA Methodology Conference 2004

16 April 2004

Abstract

An “effect display” is a graphical or tabular summary of a statistical model based on high-order terms in the model. Effect displays have previously been defined by Fox (1987, 2003) for generalized linear models (including linear model). After reviewing and illustrating effect displays for generalized linear models, we extend the displays to two models commonly used for polytomous categorical data: the multinomial logit model and the proportional-odds logit model. While most details of effect displays for these two models are straightforward, the derivation of standard errors is more challenging. We provide formulas for standard errors and develop examples.

1 Introduction

Effect displays, in the sense of Fox (1987, 2003), are tabular or — more commonly — graphical summaries of statistical models. Fox (1987) introduces effects displays for generalized linear models (including linear models); Fox (2003) refines these methods and provides software for their essentially automatic implementation.

The general idea underlying effect displays — to represent a statistical model by showing carefully selected portions of its response surface — is not limited to generalized linear models, however, nor even to models that incorporate linear predictors. Moreover, the essential idea of effect displays is not wholly original with Fox (1987). For example, adjusted means in analysis of covariance (introduced by Fisher, 1936) are a precursor to more general effect displays; see Fox (1987, 2003) for additional discussion, comparison, and references.

The primary purpose of this paper is to extend effect displays to the multinomial logit model and to the proportional-odds logit model, statistical models that find common application in social research. As we will show, this extension is largely straightforward, although the derivation of standard errors is challenging, particularly in the proportional-odds model. We begin by reviewing effect displays for generalized linear models, using as examples a binary logit model and a linear model. We then present results for the multinomial and proportional-odds logit models. Finally, we develop examples of effect displays for multinomial and proportional-odds logit models.

2 Effect Displays for Generalized Linear Models: Background and Preliminary Examples

A general principle of interpretation for statistical models containing terms that are marginal to others (in the sense of 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. In conformity with this principle, 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, which are marginal to the cubic term) are absorbed into the term, allowing the predictors appearing in the high-order term to range over their values. The values of other predictors are fixed at typical values: for example, a covariate could be fixed at its mean or median, a factor 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). Consider, 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 nevertheless illuminating to combine the three high-order terms and their lower-order relatives (see Fox, 2003, and the example developed in Section 2.1).

Let us turn now to the generalized linear model (e.g., McCullagh and Nelder, 1998, or Firth, 1991) with linear predictor $\eta = \mathbf{X}\beta$ and link function $g(\mu) = \eta$, where μ is the expectation of the response vector \mathbf{y} . Here, everything falls into place very simply: We have an estimate $\hat{\beta}$ of β , along with the estimated covariance matrix $\widehat{V}(\hat{\beta})$ of $\hat{\beta}$.

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 $\hat{\eta}^* = \mathbf{X}^*\hat{\beta}$ represent the effect in question, and a table or graph of these values — or, alternatively, of the fitted values transformed to the scale of the response, $g^{-1}(\hat{\eta}^*)$ — is an effect display. The standard errors of $\hat{\eta}^*$, available as the square-root diagonal entries of $\mathbf{X}^*\widehat{V}(\hat{\beta})\mathbf{X}^{*'}$, 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.

In an application, as we will illustrate presently, we prefer plotting on the scale of the linear predictor (where the structure of the model — for example, with respect to linearity — is preserved) but labelling the response axis on the scale of the response. This approach has the advantage of making the configuration of the display invariant with respect to the values at which the omitted predictors are held constant, in the sense that only the labelling of the response axis changes with a different selection of these values.¹

2.1 A Binary Logit Model: Toronto Arrests for Marijuana Possession

Following Fox (2003), 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. (The data discussed here are part of a larger data set featured in a series of articles in the *Toronto Star* newspaper.) Under these circumstances police have the option of releasing an arrestee with

¹As David Firth has pointed out to us, however, this invariance does not hold with respect to standard errors, which *are* affected by the fixed elements of \mathbf{X}^* , a fact that follows from considering effects as fitted values. Standard errors will tend to be smaller for components of \mathbf{x}' near the center of the data.

a summons to appear in court — similar to a traffic ticket; alternatively, the individual may be brought to the police station for questioning and possible indictment. The principal question of interest is whether and how the probability of release is influenced by the subject’s sex, race, age, employment status, and citizenship, the year in which the arrest took place, and the subject’s previous police record. Most of these variables are self-explanatory, with the following exceptions:

- Race appears in the model as “color,” and is coded as either “black” or “white.” The original data included the additional categories “brown” and “other,” but their meaning is ambiguous and their use relatively infrequent. Moreover, the motivation for collecting the data was to determine whether blacks and whites are treated differently by the police.
- The observations span the years 1997 through (part of) 2002. A few arrests in 1996 were eliminated. In the analysis reported below, year is treated as a factor (i.e., as a categorical predictor).
- When suspects are stopped by the police, their names are checked in six data bases — of previous arrests, previous convictions, parole status, and so on. The variable “checks” records the number of data bases on which an individual’s name appeared.

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. The effects of age and checks appear to be reasonably linear on the logit scale and are modelled as such.

Estimated coefficients and their standard errors are shown in Table 1. Where predictors are represented by dummy regressors, the category coded one is given in parentheses; for year, the baseline category is 1997. A fundamental point to be made with respect to Table 1 is that it is difficult to tell from the coefficients alone how the predictors combine to influence the response. This difficulty is primarily a function of the complex structure of the model — that is, the interactions of color with year and age — but partly due to the fact that the coefficients are effects on the logit scale.² It is true that with some mental arithmetic we can draw certain conclusions from the table of coefficients. For example, the fitted probability of release declines with age for whites but increases with age for blacks. Grasping the color-by-year interaction is more difficult, however, as is discerning the combined effect of these three predictors.

Two effect displays for the model fit to the Toronto marijuana-arrests data appear in Figures 1 and 2: Figure 1 depicts the interaction between color and age. Note that the lines in this graph are plotted on the logit scale (i.e., the scale of the linear predictor), but the vertical axis of the graph is labelled on the probability scale (the scale of the response); the broken lines give point-wise 95-percent confidence envelopes around the fitted values. Figure 2 combines the color-by-age interaction with the color-by-year interaction. Because there is no three-way interaction (and no interaction between age and year), the lines for blacks are parallel across the six panels of the graph, as are the lines for whites. We believe that a graphical representation such as Figure 2 effectively communicates what the model has to say about how color, age, and year combine to influence the probability of release.

²A common device, which speaks partly to the second problem but not the first, is to exponentiate the coefficients in the logit model. The exponentiated coefficients are interpretable as multiplicative effects on the relative odds of the response.

<i>Coefficient</i>	<i>Estimate</i>	<i>Standard Error</i>
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
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

Table 1: Maximum-likelihood estimates and standard errors for coefficients in the logit model for the Toronto marijuana-arrests data.

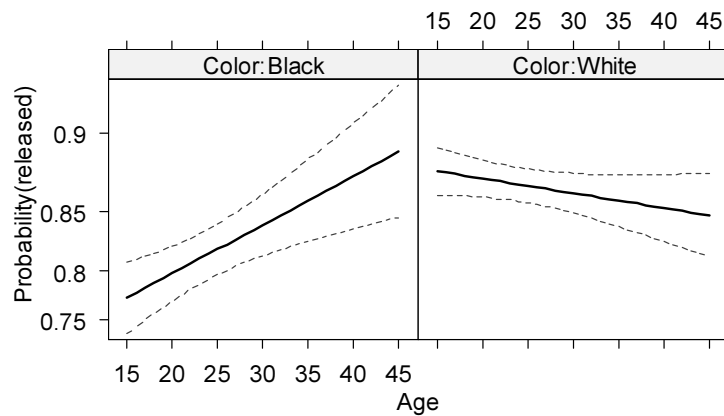


Figure 1: Effect display for the interaction of color and age in the logit model fit to the Toronto marijuana-arrests data. The vertical axis is labelled on the probability scale, and a 95-percent point-wise confidence envelope is drawn around the estimated effect. This graph, and those in Figures 2 and 3, are produced by the software described in Fox (2003).

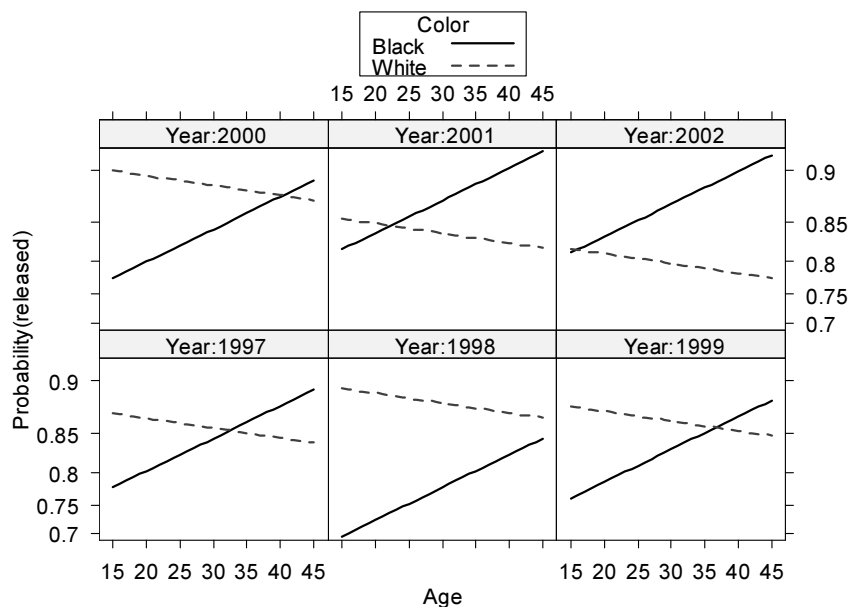


Figure 2: An effect display that combines the color-by-year and color-by-age interactions.

2.2 A Linear Model: Canadian Occupational Prestige

The data for our second example, also adapted from Fox (2003), pertain to the rated prestige of 102 Canadian occupations. The prestige of the occupations is regressed on three predictors, all derived from the 1971 Census of Canada: the average income of occupational incumbents, in dollars (represented in the model as the log of income); the average education of occupational incumbents, in years (represented by a B-spline with three degrees of freedom); and the percentage of occupational incumbents who were women (represented by an orthogonal polynomial of degree two). Estimated coefficients and standard errors for this model are shown in Table 2.

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. The coefficient of log income, for example, would be more easily interpreted had we used logs to the base two rather than

<i>Coefficient</i>	<i>Estimate</i>	<i>Standard Error</i>
Constant	-72.92	15.49
log Income	12.67	1.84
Education (1)	-8.20	7.8
Education (2)	25.66	5.50
Education (3)	30.42	4.59
Women (linear)	11.98	9.38
Women (quadratic)	18.47	6.83

Table 2: Coefficients for the regression of occupational prestige on the income and education levels of the occupations and on the percentage of occupational incumbents who are women. Education is represented in the model by a three degree-of-freedom B-spline, education by a second-order orthogonal polynomial.

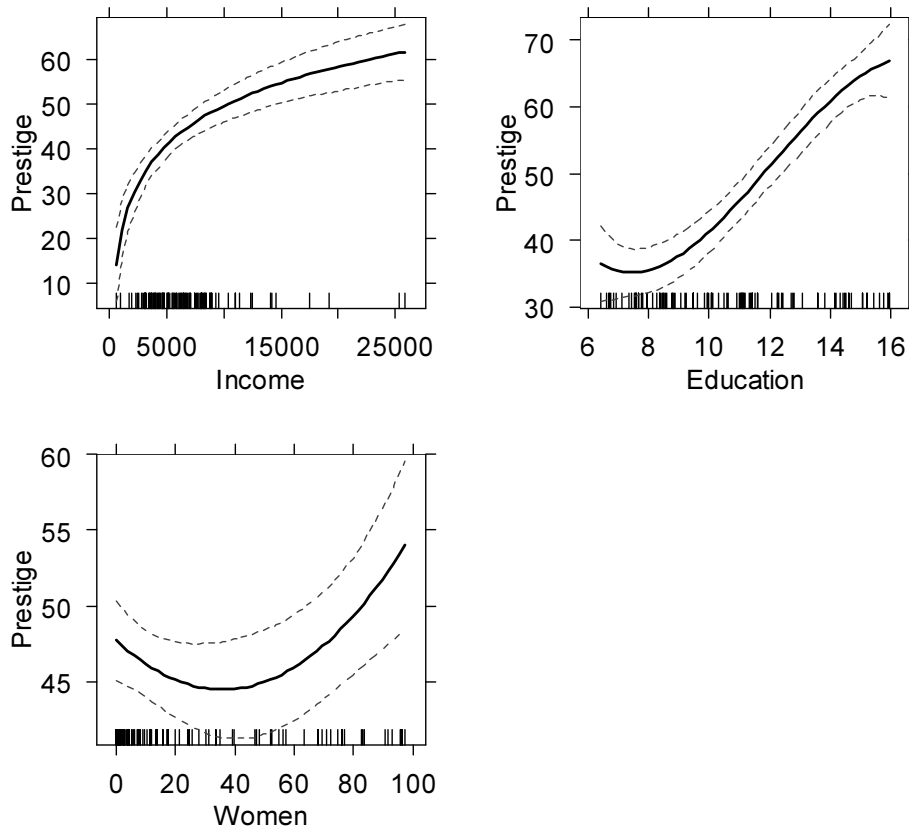


Figure 3: Effect plots for the predictors of prestige in the Canadian occupational prestige data. The model includes the log of income, a three-degree-of-freedom B-spline in education, and a quadratic in the percentage of occupational incumbents who are women. The “rug plot” (one-dimensional scatterplot) at the bottom of each graph shows the distribution of the corresponding predictor.

natural logs. The coefficients corresponding to the different elements of the B-spline basis do not have straightforward individual interpretations. Finally, although we can see from the coefficients for the orthogonal polynomial fit to the percentage of women that the linear trend in this predictor is non-significant while the quadratic trend is highly significant, these two coefficients are best interpreted in combination. Effect displays for the three predictors in the model appear in Figure 3. Note that here we prefer to plot income on the natural scale rather than using a log horizontal axis.

3 Effect Displays for the Multinomial Logit and Proportional-Odds Logit Models: Basic Results

3.1 The Multinomial Logit Model

The multinomial logit model is arguably the most widely used statistical model for polytomous (multi-category) response variables (e.g., Powers and Xie, 2000: Chapter 7; Fox, 1997: Chapter 15). Letting μ_{ij} denote the probability that observation i belongs to response category j of m

categories, the model is given by

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

where $\mathbf{x}'_i = (1, x_{i2}, \dots, x_{ip})$ is the model vector for observation i and $\beta_j = (\beta_1, \beta_2, \dots, \beta_p)'$ is the parameter vector for response category j . Observations may represent individuals, who therefore fall into a particular category of the response, or a vector of category counts for a multinomial observation (as in a contingency table, where both the predictors and the explanatory variables are discrete); the first case is a special case of the second, setting all of the multinomial total counts (i.e., the “multinomial denominators”) n_i to 1.

As it stands, model 1 is over-parametrized, because of the constraint that the probabilities for each observation sum to one: $\sum_{j=1}^m \mu_{ij} = 1$. The resulting indeterminacy can be handled by a normalization, placing a linear constraint on the parameters, $\sum_{j=1}^m a_j \beta_j = \mathbf{0}$, where the a_j are constants, not all zero. There is an important sense in which the choice of constraint is inessential: Fitted probabilities, $\hat{\mu}_{ij}$, and hence the likelihood, under the model are unaffected by the constraint. The meaning of specific parameters depends upon the constraint, however, and as we will explain, adds to the difficulty of directly interpreting coefficient estimates for the model. The most common constraint is to set one of the β_j to zero (i.e., to set one of the a_j to 1 and the rest to 0); for convenience, we will set $\beta_m = \mathbf{0}$, allowing us to rewrite equation 1 as

$$\begin{aligned} \mu_{ij} &= \frac{\exp(\mathbf{x}'_i \beta_j)}{1 + \sum_{l=1}^{m-1} \exp(\mathbf{x}'_l \beta_j)} \quad \text{for } j = 1, \dots, m-1 \\ \mu_{im} &= 1 - \sum_{l=1}^{m-1} \mu_{il} \quad \text{(for category } m) \end{aligned} \quad (2)$$

Algebraic manipulation of model 2 suggests an interpretation of the coefficients of the model:

$$\log \frac{\mu_{ij}}{\mu_{im}} = \mathbf{x}'_i \beta_j \quad \text{for } j = 1, \dots, m-1$$

and thus the coefficient vector β_j is for the relative log-odds of membership in category j versus the “baseline” category m . We can, moreover, express the relative log-odds of membership in *any* pair of categories in terms of *differences* in their coefficient vectors:

$$\log \frac{\mu_{ij}}{\mu_{ij'}} = \mathbf{x}'_i (\beta_j - \beta_{j'}) \quad \text{for } j, j' \neq m$$

All this is well and good, but it does not produce intuitively easy-to-grasp coefficients, since pairwise comparison of the categories of the response is not in itself a natural manner in which to think about a polytomous variable. This difficulty of interpretation pertains even to models in which the structure of the model vector \mathbf{x}' is simple.

Our strategy for building effect displays for the multinomial logit model 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. The fitted values on the probability scale, $\hat{\mu}_{ij}$, are given by model 2, substituting estimates $\hat{\beta}_j$ for the parameter vectors β_j .

Finding standard errors for fitted values on the probability scale is more of a challenge. As is obvious from model 2, the fitted probabilities are nonlinear functions of the model parameters.

We did not encounter this difficulty in the binary logit model because we could work on the scale of the linear predictor, translating the end-points of confidence intervals to the probability scale (or equivalently, relabelling the logit axis). In the multinomial logit model, however, as noted, the linear predictor $\eta_{ij} = \mathbf{x}'_i \beta_j$ is for the logit comparing category j to category m , not for the logit comparing category j to its complement, $\log [\mu_{ij}/(1 - \mu_{ij})]$.

Suppose that we compute the fitted value at \mathbf{x}'_0 (e.g., a focal point in an effect display). Differentiating μ_{0j} with respect to the model parameters yields³

$$\begin{aligned}\frac{\partial \mu_{0j}}{\partial \beta_j} &= \frac{\exp(\mathbf{x}'_0 \beta_j) \left[1 + \sum_{j'=1, j' \neq j}^{m-1} \exp(\mathbf{x}'_0 \beta_{j'}) \right] \mathbf{x}_0}{\left[1 + \sum_{j'=1}^{m-1} \exp(\mathbf{x}'_0 \beta_{j'}) \right]^2} \\ \frac{\partial \mu_{0j}}{\partial \beta_{j' \neq j}} &= - \frac{\left\{ \exp \left[\mathbf{x}'_0 (\beta_{j'} + \beta_j) \right] \right\} \mathbf{x}_0}{\left[1 + \sum_{j'=1}^{m-1} \exp(\mathbf{x}'_0 \beta_{j'}) \right]^2} \\ \frac{\partial \mu_{0m}}{\partial \beta_j} &= - \frac{\exp(\mathbf{x}'_0 \beta_j) \mathbf{x}_0}{\left[1 + \sum_{j'=1}^{m-1} \exp(\mathbf{x}'_0 \beta_{j'}) \right]^2}\end{aligned}$$

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

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

Here, $r = p(m - 1)$ represents the total number of parameters in the combined parameter vectors. $\widehat{\mathcal{V}}(\widehat{\beta})$ is typically computed along with $\widehat{\beta}$ when the model is estimated. Then, by the delta method (e.g., Schervish, 1995: Section 7.1.3),

$$\widehat{\mathcal{V}}(\widehat{\mu}_{0j}) \simeq \sum_{s=1}^r \sum_{t=1}^r v_{st} \frac{\partial \widehat{\mu}_{0j}}{\partial \widehat{\beta}_s} \frac{\partial \widehat{\mu}_{0j}}{\partial \widehat{\beta}_t} \quad (3)$$

(where \simeq denotes approximation).

Because the $\widehat{\mu}_{0j}$ are bounded by 0 and 1, confidence intervals on the probability scale are problematic, especially for values near the boundaries. We therefore suggest the following refinement: Re-express the category probabilities μ_{0j} as logits,

$$\lambda_{0j} = \log \frac{\mu_{0j}}{1 - \mu_{0j}} \quad (4)$$

Note that these are *not* the paired-category logits to which the parameters of the multinomial logit model 2 directly pertain. Differentiating equation 4 with respect to μ_{0j} produces

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

and, consequently, by a second application of the delta method,

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

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

³We are grateful to Georges Monette for checking these derivatives.

3.2 The Proportional-Odds Logit Model

The proportional-odds logit model is a common model for an ordinal response variable (e.g., Powers and Xie, 2000: Chapter 6; Fox, 1997: Chapter 15). The model is often motivated as follows: 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:

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

We cannot observe ξ directly, but instead implicitly dissect its range into m class intervals at the (unknown) cut-points $\alpha_1 < \alpha_2 < \dots < \alpha_{m-1}$, producing the observed ordinal response variable y . That is,

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

The cumulative probability distribution of y_i is given by

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

for $j = 1, 2, \dots, m-1$. 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:

$$\begin{aligned}\text{logit}[\Pr(y_i > j)] &= \log_e \frac{\Pr(y_i > j)}{\Pr(y_i \leq j)} \\ &= -\alpha_j + \beta' \mathbf{x}_i\end{aligned}\tag{5}$$

for $j = 1, 2, \dots, m-1$. (The similar ordered probit model is produced by assuming that the ε_i are normally distributed.)

Model 5 is over-parametrized: Since the β vector typically includes a constant, say β_1 , we have $m-1$ regression equations, the intercepts of which are expressed in terms of m (i.e., one too many) parameters. A solution is to eliminate the constant from β . Setting $\beta_1 = 0$ in this manner in effect establishes the origin of the latent continuum ξ ; we already implicitly established the scale of ξ by fixing the variance of the error to the variance of the standard logistic distribution ($\pi^2/3$). For convenience, we will absorb the negative sign into the intercept, rewriting the model as

$$\text{logit}[\Pr(y_i > j)] = \alpha_j + \beta' \mathbf{x}_i, \text{ for } j = 1, 2, \dots, m-1$$

Then the cut-points are the negatives of the intercepts α_j .

When it adequately represents the data, the proportional-odds model is more parsimonious than the multinomial logit model (and other models for unordered polytomies): While the proportional-odds model has $m + p - 2$ independent parameters, the multinomial logit model has $p(m - 1)$ independent parameters.

We consider two strategies for constructing effect displays for the proportional-odds model. The more straightforward strategy is to plot on the scale of the latent continuum, using the estimated cut-points, $-\widehat{\alpha}_j$, to show the division of the continuum into ordered categories. There is not much more to say about this approach, since — other than marking the cut-points — one proceeds exactly as for a linear model (as illustrated in the example in Section 4.2).

The second approach is to display fitted probabilities of category membership, as we did for the multinomial logit model. Suppose that we need the fitted probabilities at \mathbf{x}'_0 (where the constant regressor has been removed from the design vector \mathbf{x}' , and the intercept from the parameter vector β). Let $\eta_0 = \mathbf{x}'_0\beta$, and let $\mu_{0j} = \Pr(Y_0 = j)$. Then

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

As in the case of the multinomial logit model, we derive approximate standard errors by the delta method. The necessary derivatives are messier here, however:

$$\begin{aligned}
\frac{\partial \mu_{01}}{\partial \alpha_1} &= -\frac{\exp(\alpha_1 + \eta_0)}{[1 + \exp(\alpha_1 + \eta_0)]^2} \\
\frac{\partial \mu_{01}}{\partial \alpha_j} &= 0, \quad j = 2, \dots, m-1 \\
\frac{\partial \mu_{01}}{\partial \beta} &= -\frac{\exp(\alpha_1 + \eta_0) \mathbf{x}_0}{[\exp(\alpha_1 + \eta_0)]^2} \\
\frac{\partial \mu_{0j}}{\partial \alpha_{j-1}} &= \frac{\exp(\alpha_{j-1} + \eta_0)}{[1 + \exp(\alpha_{j-1} + \eta_0)]^2} \\
\frac{\partial \mu_{0j}}{\partial \alpha_j} &= -\frac{\exp(\alpha_j + \eta_0)}{[1 + \exp(\alpha_j + \eta_0)]^2} \\
\frac{\partial \mu_{0j}}{\partial \alpha_{j'}} &= 0, \quad j' \neq j, j-1 \\
\frac{\partial \mu_{0j}}{\partial \beta} &= \frac{\exp(\eta_0) [\exp(\alpha_j) - \exp(\alpha_{j-1})] [\exp(\alpha_{j-1} + \alpha_j + 2\eta_0) - 1] \mathbf{x}_0}{[1 + \exp(\alpha_{j-1} + \eta_0)]^2 [1 + \exp(\alpha_j + \eta_0)]^2} \\
\frac{\partial \mu_{0m}}{\partial \alpha_{m-1}} &= \frac{\exp(\alpha_{m-1} + \eta_0)}{[1 + \exp(\alpha_{m-1} + \eta_0)]^2} \\
\frac{\partial \mu_{0m}}{\partial \alpha_j} &= 0, \quad j = 1, \dots, m-2 \\
\frac{\partial \mu_{0m}}{\partial \beta} &= \frac{\exp(\alpha_{m-1} + \eta_0) \mathbf{x}_{0Stak}}{[1 + \exp(\alpha_{m-1} + \eta_0)]^2}
\end{aligned}$$

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

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

where $r = m + p - 2$. Then, as for the multinomial logit model,

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

and

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

where

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

are the individual-category logits.

4 Examples

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

The example in this section is adapted from work by Andersen, Heath and Sinnott's (2002) on political knowledge and electoral choices in Britain (see also Andersen, Tilley and Heath, in press).

The data are from the 1997-2001 British Election Panel Study (BEPS). Although the same respondents were questioned at eight points in time, we use information only from the final wave of

the study, which was conducted shortly following the 2001 British election. After removing cases with missing data, the sample size is 2206.

We fit a multinomial logit model to describe how attitude towards European integration—an important issue during the 2001 British election—and knowledge of the major political parties’ stances on Europe interact in their effect on party choice. The variables in the model are as follows:

- The response variable is party choice, which has three categories: Labour, Conservative, and Liberal Democrat. Those who voted for other parties are excluded from the analysis.
- “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 are shown in Table 3.

We have already argued that interpreting coefficients in logit models is not simple, especially in the presence of interactions. Interpretation of the multinomial logit model is further complicated because the coefficients refer to contrasts of categories of the response variable with a baseline category. Nonetheless, we can see even from the coefficients that attitude towards Europe was related to party choice and that this relationship differed according to level of political knowledge. An analysis of deviance confirms that the interaction between attitude towards Europe and political knowledge is statistically significant. As was the case with the binary logit model, however, further interpretation is simplified by plotting this interaction as an effect display.

Figure 4 displays the relationship between attitude towards Europe and the fitted probability of voting for each of the three parties at the several levels of political knowledge (ranging from 0 to 3). It is much easier to interpret the interaction between attitude and knowledge in this effect plot than directly from the coefficients: At the lowest level of knowledge, there is apparently no relationship between attitude towards Europe and party choice. In contrast, as knowledge increases, voters are progressively more likely to match their attitudes to party platforms — that is, the more Eurosceptic voters are, the more likely they are to support the Conservative Party and the less likely they are to support Labour or the Liberal Democrats.

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

We now turn to an application of effect displays to a proportional-odds logit model. Data for this example are taken from the World Values Survey of 1995-97 (Inglehart et al., 2000). We use a subset of the World Values Survey, focusing on four countries (with sample sizes in parentheses): Australia (1874), Norway (1127), Sweden (1003), and the United States (1377). Although the variables that we employ are available for more than 40 countries, we restrict attention to these four nations to simplify the example. The variables in the model are as follows:

<i>Coefficient</i>	<i>Labour/Liberal Democrat</i>	
	<i>Estimate</i>	<i>Standard Error</i>
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 Economic Position	0.171	0.082
Evaluation of Blair (Labour leader)	0.546	0.071
Evaluation of Hague (Conservative leader)	-0.088	0.064
Evaluation of Kennedy (Liberal Democrat leader)	-0.416	0.072
Europe	-0.070	0.040
Political Knowledge	-0.502	0.155
Europe \times Knowledge	0.024	0.021

<i>Coefficient</i>	<i>Conservative/Liberal Democrat</i>	
	<i>Estimate</i>	<i>Standard Error</i>
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 Economic Position	-0.008	0.101
Evaluation of Blair (Labour leader)	-0.278	0.079
Evaluation of Hague (Conservative leader)	0.781	0.079
Evaluation of Kennedy (Liberal Democrat leader)	-0.656	0.086
Europe	-0.068	0.049
Political Knowledge	-1.160	0.219
Europe \times Knowledge	0.183	0.028

Table 3: Coefficients for a multinomial logit model regressing party choice on attitude towards European integration, political knowledge and other explanatory variables.

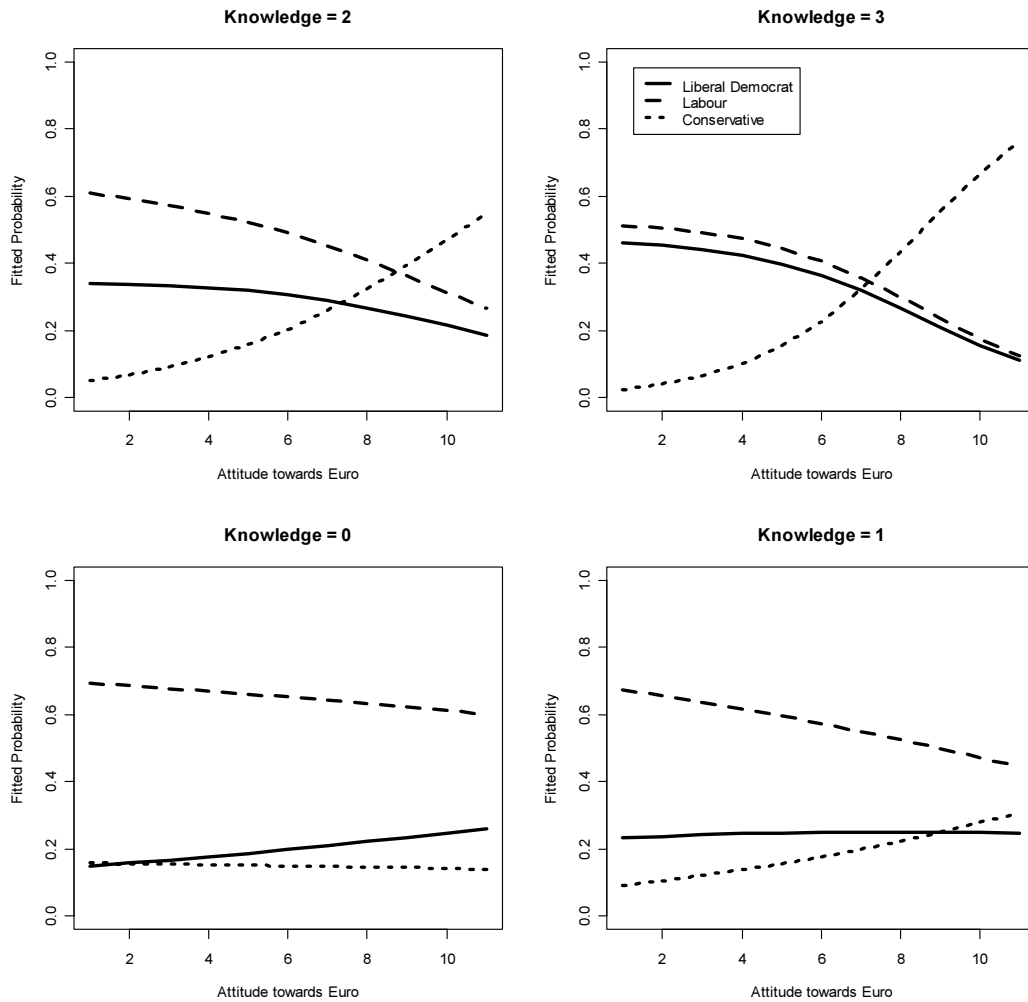


Figure 4: Display of the interaction between attitude towards Europe and political knowledge, showing the effects of these variables on the fitted probability of voting for each of the three major British parties in 2001.

<i>Coefficient</i>	<i>Estimate</i>	<i>Standard Error</i>
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
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
<i>Cut-points</i>		
Too Little About Right	0.449	0.106
About Right Too Much	2.262	0.111

Table 4: Coefficients for a proportional-odds logit model regressing attitude towards government efforts to help people in poverty on gender, age, religion, education, and country. Age is represented in the model by a cubic orthogonal polynomial, and interactions between age and country are included in the model.

- The response variable is produced from answers to the question, “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?” We order the responses as 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 fit to the data are displayed in Table 4.

The complexity of the nonlinear trend for age, its interaction with country, and coefficients for adjacent-category logits make it difficult to interpret directly the parameter estimates associated with age. We therefore again turn to effect displays. Figure 5 plots fitted probabilities for each category of the response variable in the same manner as for the multinomial logit model of Section 4.1. Because country takes on only four values while age is continuous, we construct a separate plot for each country, placing age on the horizontal axis. There are three fitted lines in each plot — representing the fitted probability of choosing each response category. Although this graph is informative — we see, for example, that age differences are relatively muted in the U.S., and that respondents there are less likely than others to feel that the government is not doing enough for the poor — the display does not take advantage of the parsimony of the proportional-odds model.

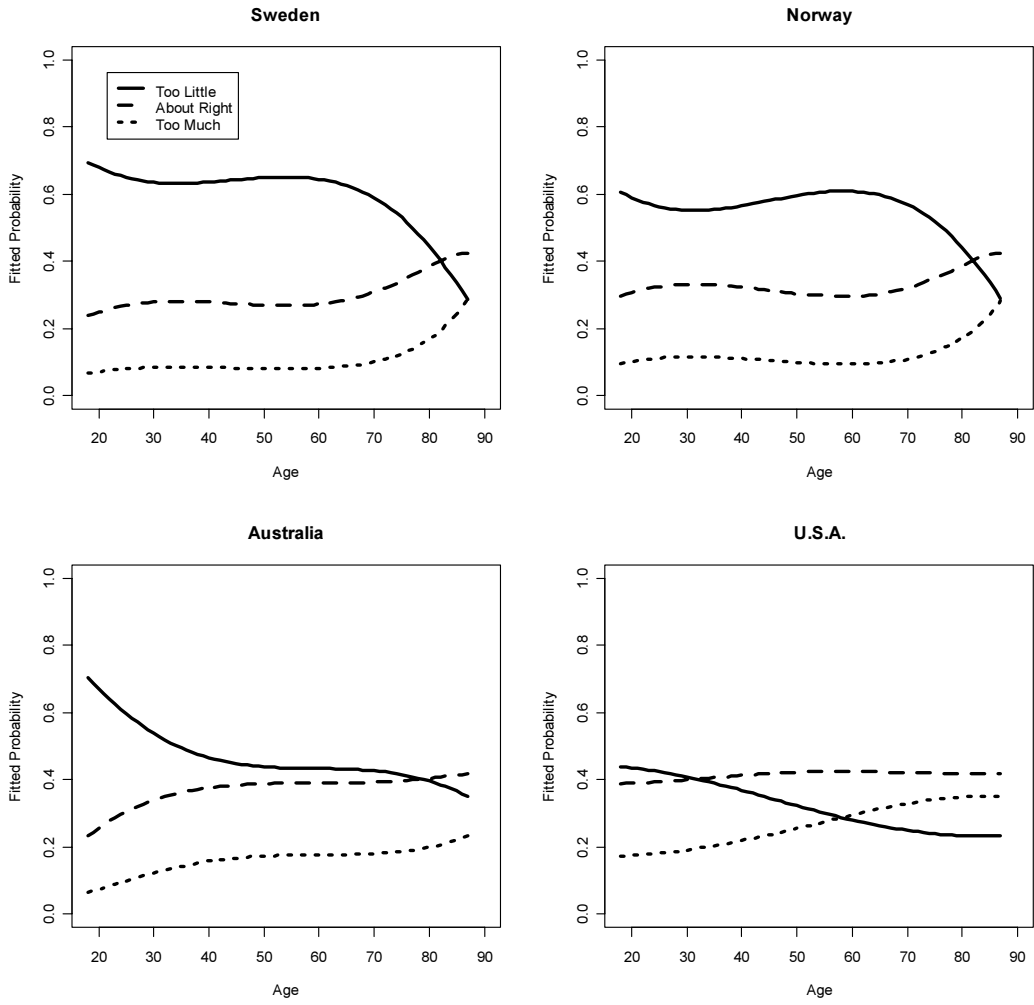


Figure 5: Display of the interaction between age and country, showing the effects of these variables on attitude towards government efforts to help people in poverty; the graphs indicate the fitted probability for each of the three categories of the response variable.

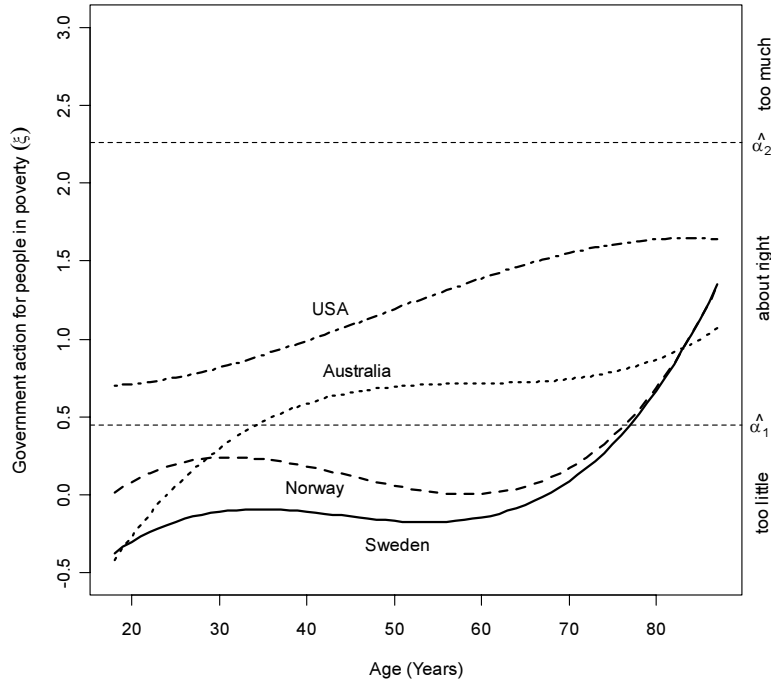


Figure 6: Plotting the interaction between age and country on the latent attitude continuum, ξ . The horizontal lines at $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are the cut-points between adjacent categories of the response.

Alternatively, and more simply, one can capitalize on the structure of the proportional-odds model to plot the fitted response on the scale of the latent attitude continuum. We pursue this strategy in Figure 6, in which there is only one line for each country. The estimated cut-points from the proportional-odds model are displayed as horizontal lines, dividing the latent continuum into three categories. Notice that none of the fitted curves exceeds the second cut-point, and it is therefore unnecessary to include this cut-point in the graph; we do so to show explicitly that “too much” is never the modal response.

References

- Andersen, R., Heath, A., & Sinnott, R. (2002). Political knowledge and electoral choice. *British Elections and Parties Review*, 12, 11–27.
- Andersen, R., Tilley, J., & Heath, A. (in press). Political knowledge and enlightened preferences. *British Journal of Political Science*.
- Firth, D. (1991). Generalized linear models. In D. V. Hinkley, N. Reid, & E. J. Snell (Eds.), *Statistical theory and modeling: In honour of Sir David Cox, FRS* (pp. 55–82). London: Chapman and Hall.
- Fisher, R. A. (1936). *Statistical methods for research workers, 6th edition*. Edinburgh: Oliver and Boyd.

- Fox, J. (1987). Effect displays for generalized linear models. In C. C. Clogg (Ed.), *Sociological methodology 1987* (pp. 347–361). Washington DC: American Sociological Association.
- Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Thousand Oaks, CA: Sage.
- Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software*, 8(15), 1–27.
- Inglehart, R. e. A. (2000). *World values surveys and european value surveys, 1981–1984, 1990–1993, and 1995–1997 [computer file]*. Ann Arbor MI: Institute for Social Research [producer], Inter-University Consortium for Political and Social Research [distributor].
- McCullagh, P., & Nelder, J. A. (1998). *Generalized linear models, second edition*. London: Chapman and Hall.
- Nelder, J. A. (1977). A reformulation of linear models [with commentary]. *Journal of the Royal Statistical Society, Series A*, 140, 48–76.
- Powers, D. A., & Xie, Y. (2000). *Statistical methods for categorical data analysis*. San Diego: Academic Press.
- Schervish, M. J. (1995). *Theory of statistics*. New York: Springer.