Notes

Logit and Probit Models

1. Topics

- Models for dichotmous data
- Models for polytomous data (as time permits)
- Implementation of logit and probit models in R

2. Models for Dichotomous Data

- To understand why logit and probit models for qualitative data are required, let us begin by examining a representative problem, attempting to apply linear regression to it:
  - In September of 1988, 15 years after the coup of 1973, the people of Chile voted in a plebiscite to decide the future of the military government. A ‘yes’ vote would represent eight more years of military rule; a ‘no’ vote would return the country to civilian government. The no side won the plebiscite, by a clear if not overwhelming margin.
  - Six months before the plebiscite, FLACSO/Chile conducted a national survey of 2,700 randomly selected Chilean voters.
    - Of these individuals, 868 said that they were planning to vote yes, and 889 said that they were planning to vote no.
    - Of the remainder, 558 said that they were undecided, 187 said that they planned to abstain, and 168 did not answer the question.

  - I will look only at those who expressed a preference.

- Figure 1 plots voting intention against a measure of support for the status quo.
  - Voting intention appears as a dummy variable, coded 1 for yes, 0 for no.
  - Support for the status quo is a scale formed from a number of questions about political, social, and economic policies: High scores represent general support for the policies of the miliary regime.
- Does it make sense to think of regression as a conditional average when the response variable is dichotomous?
  - An average between 0 and 1 represents a 'score' for the dummy response variable that cannot be realized by any individual.

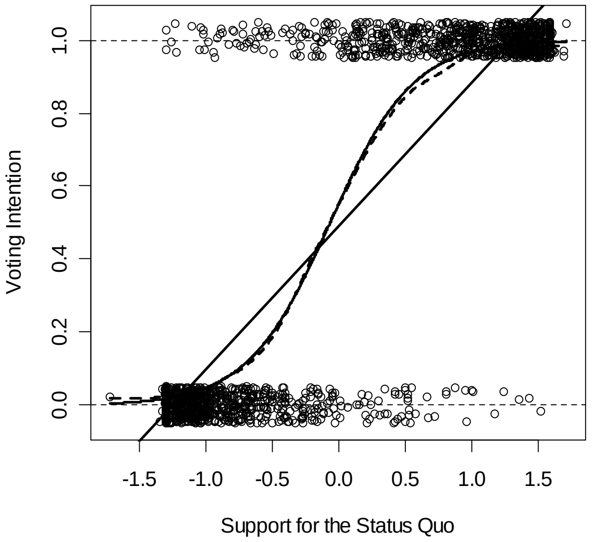

Figure 1. The Chilean plebiscite data: The solid straight line is a linear least-squares fit; the solid curved line is a logistic-regression fit; and the broken line is from a nonparametric kernel regression with a span of .4.The individual observations are all at 0 or 1 and are vertically jittered.

- In the population, the conditional average $E(Y|x_i)$ is the proportion of 1's among those individuals who share the value x_i for the explanatory variable — the conditional probability π_i of sampling a 'yes' in this group:
$$\pi_i \equiv \Pr(Y_i) \equiv \Pr(Y = 1|X = x_i)$$
and thus,
$$E(Y|x_i) = \pi_i(1) + (1 - \pi_i)(0) = \pi_i$$

- If X is discrete, then in a sample we can calculate the conditional proportion for Y at each value of X.
  - The collection of these conditional proportions represents the sample nonparametric regression of the dichotomous Y on X.
  - In the present example, X is continuous, but we can nevertheless resort to strategies such as local averaging, as illustrated in the figure.

2.1 The Linear-Probability Model

▶ Although non-parametric regression works here, it would be useful to capture the dependency of Y on X as a simple function, particularly when there are several explanatory variables.

▶ Let us first try linear regression with the usual assumptions:
$$Y_i = \alpha + \beta X_i + \varepsilon_i$$
where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and ε_i and ε_j are independent for $i \neq j$.
- If X is random, then we assume that it is independent of ε.

▶ Under this model, $E(Y_i) = \alpha + \beta X_i$, and so
$$\pi_i = \alpha + \beta X_i$$
- For this reason, the linear-regression model applied to a dummy response variable is called the *linear probability model*.

▶ This model is untenable, but its failure points the way towards more adequate specifications:

- *Non-normality:* Because Y_i can take on only the values of 0 and 1, the error ε_i is dichotomous as well — not normally distributed:
  - If $Y_i = 1$, which occurs with probability π_i, then
$$\begin{aligned}\varepsilon_i &= 1 - E(Y_i)\\ &= 1 - (\alpha + \beta X_i)\\ &= 1 - \pi_i\end{aligned}$$
  - Alternatively, if $Y_i = 0$, which occurs with probability $1 - \pi_i$, then
$$\begin{aligned}\varepsilon_i &= 0 - E(Y_i)\\ &= 0 - (\alpha + \beta X_i)\\ &= 0 - \pi_i\\ &= -\pi_i\end{aligned}$$
  - Because of the central-limit theorem, however, the assumption of normality is not critical to least-squares estimation of the normal-probability model.

- *Non-constant error variance:* If the assumption of linearity holds over the range of the data, then $E(\varepsilon_i) = 0$.
  - Using the relations just noted,
$$\begin{aligned}V(\varepsilon_i) &= \pi_i(1-\pi_i)^2 + (1-\pi_i)(-\pi_i)^2\\ &= \pi_i(1-\pi_i)\end{aligned}$$
  - The heteroscedasticity of the errors bodes ill for ordinary-least-squares estimation of the linear probability model, but only if the probabilities π_i get close to 0 or 1.
- *Nonlinearity:* Most seriously, the assumption that $E(\varepsilon_i) = 0$ — that is, the assumption of linearity — is only tenable over a limited range of X-values.
  - If the range of the X's is sufficiently broad, then the linear specification cannot confine π to the unit interval $[0, 1]$.
  - It makes no sense, of course, to interpret a number outside of the unit interval as a probability.

  - · This difficulty is illustrated in the plot of the Chilean plebiscite data, in which the least-squares line produces fitted probabilities below 0 at low levels and above 1 at high levels of support for the status-quo.

- ▶ Dummy *regressor* variables do not cause comparable difficulties because the general linear model makes no distributional assumptions about the X's.
- ▶ Nevertheless, if π doesn't get too close to 0 or 1, the linear-probability model estimated by least-squares frequently provides results similar to those produced by more generally adequate methods.
- ▶ One solution — though not a good one — is simply to constrain π to the unit interval:
$$\pi = \begin{cases} 0 & \text{for } 0 > \alpha + \beta X \\ \alpha + \beta X & \text{for } 0 \le \alpha + \beta X \le 1 \\ 1 & \text{for } \alpha + \beta X > 1 \end{cases}$$

- ▶ The *constrained linear-probability* model fit to the Chilean plebiscite data by maximum likelihood is shown in Figure 2. Although it cannot be dismissed on logical grounds, this model has certain unattractive features:
  - • *Instability:* The critical issue in estimating the linear-probability model is identifying the X-values at which π reaches 0 and 1, since the line $\pi = \alpha + \beta X$ is determined by these two points. As a consequence, estimation of the model is inherently unstable.
  - • *Impracticality:* It is much more difficult to estimate the constrained linear-probability model when there are several X's.
  - • *Unreasonableness:* Most fundamentally, the abrupt changes in slope at $\pi = 0$ and $\pi = 1$ are unreasonable. A smoother relationship between π and X, is more generally sensible.

2.2 Transformations of π: Logit and Probit Models

▶ To insure that π stays between 0 and 1, we require a positive monotone (i.e., non-decreasing) function that maps the 'linear predictor' $\eta = \alpha + \beta X$ into the unit interval.

- A transformation of this type will retain the fundamentally linear structure of the model while avoiding probabilities below 0 or above 1.
- Any cumulative probability distribution function meets this requirement:
$$\pi_i = P(\eta_i) = P(\alpha + \beta X_i)$$
where the CDF $P(\cdot)$ is selected in advance, and α and β are then parameters to be estimated.
- If we choose $P(\cdot)$ as the cumulative rectangular distribution then we obtain the constrained linear-probability model.
- An *a priori* reasonable $P(\cdot)$ should be both smooth and symmetric, and should approach $\pi = 0$ and $\pi = 1$ as asymptotes.

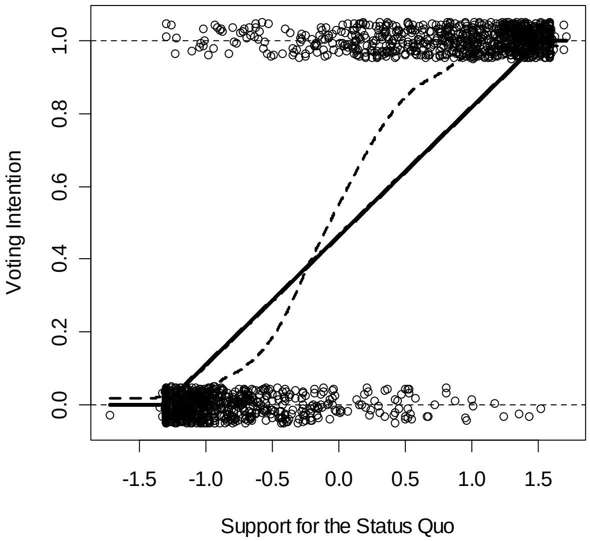

Figure 2. The solid line shows the constrained linear-probability model fit by maximum likelihood to the Chilean plebiscite data; the broken line is for a nonparametric kernel regression.

- Moreover, it is advantageous if $P(\cdot)$ is strictly increasing, permitting us to rewrite the model as
$$P^{-1}(\pi_i) = \eta_i = \alpha + \beta X_i$$
where $P^{-1}(\cdot)$ is the inverse of the CDF $P(\cdot)$, i.e., the quantile function.
  - · Thus, we have a linear model for a transformation of π, or — equivalently — a nonlinear model for π itself.

▶ The transformation $P(\cdot)$ is often chosen as the CDF of the unit-normal distribution
$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}Z^2} dZ$$
or, even more commonly, of the *logistic distribution*
$$\Lambda(z) = \frac{1}{1 + e^{-z}}$$
where $\pi \approx 3.141$ and $e \approx 2.718$ are the familiar mathematical constants.

- Using the normal distribution $\Phi(\cdot)$ yields the *linear probit model*:
$$\begin{aligned}\pi_i &= \Phi(\alpha + \beta X_i) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta X_i} e^{-\frac{1}{2}Z^2} dZ\end{aligned}$$
- Using the logistic distribution $\Lambda(\cdot)$ produces the *linear logistic-regression* or *linear logit model*:
$$\begin{aligned}\pi_i &= \Lambda(\alpha + \beta X_i) \\ &= \frac{1}{1 + e^{-(\alpha+\beta X_i)}}\end{aligned}$$
- Once their variances are equated, the logit and probit transformations are so similar that it is not possible in practice to distinguish between them, as is apparent in Figure 3.
- Both functions are nearly linear between about $\pi = .2$ and $\pi = .8$. This is why the linear probability model produces results similar to the logit and probit models, except when there are extreme values of π_i.

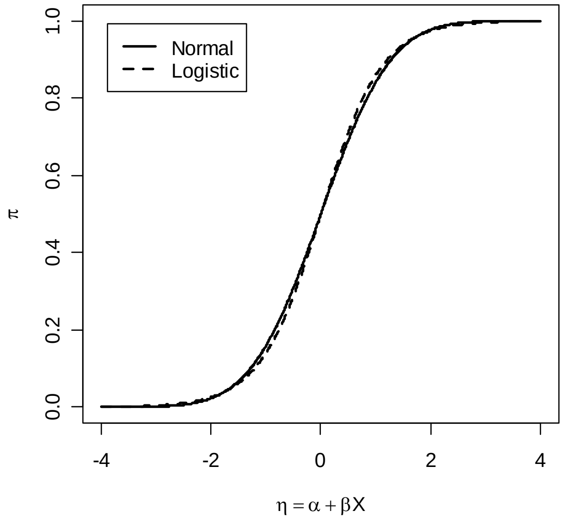

Figure 3. The normal and logistic cumulative distribution functions (as a function of the linear predictor and with variances equated).

▶ Despite their similarity, there are two practical advantages of the logit model:

1. *Simplicity:* The equation of the logistic CDF is very simple, while the normal CDF involves an unevaluated integral.
   - This difference is trivial for dichotomous data, but for polytomous data, where we will require the *multivariate* logistic or normal distribution, the disadvantage of the probit model is more acute.
2. *Interpretability:* The inverse linearizing transformation for the logit model, $\Lambda^{-1}(\pi)$, is directly interpretable as a *log-odds*, while the inverse transformation $\Phi^{-1}(\pi)$ does not have a direct interpretation.
   - Rearranging the equation for the logit model,
   $$\frac{\pi_i}{1-\pi_i} = e^{\alpha+\beta X_i}$$
   - The ratio $\pi_i/(1-\pi_i)$ is the *odds* that $Y_i = 1$, an expression of relative chances familiar to gamblers.

- Taking the log of both sides of this equation,

$$\log_e \frac{\pi_i}{1-\pi_i} = \alpha + \beta X_i$$

- The inverse transformation $\Lambda^{-1}(\pi) = \log_e[\pi/(1-\pi)]$, called the *logit* of π, is therefore the log of the odds that Y is 1 rather than 0.
- The logit is symmetric around 0, and unbounded both above and below, making the logit a good candidate for the response-variable side of a linear model:

Probability π	Odds $\dfrac{\pi}{1-\pi}$	Logit $\log_e \dfrac{\pi}{1-\pi}$
$.01$	$1/99 = 0.0101$	-4.60
$.05$	$5/95 = 0.0526$	-2.94
$.10$	$1/9 = 0.1111$	-2.20
$.30$	$3/7 = 0.4286$	-0.85
$.50$	$5/5 = 1$	0.00
$.70$	$7/3 = 2.333$	0.85
$.90$	$9/1 = 9$	2.20
$.95$	$95/5 = 19$	2.94
$.99$	$99/1 = 99$	4.60

- The logit model is also a multiplicative model for the odds:

$$\begin{aligned}\frac{\pi_i}{1-\pi_i} &= e^{\alpha+\beta X_i} = e^{\alpha}e^{\beta X_i} \\ &= e^{\alpha}\left(e^{\beta}\right)^{X_i}\end{aligned}$$

  - So, increasing X by 1 changes the logit by β and multiplies the odds by e^{β}.
  - For example, if $\beta = 2$, then increasing X by 1 increases the odds by a factor of $e^2 \approx 2.718^2 = 7.389$.

- Still another way of understanding the parameter β in the logit model is to consider the slope of the relationship between π and X.
  - Since this relationship is nonlinear, the slope is not constant; the slope is $\beta\pi(1-\pi)$, and hence is at a maximum when $\pi = 1/2$, where the slope is $\beta/4$:

π	$\beta\pi(1-\pi)$
$.01$	$\beta \times .0099$
$.05$	$\beta \times .0475$
$.10$	$\beta \times .09$
$.20$	$\beta \times .16$
$.50$	$\beta \times .25$
$.80$	$\beta \times .16$
$.90$	$\beta \times .09$
$.95$	$\beta \times .0475$
$.99$	$\beta \times .0099$

  - The slope does not change very much between $\pi = .2$ and $\pi = .8$, reflecting the near linearity of the logistic curve in this range.

▶ The least-squares line fit to the Chilean plebescite data has the equation

$$\widehat{\pi}_{\text{yes}} = 0.492 + 0.394 \times \text{ Status-Quo}$$

- This line is a poor summary of the data.

- The logistic-regression model, fit by the method of maximum-likelihood, has the equation

$$\log_e \frac{\widehat{\pi}_{\text{yes}}}{\widehat{\pi}_{\text{no}}} = 0.215 + 3.21 \times \text{ Status-Quo}$$

  - The logit model produces a much more adequate summary of the data, one that is very close to the nonparametric regression.
  - Increasing support for the status-quo by one unit multiplies the odds of voting yes by $e^{3.21} = 24.8$.
  - Put alternatively, the slope of the relationship between the fitted probability of voting yes and support for the status-quo at $\widehat{\pi}_{\text{yes}} = .5$ is $3.21/4 = 0.80$.

2.3 An Unobserved-Variable Formulation

- An alternative derivation posits an underlying regression for a continuous but unobservable response variable ξ (representing, e.g., the ‘propensity’ to vote yes), scaled so that

$$Y_i = \begin{cases} 0 \text{ when } \xi_i \leq 0 \\ 1 \text{ when } \xi_i > 0 \end{cases}$$

  - That is, when ξ crosses 0, the observed discrete response Y changes from ‘no’ to ‘yes.’
  - The latent variable ξ is assumed to be a linear function of the explanatory variable X and the unobservable error variable ε:

$$\xi_i = \alpha + \beta X_i - \varepsilon_i$$

- We want to estimate α and β, but cannot proceed by least-squares regression of ξ on X because the latent response variable is not directly observed.

- Using these equations,

$$\begin{aligned}\pi_i &\equiv \Pr(Y_i = 1) = \Pr(\xi_i > 0) = \Pr(\alpha + \beta X_i - \varepsilon_i > 0)\\ &= \Pr(\varepsilon_i < \alpha + \beta X_i)\end{aligned}$$

  - If the errors are independently distributed according to the unit-normal distribution, $\varepsilon_i \sim N(0,1)$, then

    $$\pi_i = \Pr(\varepsilon_i < \alpha + \beta X_i) = \Phi(\alpha + \beta X_i)$$

    which is the probit model.
  - Alternatively, if the ε_i follow the similar logistic distribution, then we get the logit model

    $$\pi_i = \Pr(\varepsilon_i < \alpha + \beta X_i) = \Lambda(\alpha + \beta X_i)$$

- We will return to the unobserved-variable formulation when we consider models for ordinal categorical data.

2.4 Logit and Probit Models for Multiple Regression

- To generalize the logit and probit models to several explanatory variables we require a linear predictor that is a function of several regressors.
  - For the logit model,

    $$\begin{aligned}\pi_i &= \Lambda(\eta_i) = \Lambda(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})\\ &= \Lambda(\mathbf{x}_i'\boldsymbol{\beta})\\ &= \frac{1}{1 + e^{-\mathbf{x}_i'\boldsymbol{\beta}}}\end{aligned}$$

    or, equivalently,

    $$\begin{aligned}\log_e \frac{\pi_i}{1 - \pi_i} &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}\\ &= \mathbf{x}_i'\boldsymbol{\beta}\end{aligned}$$

  - For the probit model,

    $$\pi_i = \Phi(\eta_i) = \Phi(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$$

- The X's in the linear predictor can be as general as in the general linear model, including, for example:

- quantitative explanatory variables;
- transformations of quantitative explanatory variables;
- polynomial regressors formed from quantitative explanatory variables;
- dummy regressors representing qualitative explanatory variables; and
- interaction regressors.

▶ Interpretation of the partial regression coefficients in the general logit model is similar to the interpretation of the slope in the logit simple-regression model, with the additional provision of holding other explanatory variables in the model constant.

- Expressing the model in terms of odds,

$$\begin{aligned}\frac{\pi_i}{1-\pi_i} &= e^{(\alpha+\beta_1 X_{i1}+\cdots+\beta_k X_{ik})} \\ &= e^{\alpha}\left(e^{\beta_1}\right)^{X_{i1}}\cdots\left(e^{\beta_k}\right)^{X_{ik}}\end{aligned}$$

- Thus, e^{β_j} is the multiplicative effect on the odds of increasing X_j by 1, holding the other X's constant.

- Similarly, $\beta_j/4$ is the slope of the logistic regression surface in the direction of X_j at $\pi = .5$.

▶ The general linear logit and probit models can be fit to data by the method of maximum likelihood.

▶ Hypothesis tests and confidence intervals follow from general procedures for statistical inference in maximum-likelihood estimation.

- For an individual coefficient, it is most convenient to test the hypothesis H_0: $\beta_j = \beta_j^{(0)}$ by calculating the Wald statistic

$$Z_0 = \frac{B_j - \beta_j^{(0)}}{\text{SE}(B_j)}$$

where SE(B_j) is the asymptotic standard error of B_j.

  - · The test statistic Z_0 follows an asymptotic unit-normal distribution under the null hypothesis.

- Similarly, an asymptotic $100(1-a)$-percent confidence interval for β_j is given by
$$\beta_j = B_j \pm z_{a/2}\text{SE}(B_j)$$
where $z_{a/2}$ is the value from $Z \sim N(0,1)$ with a probability of $a/2$ to the right.
- Wald tests for several coefficients can be formulated from the estimated asymptotic variances and covariances of the coefficients.
- It is also possible to formulate a likelihood-ratio test for the hypothesis that several coefficients are simultaneously zero, H_0: $\beta_1 = \cdots = \beta_q = 0$. We proceed, as in least-squares regression, by fitting two models to the data:
  - The full model (model 1)
$$\text{logit}(\pi) = \alpha + \beta_1 X_1 + \cdots + \beta_q X_q + \beta_{q+1} X_{q+1} + \cdots + \beta_k X_k$$

  - and the null model (model 0)
$$\begin{aligned}\text{logit}(\pi) &= \alpha + 0X_1 + \cdots + 0X_q + \beta_{q+1}X_{q+1} + \cdots + \beta_k X_k \\ &= \alpha + \beta_{q+1}X_{q+1} + \cdots + \beta_k X_k\end{aligned}$$
  - Each model produces a maximized likelihood: L_1 for the full model, L_0 for the null model.
  - Because the null model is a specialization of the full model, $L_1 \geq L_0$.
  - The generalized likelihood-ratio test statistic for the null hypothesis is
$$G_0^2 = 2(\log_e L_1 - \log_e L_0)$$
  - Under the null hypothesis, this test statistic has an asymptotic chisquare distribution with q degrees of freedom.
- A test of the omnibus null hypothesis H_0: $\beta_1 = \cdots = \beta_k = 0$ is obtained by specifying a null model that includes only the constant, logit$(\pi) = \alpha$.

- The likelihood-ratio test can be inverted to produce confidence intervals for coefficients.
- The likelihood-ratio test is less prone to breaking down than the Wald test.

▶ An analog to the multiple-correlation coefficient can also be obtained from the log-likelihood.

- By comparing $\log_e L_0$ for the model containing only the constant with $\log_e L_1$ for the full model, we can measure the degree to which using the explanatory variables improves the predictability of Y.
- The quantity $G^2 \equiv -2\log_e L$, called the *residual deviance* under the model, is a generalization of the residual sum of squares for a linear model.
- Thus,

$$\begin{aligned} R^2 &= 1 - \frac{G_1^2}{G_0^2} \\ &= 1 - \frac{\log_e L_1}{\log_e L_0} \end{aligned}$$

is analogous to R^2 for a linear model.

2.5 Illustration: SLID Data

- To illustrate logistic regression, I will use data from the 1994 wave of the Statistics Canada Survey of Labour and Income Dynamics (the "SLID").
- Confining attention to married women between the ages of 20 and 35, I examine how the labor-force participation of these women is related to several explanatory variables:
  - the region of the country in which the woman resides;
  - the presence of children between zero and four years of age in the household, coded as absent or present;
  - the presence of children between five and nine years of age;
  - the presence of children between ten and fourteen years of age
  - family after-tax income, excluding the woman's own income (if any);
  - education, defined as number of years of schooling.

3. Models for Polytomous Data

- I will describe three general approaches to modeling polytomous data:

1. modeling the polytomy directly as a set of unordered categories, using a generalization of the dichotomous logit model;
2. constructing a set of nested dichotomies from the polytomy, fitting an independent logit or probit model to each dichotomy; and
3. extending the unobserved-variable interpretation of the dichotomous logit and probit models to ordered polytomies.

3.1 The Polytomous Logit Model

- ▶ The dichotomous logit model can be extended to a polytomy by employing the multivariate-logistic distribution. This approach has the advantage of treating the categories of the polytomy in a non-arbitrary, symmetric manner.
- ▶ The response variable Y can take on any of m qualitative values, which, for convenience, we number $1, 2, ..., m$ (using the numbers only as category labels).
  - For example, in the UK, voters can vote for (1) the Conservatives, (2) Labour, or (3) the Liberal Democrats (ignoring other parties).
- ▶ Let π_{ij} denote the probability that the ith observation falls in the jth category of the response variable; that is,

$$\pi_{ij} \equiv \Pr(Y_i = j) \text{ for } j = 1, ..., m.$$

- ▶ We have k regressors, $X_1, ..., X_k$, on which the π_{ij} depend.

- More specifically, suppose that this dependence can be modeled using the *multivariate logistic distribution*:

$$\begin{aligned}\pi_{ij} &= \frac{e^{\gamma_{0j}+\gamma_{1j}X_{i1}+\cdots+\gamma_{kj}X_{ik}}}{1+\sum_{l=2}^{m} e^{\gamma_{0l}+\gamma_{1l}X_{i1}+\cdots+\gamma_{kl}X_{ik}}}\\ &\text{for } j = 2, ..., m\\ \pi_{i1} &= 1-\sum_{j=2}^{m}\pi_{ij}\end{aligned}$$

- There is one set of parameters, $\gamma_{0j}, \gamma_{1j}, ..., \gamma_{kj}$, for each response-variable category but the first; category 1 functions as a type of baseline.
- The use of a baseline category is one way of avoiding redundant parameters because of the restriction that $\sum_{j=1}^{m} \pi_{ij} = 1$.

- Some algebraic manipulation of the model produces

$$\log_e \frac{\pi_{ij}}{\pi_{i1}} = \gamma_{0j} + \gamma_{1j}X_{i1} + \cdots + \gamma_{kj}X_{ik}$$
$$\text{for } j = 2, ..., m$$

- The regression coefficients affect the log-odds of membership in category j versus the baseline category.
- It is also possible to form the log-odds of membership in *any* pair of categories j and j':

$$\begin{aligned}\log_e \frac{\pi_{ij}}{\pi_{ij'}} &= \log_e \left(\frac{\pi_{ij}}{\pi_{im}} \Big/ \frac{\pi_{ij'}}{\pi_{im}} \right) \\ &= \log_e \frac{\pi_{ij}}{\pi_{im}} - \log_e \frac{\pi_{ij'}}{\pi_{im}} \\ &= (\gamma_{0j} - \gamma_{0j'}) + (\gamma_{1j} - \gamma_{1j'})X_{i1} \\ &\quad + \cdots + (\gamma_{kj} - \gamma_{kj'})X_{ik}\end{aligned}$$

  - The regression coefficients for the logit between any pair of categories are the differences between corresponding coefficients.

▶ Now suppose that the model is specialized to a dichotomous response variable. Then, $m = 2$, and

$$\begin{aligned}\log_e \frac{\pi_{i2}}{\pi_{i1}} &= \log_e \frac{\pi_{i2}}{1 - \pi_{i2}} \\ &= \gamma_{02} + \gamma_{12}X_{i1} + \cdots + \gamma_{k2}X_{ik}\end{aligned}$$

- Applied to a dichotomy, the polytomous logit model is identical to the dichotomous logit model.

3.1.1 Illustration: British Election Panel Study

- This example is adapted from work by Andersen, Heath, and Sinnott (2002) on the 2001 British election.
- The central issue addressed in the data analysis is the potential interaction between respondents' political knowledge and political attitudes in determining their vote.
- The response variable, vote, has three categories: Conservative, Labour, and Liberal Democrat.
- There are several explanatory variables:
  - Attitude toward European integration, an 11-point scale, with high scores representing a negative attitude (so-called "Euro-sceptism").
  - Knowledge of the platforms of the three parties on the issue of European integration, with integer scores ranging from 0 through 3. (Labour and the Liberal Democrats supported European integration, the Conservatives were opposed.)

  - Other variables included in the model primarily as "controls"—age, gender, perceptions of national and household economic conditions, and ratings of the three party leaders.

3.2 Nested Dichotomies

- Perhaps the simplest approach to polytomous data is to fit separate models to each of a set of dichotomies derived from the polytomy.
  - These dichotomies are *nested,* making the models statistically independent.
  - Logit models fit to a set of nested dichotomies constitute a model for the polytomy, but are not equivalent to the polytomous logit model previously described.
- A nested set of $m - 1$ dichotomies is produced from an m-category polytomy by successive binary partitions of the categories of the polytomy.
  - Two examples for a four-category variable are shown in Figure 4.
    - In part (a), the dichotomies are {12, 34}, {1, 2}, and {3, 4}.
    - In part (b), the nested dichotomies are {1, 234}, {2, 34}, and {3, 4}.

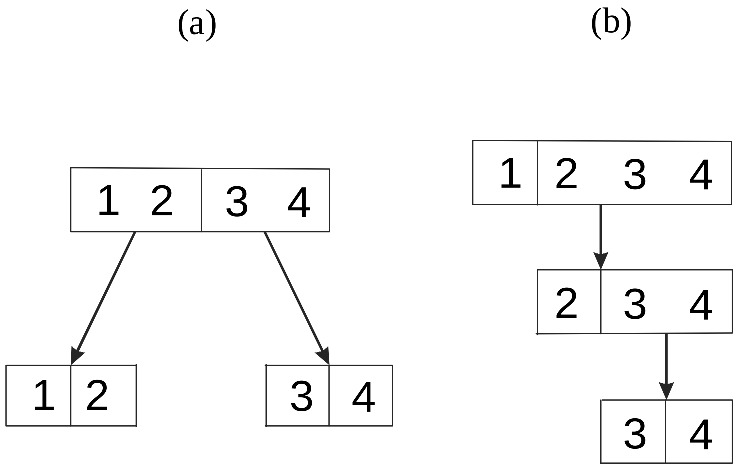

Figure 4. Alternative sets of nested dichotomies for a four-category response.

- Because the results of the analysis and their interpretation depend upon the set of nested dichotomies that is selected, this approach to polytomous data is reasonable only when a particular choice of dichotomies is substantively compelling.
- Nested dichotomies are attractive when the categories of the polytomy represent ordered progress through the stages of a process (called *continuation dichotomies*).
  - Imagine that the categories in (b) represent adults' attained level of education: (1) less than high school; (2) high-school graduate; (3) some post-secondary; (4) post-secondary degree.
  - Since individuals normally progress through these categories in sequence, the dichotomy {1, 234) represents the completion of high school; {2, 34} the continuation to post-secondary education, conditional on high-school graduation; and {3, 4} the completion of a degree conditional on undertaking a post-secondary education.

3.3 Ordered Logit and Probit Models

- Imagine that there is a latent variable ξ that is a linear function of the X's plus a random error:

$$\xi_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

  - Suppose that instead of dividing the range of ξ into two regions to produce a dichotomous response, the range of ξ is dissected by $m-1$ boundaries or *thesholds* into m regions.
  - Denoting the thresholds by $\alpha_1 < \alpha_2 < \cdots < \alpha_{m-1}$, and the resulting response by Y, we observe

$$Y_i = \begin{cases} 1 & \text{if } \xi_i \leq \alpha_1 \\ 2 & \text{if } \alpha_1 < \xi_i \leq \alpha_2 \\ \cdot & \\ \cdot & \\ \cdot & \\ m-1 & \text{if } \alpha_{m-2} < \xi_i \leq \alpha_{m-1} \\ m & \text{if } \alpha_{m-1} < \xi_i \end{cases}$$

- The thresholds, regions, and corresponding values of ξ and Y are represented graphically in Figure 5.

▶ Using the model for the latent variable, along with category thresholds, we can determine the cumulative probability distribution of Y:

$$\begin{aligned}\Pr(Y_i \leq j) &= \Pr(\xi_i \leq \alpha_j)\\ &= \Pr(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i \leq \alpha_j)\\ &= \Pr(\varepsilon_i \leq \alpha_j - \alpha - \beta_1 X_{i1} - \cdots - \beta_k X_{ik})\end{aligned}$$

- If the errors ε_i are independently distributed according to the standard normal distribution, then we obtain the ordered probit model.

- If the errors follow the similar logistic distribution, then we get the ordered logit model:

$$\begin{aligned}\text{logit}[\Pr(Y_i \leq j)] &= \log_e \frac{\Pr(Y_i \leq j)}{\Pr(Y_i > j)}\\ &= \alpha_j - \alpha - \beta_1 X_{i1} - \cdots - \beta_k X_{ik}\end{aligned}$$

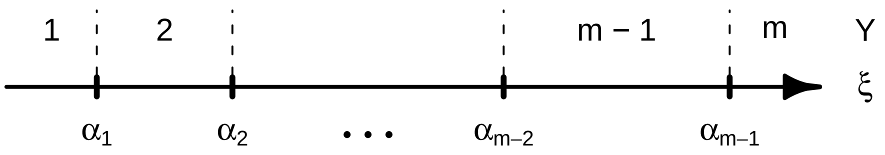

Figure 5. The thresholds $\alpha_1 < \alpha_2 < \cdots < \alpha_{m-1}$ divide the latent continuum ξ into m regions, corresponding to the values of the observable variable Y.

- Equivalently,
$$\text{logit}[\Pr(Y_i > j)] = \log_e \frac{\Pr(Y_i > j)}{\Pr(Y_i \le j)} = (\alpha - \alpha_j) + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$
for $j = 1, 2, ..., m - 1$.
- The logits in this model are for cumulative categories — at each point contrasting categories above category j with category j and below.
- The slopes for each of these regression equations are identical; the equations differ only in their intercepts.
  - · The logistic regression surfaces are therefore horizontally parallel to each other, as illustrated in Figure 6 for $m = 4$ response categories and a single X.
- For a fixed set of X's, any two different cumulative log-odds — say, at categories j and j' — differ only by the constant $(\alpha_j - \alpha_{j'})$.

  - · The odds, therefore, are proportional to one-another, and for this reason, the ordered logit model is called the *proportional-odds model*.

▶ There are $(k + 1) + (m - 1) = k + m$ parameters to estimate in the proportional-odds model, including the regression coefficients $\alpha, \beta_1, ..., \beta_k$ and the category thresholds $\alpha_1, ..., \alpha_{m-1}$.
- There is an extra parameter in the regression equations, since each equation has its own constant, $-\alpha_j$, along with the common constant α.
- A simple solution is to set $\alpha = 0$, producing
$$\text{logit}[\Pr(Y_i > j)] = -\alpha_j + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

▶ Figure 7 illustrates the proportional-odds model for $m = 4$ response categories and a single X.
- The conditional distribution of the latent response variable ξ is shown for two representative values of the explanatory variable, x_1 and x_2.

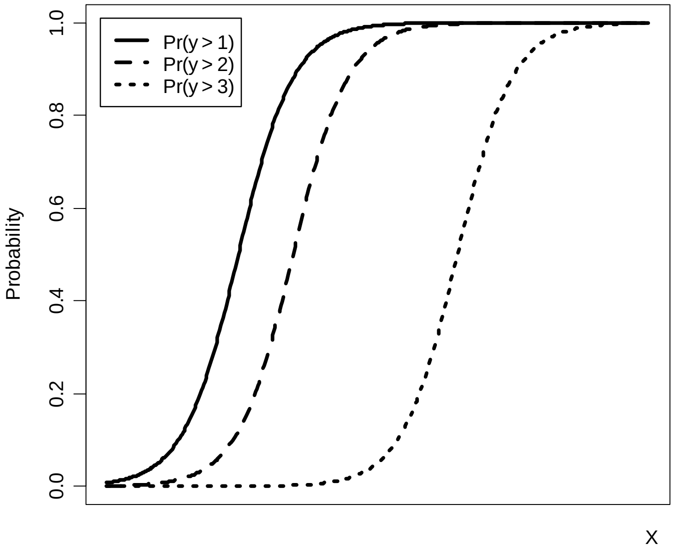

Figure 6. The proportional-odds model for four response categories and a single explanatory variable X.

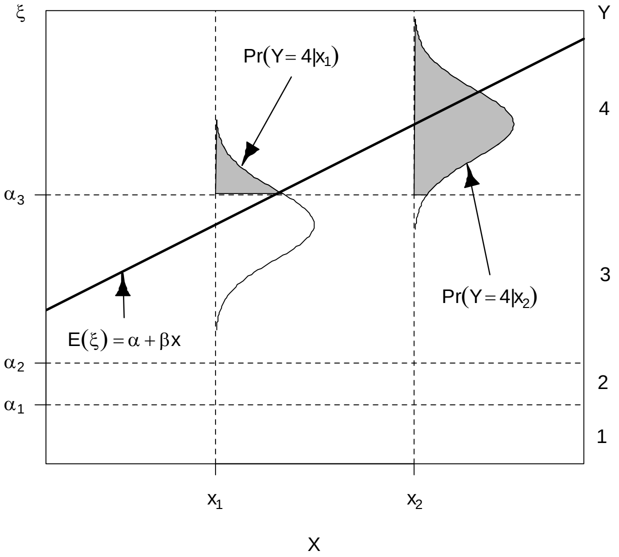

Figure 7. The proportional-odds model for four response categories and a single explanatory variable X.

3.3.1 Illustration: World Value Survey

- ▶ To illustrate the use of the proportional-odds model, I draw on data from the World Values Survey (WVS) of 1995–97
  - Although the WVS collects data in many countries, to provide a manageable example, I will restrict attention to only four: Australia, Sweden, Norway, and the United States. The combined sample size for these four countries is 5381.
- ▶ The response variable in the analysis is the answer to the question, “Do you think that what the government is doing for people in poverty is about the right amount, too much, or too little.” There are, therefore, three ordered categories: *too little*, *about right*, *too much*.

- ▶ There are several explanatory variables:
  - gender (represented by a dummy variable coded 1 for *men* and 0 for *women*);
  - whether or not the respondent belonged to a religion (coded 1 for *yes*, 0 for *no*);
  - whether or not the respondent had a university degree (coded 1 for *yes* and 0 for *no*);
  - age (in years, ranging from 18 to 87); preliminary analysis of the data suggested a roughly linear age effect;
  - country (entered into the model as a set of three dummy regressors, with *Australia* as the base-line category).

3.4 Comparison of the Three Approaches

▶ The three approaches to modeling polytomous data — the polytomous logit model, logit models for nested dichotomies, and the proportional-odds model — address different sets of log-odds, corresponding to different dichotomies constructed from the polytomy.

▶ Consider, for example, the ordered polytomy {1, 2, 3, 4}:

- Treating category 1 as the baseline, the coefficients of the polytomous logit model apply directly to the dichotomies {1, 2}, {1, 3}, and {1, 4}, and indirectly to any pair of categories.
- Forming continuation dichotomies (one of several possibilities), the nested-dichotomies approach models {1, 234}, {2, 34}, and {3, 4}.
- The proportional-odds model applies to the dichotomies {1, 234}, {12, 34}, and {123, 4}, imposing the restriction that only the intercepts of the three regression equations differ.

▶ Which of these models is most appropriate depends partly on the structure of the data and partly upon our interest in them.

4. Logit and Probit Models in R

- Dichotomous logit model:

```
    glm(lhs ~ rhs, family=binomial, data,
        subset, weights, na.action, contrasts)
```

  where `lhs` can take several forms (e.g., a numeric variable with values `0` and `1`, a logical variable with values `TRUE` and `FALSE`, a dichotomous factor), and `rhs` gives the terms in the model as in `lm()`.
  - The arguments `data`, `subset`, `na.action`, and `contrasts` are as in `lm()`.
  - The `weights` argument has a different meaning, representing number of trials for binomial data (in which case, e.g., `lhs` can give the observed proportion of "successes" for each binomial observation).
  - For binomial data, the counts of "successes" and "failures" can also be given as a two-column matrix on the `lhs` of the model formula.

- Dichotomous probit model:

```
glm(lhs ~ rhs,family=binomial(link=probit), data,
    subset, weights, na.action, contrasts)
```

  — just like the logit model except for the explicit specification of the probit "link."

- Multinomial logit model:

```
    multinom(lhs ~ rhs, data,
        subset, weights, na.action, contrasts
```

  using the `multinom()` function from the **nnet** package; `lhs` should be a factor, and the weights argument can give "case weights."

- Proportional-odds model:

```
      polr(lhs ~ rhs, data,
          subset, weights, na.action, contrasts)
```

  using the `polr()` function from the **MASS** package; `lhs` should be an ordered factor or factor, and the weights argument can give “case weights.”

- Not all of the arguments of these functions are shown: see the relevant help pages for details — e.g., `?polr`.