8

Ordered Categorical Variable Models

8.1 MOTIVATION

Many variables that social scientists study are neither continuous nor binary, but rather consist of a (usually small) set of categories that are ranked from low to high. The most familiar example is the five point scale that is widely used in surveys such as the American National Election Study (ANES). Survey designers frequently phrase questions so that respondents must choose among five ordered choices ranging from strongly disagree (at the low end) to strongly agree (at the high end). Intermediate categories are Agree, Don't Know, and Disagree. This assumes an underlying dimension, y^*, in which the various responses are ordered, but we lack a meaningful metric for distance. How far is "disagree" from "strongly agree"?

$$y^* \longleftarrow \text{Strongly Agree} \;\text{——}\; \text{Agree} \;\text{——}\; \text{Don't Know} \;\text{——}\; \text{Disagree} \;\text{——}\; \text{Strongly Disagree} \longrightarrow y^*$$

Sometimes data for which we do have a meaningful distance metric are reported in ordered bins or categories. Age, for example, is often measured in chunks (under 18, 18–24, etc.). Income, especially when solicited as a self-reported value on a survey, is another variable frequently measured in coarse, ordinal categories. Even if the underlying variable is continuous and might be measured at the interval (or even ratio) level, binning the data into categories usually fails to preserve the metric information. Ordinal data instead represent ranks such that we can use inequality operations but not arithmetic ones. We can tell which of two responses is greater, but we cannot say by how much.

If we could measure the distances between adjacent categories, then we could treat the categorization scheme as if it were a continuous variable. In the absence of such information, treating ordered categories as continuous – by

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press

calculating a mean or fitting an OLS model – requires strong assumptions. There are numerous published studies where the authors proceed anyway, sometimes leaving these assumptions as implicit.

Treating ordered data as continuous can lead to problems similar to those we identified with the linear probability model for binary data. Fortunately the likelihood framework allows us to relax assumptions about the metric content of our data. We do this by positing a latent but unobserved continuous variable that is only reported in discrete bins if the underlying value falls between particular cutpoints. The challenge becomes one of estimating the thresholds that divide one category from another.

8.2 THE ORDERED LOGIT MODEL

In modeling ordinal data we generalize the latent variable model introduced in Chapter 3. Suppose our observed data, Y, can take on one of M possible ordered categorical values. We again imagine that the observed Y is an imperfect or imprecise observation of some underlying continuous latent variable, Y^*, such that $y_i < y_j \Rightarrow y_i^* < y_j^*$. Recall that in constructing the latent variable model for the binary case, we assumed, arbitrarily, that we observed a "0" if $y_i^* \leq 0$ and "1" otherwise. In other words, we fixed a threshold, $\tau = 0$. In the binary case where $M = 2$ we needed only one threshold. With $M > 2$ we need to estimate not only the regression parameters but also the thresholds which will divide the underlying latent variable into the observed ordered categories. As usual our approach involves specifying stochastic and systematic components:

$$\begin{aligned}
&Y_i^* \sim f_L(\mu_i, 1) \\
&\quad \Updownarrow \\
&Y_i^* = \mu_i + \varepsilon_i, \quad \varepsilon_i \sim f_L(0,1) \\
&\mu_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} \\
&Y_i = m \iff \tau_{m-1} < Y_i^* \leq \tau_m \quad \forall \quad m \in \{1, \ldots, M\},
\end{aligned}$$

where τ_m are the threshold parameters that divide the unobserved, latent variable into observed, modeled categories. We commonly assume that $\tau_0 = -\infty$ (the probability of being less than the lowest category is 0) and $\tau_M = +\infty$ (the probability of being in some category is 1). This leads to a stochastic component with the following form:

$$\begin{aligned}
\Pr(Y_i = m|\mathbf{x}_i) &= \Pr(\tau_{m-1} < Y_i^* \leq \tau_m|\mathbf{x}_i) \\
&= \Pr(\tau_{m-1} < \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i \leq \tau_m) \\
&= \Pr(\tau_{m-1} - \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} < \varepsilon_i \leq \tau_m - \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta}) \\
&= \Pr(\varepsilon_i \leq \tau_m - \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta}) - \Pr(\varepsilon_i \leq \tau_{m-1} - \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta}).
\end{aligned}$$

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press

The assumption that the errors are logistic implies that[1]

$$\begin{aligned}\Pr(\varepsilon_i \leq \tau_m - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}) = \Lambda(\tau_m - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}) &= \frac{\exp(\tau_m - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})}{1+\exp(\tau_m - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})} \\ &= \Pr(Y^* \leq \tau_m) = \Pr(Y \leq m),\end{aligned}$$

where Λ is the logistic cumulative distribution function (cdf).

It follows that we simply difference the logistic cdf for adjacent categories to find the probability that Y_i falls in any specific category m:

$$\Pr(Y_i = m) = \frac{\exp(\tau_m - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})}{1+\exp(\tau_m - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})} - \frac{\exp(\tau_{m-1} - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})}{1+\exp(\tau_{m-1} - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})}.$$

The cutpoints serve an important purpose: they slice up the density of the underlying latent variable into categories. Suppose we have five observed categories, following the Likert scale widely employed in surveys, where $m \in \{\text{SD}, \text{D}, \text{DK}, \text{A}, \text{SA}\}$.

$$\Pr(Y_i = m) = \begin{cases} \Lambda(\tau_{\text{SD-D}} - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}) & \text{for} \quad m = \text{SD} \\ \Lambda(\tau_{\text{D-DK}} - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}) - \Lambda(\tau_{\text{SD-D}} - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}) & \text{for} \quad m = \text{D} \\ \Lambda(\tau_{\text{DK-A}} - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}) - \Lambda(\tau_{\text{D-DK}} - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}) & \text{for} \quad m = \text{DK} \\ \Lambda(\tau_{\text{A-SA}} - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}) - \Lambda(\tau_{\text{DK-A}} - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}) & \text{for} \quad m = \text{A} \\ 1 - \Lambda(\tau_{\text{A-SA}} - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}) & \text{for} \quad m = \text{SA}. \end{cases} \tag{8.1}$$

This is illustrated in Figure 8.1, which shows example cutpoints defining the observed categories in a hypothetical example. The figure also illustrates how covariates and regression parameters are translated into the observed categories. The two curves represent the distributions for two hypothetical observations, i and j. We can see that $\mathbf{x}_j^{\mathsf{T}}\boldsymbol{\beta} > \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}$. This has the effect of shifting the probability mass to the right, implying more of the mass is above the higher thresholds for observation j. This means that Y_j is more likely to be in the higher categories than Y_i.

Equation 8.1 is the basic probability statement for the model, which yields a likelihood that is simply the product of the binary logit models that switches between adjacent categories for each observation:

$$\begin{aligned}&\mathcal{L}(\{\beta_1, \dots, \beta_k\}, \{\tau_1, \dots, \tau_m\} | \mathbf{Y}, \mathbf{X}) \\ &\quad = \prod_{i=1}^{n} \prod_{m=1}^{M} \left[\Lambda(\tau_m - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}) - \Lambda(\tau_{m-1} - \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta})\right]^{\mathbb{1}_{im}},\end{aligned}$$

[1] Recall that $\Pr(V \leq v) = \int_{-\infty}^{v} f(r)dr$, which implies that $\Pr(u < V \leq w) = \int_{-\infty}^{w} f(r)dr - \int_{-\infty}^{u} f(r)dr = F(w) - F(u)$ where $f(\cdot)$ is the density function, and $F(\cdot)$ is the cumulative distribution function for the random variable V.

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press

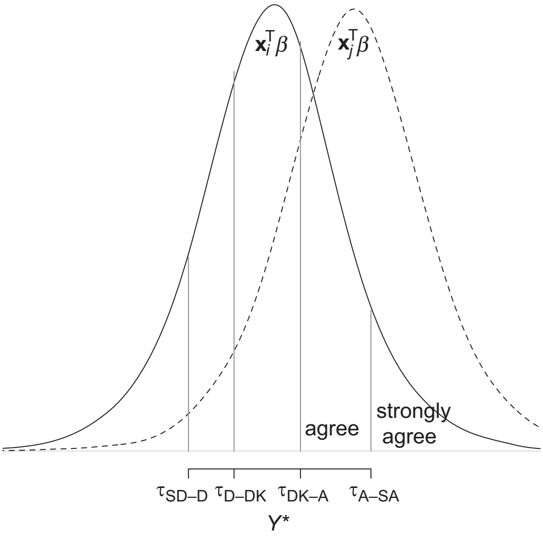

FIGURE 8.1 Visualizing how cutpoints divide the density of a continuous latent variable into discrete ordered categories. The curves represent Y^* for two observations, i and j.

where $\mathbb{1}_{im}$ is an indicator variable equal to 1 if and only if $y_i = m$ and 0 otherwise.

This translates easily to the log-likelihood as

$$\log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\tau}|\mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{n} \sum_{m=1}^{M} \mathbb{1}_{im} \log[\Lambda(\tau_m - \mathbf{x}_i^\top \boldsymbol{\beta}) - \Lambda(\tau_{m-1} - \mathbf{x}_i^\top \boldsymbol{\beta})]. \quad (8.2)$$

One additional mathematical consideration involves the question of how to anchor the scale of the latent variable or, put another way, whether to estimate an intercept. To see this, note that we can conceptualize the τ parameters as a series of intercepts, i.e., probabilities of being in each category when all independent variables are set to 0.0. If we estimate an intercept along with a threshold for each of the categories (i.e., $M-1$ cut points), then the model is not identified. In order to identify the model either the intercept or a cutpoint must be dropped. $\mathcal{R}$'s `polr` function (from the `MASS` library) omits the intercept. These differences are purely technical; they have no bearing on the model's fit to the data or the interpretation of the regression parameters.

As with the standard logit model, the ordered logit likelihood can be maximized with respect to the regression parameters and cutpoints. As usual we calculate the variance-covariance matrix from the inverse of the Hessian matrix.

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press

8.2.1 What about the Ordered *Probit* Model?

Just as in the binary case, the distinction between ordered logit and probit is driven by assumptions about the stochastic component of the model or, equivalently, the error process. In the ordered logit we assumed that the errors followed a standard logistic distribution. If we assume the errors follow the standard normal distribution, then we have an ordered probit model. Using the convention that Φ denotes the cumulative distribution function for the standard normal, we can express the log-likelihood for an ordered probit model as

$$\log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\tau} | \mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{n} \sum_{m=1}^{M} \mathbb{1}_{im} \log \left[\Phi(\tau_m - \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta}) - \Phi(\tau_{m-1} - \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta}) \right]. \tag{8.3}$$

Just as in the binary case, the choice between logit and probit is virtually always inconsequential.

8.2.2 Results and Interpretation

Results from ordered logit and probit models are frequently presented using the standard BUTON of coefficient and standard error estimates. This is unfortunate since the point estimates from ordered logit models are even less straightforward on their own than in the binary logit case. This is largely because these models have multiple categories for the dependent variable. Each category is itself modeled as a nonlinear function of covariates and parameters.

Scholars will also sometimes report odds ratios, although these remain difficult to explain in front of others (e.g., in job talks). For completeness, however, the change in the odds of $Y \leq m$ (versus greater than m) associated with a δ-unit change in covariate X_j equals $\exp(-\delta \hat{\beta}_j)$. Using tabular displays for interpreting things like odds ratios for ordered outcomes becomes quite complicated because each level of the dependent variable must be discussed separately.

Another way of looking at model implications is to consider marginal effects in the form of the partial derivative of the probability of any particular outcome value with respect to a particular covariate $j \in \{1, 2, \ldots, k\}$. This is given as

$$\begin{aligned} \left. \frac{\partial \Pr(Y = m)}{\partial x_j} \right|_{\bar{\mathbf{x}}} &= \frac{\partial \Lambda(\hat{\tau}_m - \bar{\mathbf{x}}^{\mathsf{T}} \hat{\boldsymbol{\beta}})}{\partial x_j} - \frac{\partial \Lambda(\hat{\tau}_{m-1} - \bar{\mathbf{x}}^{\mathsf{T}} \hat{\boldsymbol{\beta}})}{\partial x_j} \\ &= \hat{\boldsymbol{\beta}}_j \left[f_L(\hat{\tau}_m - \bar{\mathbf{x}}^{\mathsf{T}} \hat{\boldsymbol{\beta}}) - f_L(\hat{\tau}_{m-1} - \bar{\mathbf{x}}^{\mathsf{T}} \hat{\boldsymbol{\beta}}) \right]. \end{aligned} \tag{8.4}$$

A change in x_j shifts the density up or down the axis, while the positions of the cutpoints stay the same. For a positive $\hat{\beta}_j$ the "mass of probability" in the middle category first gets larger, then smaller as we increase x_j. Mathematically, this means that the change in the probability of observing a "middle category" response as a function of X_j, is not directly observable from the sign of $\hat{\beta}_j$.

Predicted probabilities and first differences are also straightforward to calculate, as we show below:

$$\Pr(\widehat{Y_i = m}|\mathbf{x}_i) = \Lambda(\hat{\tau}_m - \bar{\mathbf{x}}_i^\top \hat{\boldsymbol{\beta}}) - \Lambda(\hat{\tau}_{m-1} - \bar{\mathbf{x}}_i^\top \hat{\boldsymbol{\beta}}). \tag{8.5}$$

As always, the most informative presentations of results will not just rely on point estimates but also incorporate the full measure of uncertainty.

8.3 EXAMPLE: CATEGORICAL ASSESSMENTS OF POLITICAL IDEOLOGY

The ANES includes a variety of interesting questions, many of which are ordered, categorical variables. We use the pilot study from 2016 to illustrate the ordered logit model. The variables are delineated, briefly, in Table 8.1.

We fit an ordered logit model to respondents' ratings of Obama's conservatism as a function of how much one pays attention to politics, party identification, age, educational level, income, and race. Standard $\mathcal{R}$ results appear in Table 8.2, a BUTON.

What can we tell from these results? First, categories can be distinguished from each other. We see this by looking at the τ_m; the distances between the cutpoints are large relative to their standard errors. There is a significant and substantively large negative relationship between party identification and perceptions of Obama's conservatism. Party identification is coded from 0 (Strong Democrat) to 7 (Strong Republican), which implies that Republican identifiers are more likely to perceive Obama to be politically liberal than Democratic identifiers. Thus, $\exp(-1 * -0.39) = 1.47$, which shows that a respondent one "unit" more Republican is more likely to place Obama in category m or lower by a factor of 1.5. Interest in politics, income, and age

TABLE 8.1 *Selected variables and descriptors from 2016 ANES pilot study.*

Obama L-C	Barak Obama's perceived conservatism, 7-point scale (7 = very conservative)
age	respondent's age in years
education	respondent's education level, 7-point scale (7 = advanced degree)
income	respondent's family income (binned)
party ID	respondent's partisan allegiance, 7-point scale (7 = strong Republican)
follow politics	extent to which respondent follows politics, 4-point scale (1 = most of the time)
white	Did the respondent self-identify as white?

TABLE 8.2 *Ordered logit analysis of 2016 ANES data. What determined the liberal–conservative assessment of Barak Obama?*

	$\hat{\beta}$	$\sigma_{\hat{\beta}}$
follow politics	0.62	0.08
party ID	−0.39	0.03
age	−0.01	0.004
education	−0.03	0.04
income	−0.04	0.02
white	−0.10	0.14
cutpoints	$\hat{\tau}_m$	$\sigma_{\hat{\tau}_m}$
1|2	−1.58	0.34
2|3	−0.51	0.33
3|4	0.34	0.33
4|5	1.95	0.35
5|6	2.73	0.37
6|7	3.58	0.42
n	998	
$\log \mathcal{L}$	−1,400	
AIC	2,823	
BIC	2,882	

also appear to be strong predictors of respondents' perceptions of Obama's conservatism. But it is hard to be very precise about these relationships without some exponentiation, alas.

The logic of calculating first differences suggests that a plot may be useful. We examine the model's predicted probabilities of falling into each outcome category as over different levels of party identification while holding the other covariates at fixed levels. Specifically, we do this calculation for a 49-year-old white respondent with more than 12 years of schooling but no higher degree, with annual family income between $40,000–$49,000, and who claims to follow politics "most of the time."

Figure 8.2 illustrates these results, which are obtained using the $\mathcal{R}$ code displayed in the box for Example 8.1. We can then plot these predicted probabilities across the range of party ID, as displayed in Figure 8.2.

Figure 8.2 displays a common feature of models for ordered categorical data: the effect of a covariate on the probability of falling into the extreme categories is monotonic, but a covariate's effect on the probability of falling into intermediate categories is not. Figure 8.1 shows why this occurs: the covariates shift the location of each individual's latent Y_i^* but the

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press

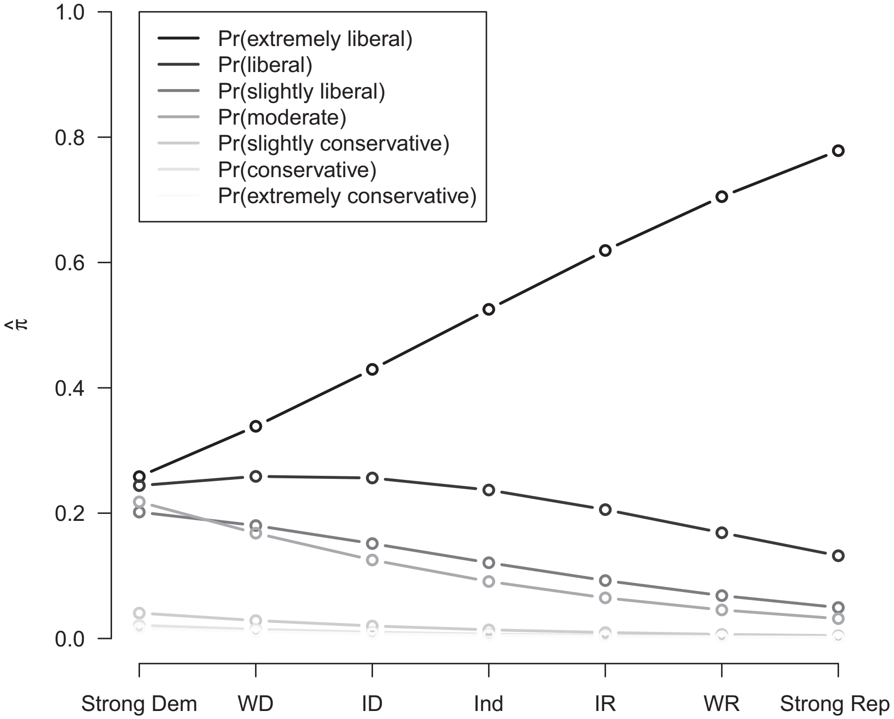

FIGURE 8.2 Party identification and perceptions of Obama's conservatism, calculated from the ordered logistic regression reported in Table 8.2.

cutpoints remain fixed. A change in a covariate that shifts the probability distribution to the right will unambiguously put more of the mass beyond the highest threshold and less of it below the lowest, but what happens in between depends on the magnitude of the shift and the point where we started.

Figure 8.2 echoes Figure 3.3, only without including any way of evaluating the uncertainty in point estimates. Including error bars or shaded regions for all predicted outcome categories here would make the figure hopelessly unreadable. We can build on the uncertainty estimates produced during model fitting, though we may be required to choose specific levels of the outcome variable to focus on for visual clarity. We do just that in Figure 8.3, presenting the differences between self-identified weak democrats and independents in finding Obama to be "moderate" and "liberal." As always in nonlinear models, we hold all other covariates at fixed values. Here we use their central tendencies. The densities reflect our estimation uncertainty around these quantities.

Code Example 8.1 produces these estimates and graphics.

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press

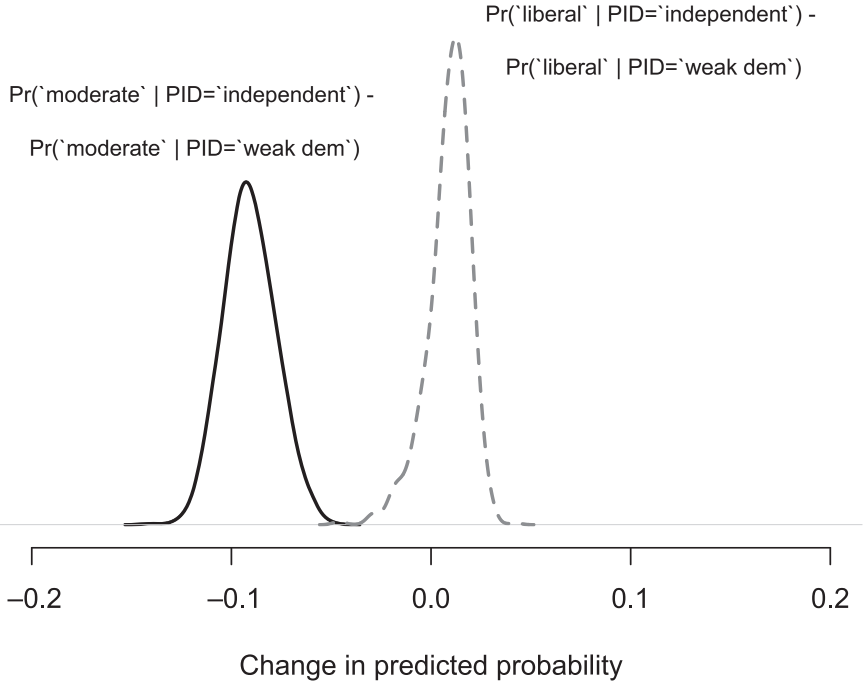

FIGURE 8.3 First differences: Comparing weak democrats against independents in their predicted probabilities of identifying Obama as "moderate" (solid) and "liberal" (broken), calculated from the ordered logistic regression reported in Table 8.2. Probability densities reflect the estimation uncertainties around these quantities.

8.4 PARALLEL REGRESSIONS

An alternative way of conceptualizing the ordered logit/probit model is as the constrained estimation of a system of logit/probit models. To see this, note that we can reexpress our ordered categorical variable, Y_i, as a series of binary variables, $\tilde{Y}_{im}$, such that $\tilde{y}_{im} = 1 \Leftrightarrow y_i \leq m$ for some category, m. We might then be interested in fitting the system of equations

$$\begin{aligned}
\Pr(Y_i \leq 1) &= \text{logit}^{-1}\left(\tau_1 + \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}_1\right)\\
&\vdots\\
\Pr(Y_i \leq M-1) &= \text{logit}^{-1}\left(\tau_{M-1} + \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}_{M-1}\right)\\
\Pr(Y_i = M) &= 1 - \text{logit}^{-1}\left(\tau_{M-1} - \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}_{M-1}\right).
\end{aligned}$$

The system of logit models fit to each of these $\tilde{Y}_m$ is called the *cumulative logit* model.

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press

$\mathcal{R}$ Code Example 8.1 *Ordered logit first differences*

```r
library(MASS)
# requires anes.sub data from repository.
attach(anes.sub)
polr.out <- polr(as.ordered(Obama.LR)~ follow.politics + pid +
                 age + education +income + white,
                 data=anes.sub, method="logistic", Hess=T)
beta <- coef(polr.out)
tau <- polr.out$zeta
#create predicted probabilities
X <- cbind(median(follow.politics), #scenario of interest
           min(pid):max(pid),       #across range of party ID
           median(age), median(education),
           median(income), median(white)
           )
p1 <- plogis(tau[1] - X %*% beta)
p2 <- plogis(tau[2] - X %*% beta) - plogis(tau[1] - X %*% beta)
p3 <- plogis(tau[3] - X %*% beta) - plogis(tau[2] - X %*% beta)
p4 <- plogis(tau[4] - X %*% beta) - plogis(tau[3] - X %*% beta)
p5 <- plogis(tau[5] - X %*% beta) - plogis(tau[4] - X %*% beta)
p6 <- plogis(tau[6] - X %*% beta) - plogis(tau[5] - X %*% beta)
p7 <- 1.0 - plogis(tau[6] - X %*% beta)
```

Note that in this general specification each equation has its own set of slope parameters, $\boldsymbol{\beta}_m$. The ordered logit and probit models estimate these equations simultaneously, with the constraint that slope parameters are equal across equations, i.e., $\boldsymbol{\beta}_m = \boldsymbol{\beta} \ \forall \ m$. The intercepts ($\tau_m$) vary across equations, as they must if we are to discriminate between categories. This assumption of common slope parameters across levels of the response variable goes by the name "parallel regressions" or, in the logit case, "proportional odds." The phrase parallel regressions describes how the impact of the covariates may shift the predicted probability curves to the right or the left but does not change the basic slope of these curves across any two categories. This means that the partial derivative for the probability that Y is in any category with respect to the covariate information should be equal:

$$\frac{\partial \Pr(y \leq m \mid x)}{\partial x} = \frac{\partial \Pr(y \leq m-1 \mid x)}{\partial x} = \frac{\partial \Pr(y \leq m-2 \mid x)}{\partial x} = \cdots$$

For better or worse, the parallel regressions assumption is rarely examined in practice, so the extent to which violations are substantively important is not well-explored. But it is possible for a covariate to increase the probability of being in category 2 relative to 1 and yet have a null or even negative relationship at other levels. An inappropriate constraint could result in missing this relationship. The parallel regressions assumption is easier to satisfy when

$\mathcal{R}$ Code Example 8.2 ***Interpreting ordered logit***

```
#scenarios
X.wd <- cbind(median(follow.politics), 1, #PID=1 weak democrat
            median(age), median(education), median(income), TRUE)
X.ind <- cbind(median(follow.politics), 3, #PID=3 independent
            median(age), median(education), median(income), TRUE)

#coefficient vectors. Note inclusion of cutpoints
draws<-mvrnorm(1000, c(coef(polr.out),polr.out$zeta),
      solve(polr.out$Hessian))
B<-draws[,1:length(coef(polr.out))]
Taus<-draws[,(length(coef(polr.out))+1):ncol(draws)]

#predicted probabilities
pi.lib.wd<- plogis(Taus[,2] - B%*%t(X.wd)) - plogis(Taus[,1] - B%*%t(X.wd))
pi.lib.ind <- plogis(Taus[,2] - B%*%t(X.ind)) - plogis(Taus[,1] - B%*%t(X.ind))
pi.mod.wd<- plogis(Taus[,4] - B%*%t(X.wd)) - plogis(Taus[,3] - B%*%t(X.wd))
pi.mod.ind <- plogis(Taus[,4] - B%*%t(X.ind)) - plogis(Taus[,3] - B%*%t(X.ind))

#differences
fd.lib<- pi.lib.ind - pi.lib.wd
fd.mod<- pi.mod.ind - pi.mod.wd

#plotting
plot(density(fd.mod, adjust=1.5), xlim=c(-0.2,0.2),ylim=c(0,50),
    xlab="Change in predicted probability", bty="n", col=1,
    yaxt="n", lwd=2, main="", ylab="")
lines(density(fd.lib, adjust=1.5), col=grey(0.5), lwd=2, lty=2)
text(x=0.11, y=42, labels="Pr(‘liberal‘ | PID=‘independent‘) -
      \n Pr(‘liberal‘ | PID=‘weak dem‘)",cex=.8)
text(x=-.12, y=35, labels="Pr(‘moderate‘ | PID=‘independent‘) -
      \n Pr(‘moderate‘ | PID=‘weak dem‘)",cex=.8)
detach(anes.sub)
```

there are just a few categories but becomes more difficult to meet as the number of categories grows.

The parallel regressions assumption can be tested in several ways. One way is to fit $m-1$ binary regressions after expanding the ordinal dependent variable into $m-1$ binary $\tilde{Y}_m$. With these regression results in hand, the equality of $\hat{\boldsymbol{\beta}}_m$ can be examined with standard Wald-type tests or simply visualized. A somewhat easier test to execute is to compare the ordered logit/probit model with a cumulative logit or a multinominal logit/probit. We provide an example of the former below. Multinomial models are discussed in Chapter 9.

8.4.1 Example: Genocide Severity

Krain (2005) examined the effectiveness of military intervention in slowing or stopping killing during genocides. His results suggested that interventions that directly challenged the perpetrator or specifically aided the targets were the only efficacious types of military responses. Krain's dependent variable, `magnitud`,

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press

TABLE 8.3 *The distribution of observations over Krain's 11-point scale of genocide magnitude.*

Category	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
Frequency	45	27	20	21	16	25	34	53	19	7	6

is an ordered categorical variable describing the magnitude of genocide, where the category coded 0 implies the lowest severity of genocide, and 5 is used to flag the highest severity levels. Table 8.3 displays the frequency of cases in each of the eleven categories.

In his analysis Krain treats the outcome variable as ordered categories, which is, strictly speaking, correct. But when there is a large number of categories it is common to use continuous regression models such as least squares, especially if there is roughly the same number of observations in each category. We revisit this decision by comparing Krain's ordered logit specification to a simpler OLS model. We also interrogate the parallel regressions assumption.

Table 8.4 displays the BUTON for an OLS and ordered logit specification. The results for the ordered logit model are identical to those presented by Krain in his article (model 1, in table 3, pg. 379). Note that the likelihood ratio, AIC, and BIC all indicate that the simpler OLS model with fewer parameters is a better in-sample fit than the ordered logit. Looking at the estimated cutpoints, we also see that they are all very close together relative to their standard errors, particularly below category 3.5. This indicates that the model has a hard time distinguishing between adjacent categories in these data. With eleven categories the assumption of a continuous response appears reasonable.

8.4.2 Parallel Regressions and Genocide Severity

We can use two different graphical heuristics to examine the parallel regressions assumption in the Krain genocide severity example.

Figure 8.4 plots the means of the regressors at different levels of the response variable.[2] If the included regressors are able to differentiate one category from another, then we should expect to see a strong trend across the levels of the dependent variable. If the parallel regressions assumption holds, then this trend should be linear, and the conditional means should line up neatly along the broken trend line. In Krain's data, this does not appear to hold. Covariates that appear as "significant" in Table 8.4 – duration of genocide (`genyr`) and state failure (`stfl`) – appear to have nonlinear relationships with the levels of Y.

A second strategy is to fit the cumulative logit regressions without any constraint and examine the stability of the coefficients across the levels of the response variable. A response with eleven categories implies ten simultaneous

[2] Note that we have rescaled the response variable to be in integer increments, e.g., $\{0, 1, \ldots, 10\}$, rather than increments of 0.5.

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press

TABLE 8.4 *The correlates of genocide severity, a reanalysis of Krain (2005).*

	OLS	Ordered logit
Intervention	0.23 (0.22)	0.13 (0.30)
Contiguity	0.39 (0.22)	0.57 (0.30)
Genocide severity	0.42 (0.06)	0.72 (0.10)
Genocide duration	−0.05 (0.02)	−0.07 (0.03)
State failures	0.44 (0.19)	0.54 (0.27)
Regime type	−0.01 (0.02)	−0.01 (0.02)
Ethnic fractionalization	0.78 (0.38)	0.49 (0.51)
Economic marginalization	0.0004 (0.001)	−0.0004 (0.002)
Cold War	0.03 (0.24)	−0.27 (0.36)
Cutpoints		$\hat{\tau}_m$
Constant	0.50 (0.41)	
0|0.5		−0.33 (0.57)
0.5|1		0.34 (0.56)
1|1.5		0.80 (0.56)
1.5|2		1.25 (0.57)
2|2.5		1.60 (0.58)
2.5|3		2.12 (0.59)
3|3.5		2.84 (0.60)
3.5|4		4.35 (0.62)
4|4.5		5.45 (0.65)
4.5|5		6.29 (0.72)
n	273	273
$\log \mathcal{L}$	−450	−565
AIC	920	1,167
BIC	960	1,236

equations using the cumulative logit conceptualization. Table 8.5 displays how the dependent variables, the $\tilde{Y}_m$, are constructed.

Figure 8.5 displays the results. The vertical axis is the $\hat{\beta}_k$. If parallel regression holds, then these estimates should be approximately equal. The plots are scaled so that the vertical axis distance roughly covers ±2 standard errors from the maximum and minimum coefficient estimates. Again, we see unusual and unstable patterns in the regression coefficients, even those like state failure (`stfl`) that appeared as significant in the ordered logit fit.

The results here are consistent with the imprecisely estimated threshold parameters reported in Table 8.4, implying that several of the categories are difficult to distinguish from one another. We might fit the OLS model, as above, or consider combining some categories together, especially where thresholds are imprecisely estimated. But either way this application violates the parallel regressions assumption.

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press

TABLE 8.5 *Coding Krain's response variable for the cumulative logit.*

		$\tilde{Y}_{im} = 1$	$\tilde{Y}_{im} = 0$
eq. 1	$\tilde{Y}_{i,0}$	0	0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5
eq. 2	$\tilde{Y}_{i,0.5}$	0,0.5	1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5
eq. 3	$\tilde{Y}_{i,1}$	0,0.5,1	1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5
eq. 4	$\tilde{Y}_{i,1.5}$	0,0.5,1, 1.5	2, 2.5, 3, 3.5, 4, 4.5, 5
eq. 5	$\tilde{Y}_{i,2}$	0,0.5,1, 1.5, 2	2.5, 3, 3.5, 4, 4.5, 5
eq. 6	$\tilde{Y}_{i,2.5}$	0,0.5,1, 1.5, 2, 2.5	3, 3.5, 4, 4.5, 5
eq. 7	$\tilde{Y}_{i,3}$	0,0.5,1, 1.5, 2, 2.5, 3	3.5, 4, 4.5, 5
eq. 8	$\tilde{Y}_{i,3.5}$	0,0.5,1, 1.5, 2, 2.5, 3, 3.5	4, 4.5, 5
eq. 9	$\tilde{Y}_{i,4}$	0,0.5,1, 1.5, 2, 2.5, 3, 3.5,4	4.5, 5
eq. 10	$\tilde{Y}_{i,4.5}$	0,0.5,1, 1.5, 2, 2.5, 3, 3.5,4, 4.5	5

8.4.3 Extensions

Several extensions to the ordered logit/probit model have been proposed to partially relax the parallel regressions assumptions and allow for other nuances like unequal variances across units. These models, going by names like *partial proportional odds* and *generalized ordered logit* have not seen exensive use in the social science for several reasons. First, an even more general and flexible approach, the multinomial model taken up in the next chapter, is widely understood. The main cost of the multinomial model is that it estimates many more parameters than an ordered alternative. With datasets growing bigger and computers becoming ever faster, the relative cost of the multinomial alternative is falling fast. Second, most of these models require theoretical or other reasons for contraining some predictors to respect the parallel regressions assumption, while others do not. Or, as with the heterokedastic probit model, some covariates must be used to model scale parameters, often with little improvement to model fit. Rarely do we have theories at this level of precision. Third, interpretation of these hybrid ordered models is more complicated, including the fact that it is possible for certain models to generate negative predicted probabilities (McCullagh and Nelder, 1989).

8.5 ORDERED CATEGORICAL VARIABLES AS REGRESSORS

Computer programs, including $\mathcal{R}$, do not care on which side of the regression equation categorical variables appear. Analysts, however, will often model an ordered variable using ordered logit or probit when it is an outcome, yet treat the same variable as if it were continuous when including it as a regressor. We did just that for both the model reported in Table 8.2 and those in Table 8.4. In the former we treated the respondents' self-reported party identification as an interval-level variable while modeling the same respondent's perception of

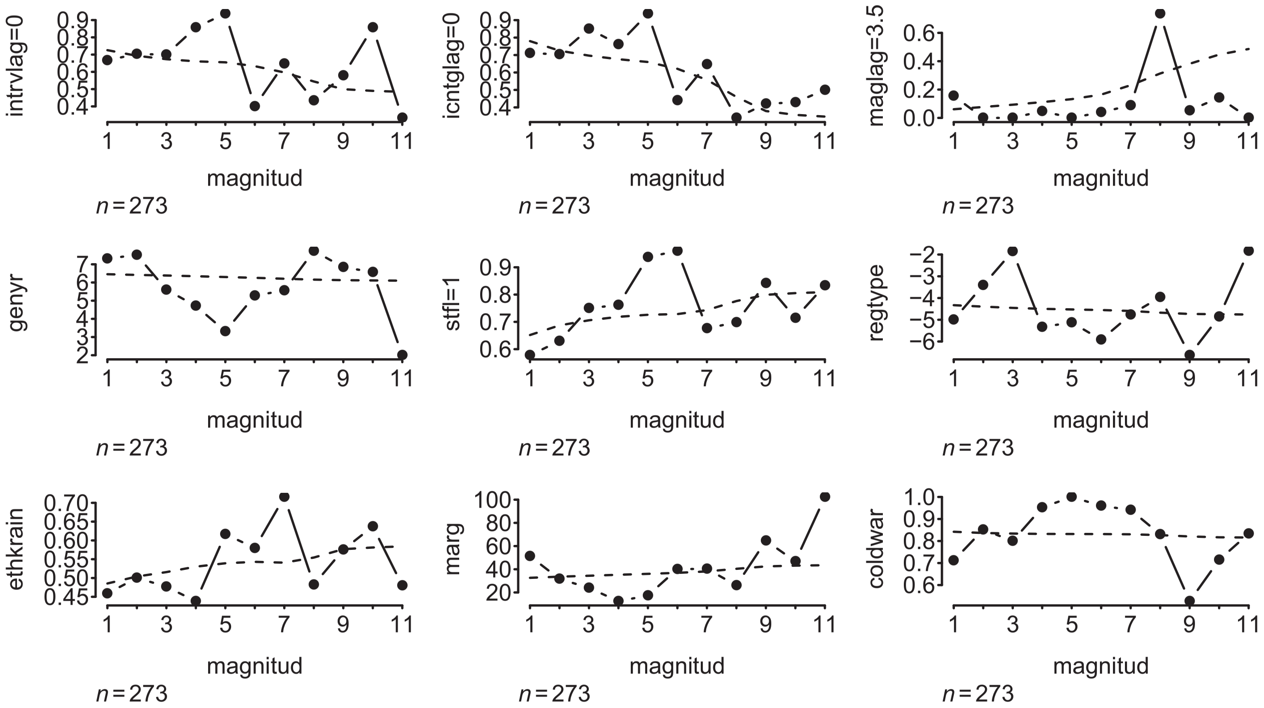

FIGURE 8.4 *Plot of the conditional means of the regressors at different levels of the response variable,* `magnitud`*. If the parallel regressions assumption holds, the means will show a strong trend across values of the response variable and line up neatly. The broken line is the loess smoother describing this trend.*

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press

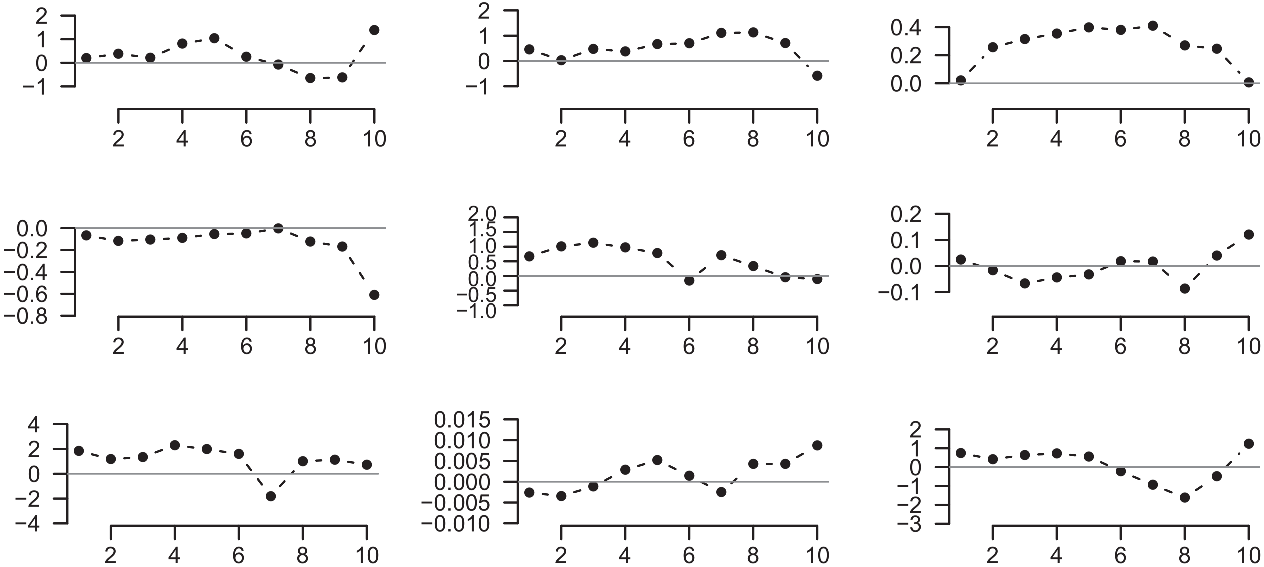

FIGURE 8.5 *Plot of the estimated regression coefficients from* M − 1 *regressions on* $\tilde{Y}_m$*, the indicators that* $Y \leq m$*. If the parallel regressions assumption holds, the estimated coefficients should be stable across levels of* $\tilde{Y}$*.*

Obama as categorical (but ordered). In the latter we reproduced Krain's model, in which he included a one-year lagged version of the genocide magnitude variable as a predictor, again treating it as continuous even though that same variable was treated as ordered categorical on the left side of the regression.

The difference between an ordered categorical variable and one that is "continuous" can be subtle. We do not wade into complicated issues of measurement and scaling here. Rather, we take an applied (some might say naïve) approach. Whether a variable is better treated as an ordered categorical variable or an interval-scaled measure will depend on the quantities of interest, coarseness or fineness of the coding scheme, available degrees of freedom, and, ultimately, predictive power. Deciding on the best modeling approach to take will benefit from the tools and heuristics described in Chapter 5.

8.5.1 An Example: Child Mortality and Democracy

Suppose we are interested in the relationship between democracy and infant mortality, a common measure of human well-being. We gather data on child mortality rates for a cross-section of countries in the year 2000. The measure we use for democracy here is the "constraints on the executive" (XCONST) variable from Polity IV (Marshall et al., 2016), an ordinal scale ranging from few restrictions to many. In Figure 8.6 we display the distribution of (log) mortality for different levels of XCONST. It appears that child mortality drops significantly when the executive is highly constrained (categories 6 and 7).

We first follow common practice and fit an OLS regression model treating XCONST as if it were simply an interval-scaled variable. We obtain the results presented in the first column of Table 8.6, in which there appears to be a substantively large and "significant" negative association between (log) child mortality and XCONST.

In the second column we treat XCONST as a purely categorical variable; all XCONST coefficients represent comparisons against XCONST=1. In this case we see that *only* category 7 (most constrained) of XCONST shows a substantively large negative relationship with child mortality relative to category 1.[3] Clearly, conclusions based on the first model were somewhat misleading. Consistent with Figure 8.6, it does not appear that the negative relationship holds equally and constantly across all levels of XCONST.

One way of thinking about what we have done so far is that in the first model, we privileged ordering over the categories, whereas in the second model, we gave greater analytical weight to the categories (particularly the comparison against category 1) and ignored the ordinal information. What

[3] But note that categories 6 and 7 can be distinguished from 2–5.

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press

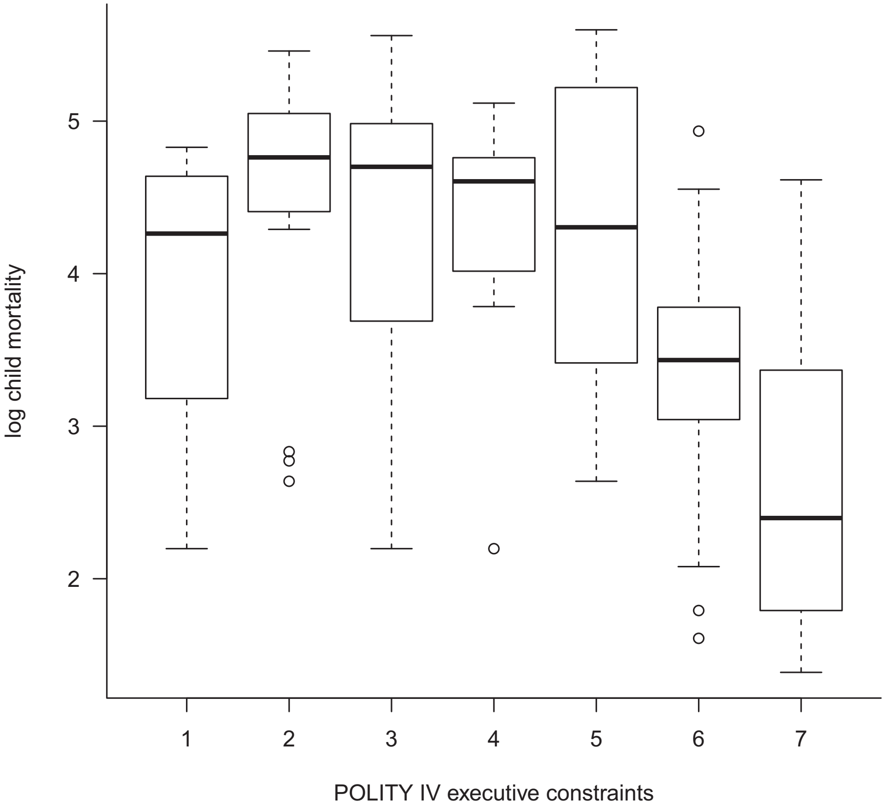

FIGURE 8.6 Child mortality and constraints on the executive.

happens if we try to balance both, respecting the categorical, ordinal nature of the variable? In $\mathcal{R}$ we simply declare the variable to be of class `ordered`. In this case we also choose to use "successive differences" contrasts, reporting results in the third column of Table 8.6. In this model, coefficient estimates represent the "effect" of going from one level of a variable to the next, e.g., the expected change in child mortality associated with going from XCONST=2 to XCONST=3. In fact, the coefficient estimates in the last column are simply the differences of coefficients for adjacent categories from the second model. Note that using this encoding of XCONST does not in any way alter the model fit, since we are not bringing any new information to bear. This encoding simply alters how we interpret the coefficients from the model. On this basis we see that moving from 5 to 6 and 6 to 7 both reduce infant mortality.

There are several other things worth noting here. First, the two models treating XCONST as a categorical variable fit the data equally well. Second, by the AIC and (penalized) R^2 criteria, the second and third models

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press

TABLE 8.6 *OLS regression of (log) infant mortality on different encodings of executive constraints (XCONST).*

	Continuous	**Categorical**	**Ordered**
XCONST	−0.30		
	(0.04)		
XCONST=2		0.59	
		(0.37)	
XCONST=3		0.40	
		(0.35)	
XCONST=4		0.30	
		(0.46)	
XCONST=5		0.37	
		(0.35)	
XCONST=6		−0.58	
		(0.37)	
XCONST=7		−1.29	
		(0.32)	
XCONST$_{2-1}$			0.59
			(0.37)
XCONST$_{3-2}$			−0.19
			(0.30)
XCONST$_{4-3}$			−0.09
			(0.41)
XCONST$_{4-4}$			0.07
			(0.41)
XCONST$_{6-5}$			−0.95
			(0.31)
XCONST$_{7-6}$			−0.72
			(0.28)
Constant	5.07	3.91	3.88
	(0.21)	(0.29)	(0.09)
n	140	140	140
Adjusted R^2	0.26	0.37	0.37
AIC	411	393	393
BIC	420	417	417

using categorical treatments of XCONST fit the data better in-sample than the one treating XCONST as continuous. On a BIC basis, the decision is less clear. Using a five-fold cross validation heuristic, we find that the first model returns an MSE of 1.07; the other two return MSEs of 0.926, an improvement of almost 14% in out-of-sample predictive performance. Better modeling of the categorical data on the right hand side of the regression has helped us better understand the relationship between the variables of interest here.

8.6 FURTHER READING

Applications

Walter (2017) uses ordered logit to analyze European Social Survey data reporting people's perceptions of labor market risk. Kriner and Shen (2016) analyze an experiment on the presence of the draft on support of military action.

Past Work

McKelvey and Zaviona (1975) present an early introduction to the ordered probit in the social sciences. Fullerton (2009) develops a typology of the various types of logit-based models for ordered categorical outcome variables.

For a nuanced exploration of an ordered categorical variable commonly treated as interval-scaled (Polity scores), see Treier and Jackman (2008).

Advanced Study

On the "generalized ordered logit" that constrains some regression parameters to be equal across categories while allowing others to vary, see Peterson and Harrell (1990); Williams (2016).

Software Notes

See Venables and Ripley (2002) for more extended discussion of contrasts in $\mathcal{R}$ and how they are constructed. Successive differences contrasts are part of the `MASS` library, as is the `polr` function. The `ordinal` package (Christensen, 2015) implements a variety of models for ordinal data, including partial proportional odds and heterokedastic ordered logit. The `oglmx` package (Carroll, 2017) implements heteroskedastic ordered logit and probit models.

https://doi.org/10.1017/9781316888544.011 Published online by Cambridge University Press