15.5 Ordered and unordered categorical regression

Logistic and probit regression can be extended to multiple categories, which can be ordered or unordered.

The ordered multinomial logit model

Consider a categorical outcome y that can take on the values 1, 2, ... , K. The ordered logistic model can be written in two equivalent ways. First, we express it as a series of logistic regressions:

$$\begin{aligned}
\Pr(y > 1) &= \text{logit}^{-1}(X\beta), \\
\Pr(y > 2) &= \text{logit}^{-1}(X\beta - c_2), \\
\Pr(y > 3) &= \text{logit}^{-1}(X\beta - c_3), \\
&\cdots \\
\Pr(y > K-1) &= \text{logit}^{-1}(X\beta - c_{K-1}).
\end{aligned} \tag{15.6}$$

The parameters c_k (called thresholds or cutpoints), are constrained to increase: $0 = c1 < c2 < \cdots < c_{K-1}$, because the probabilities in (15.6) are strictly decreasing (assuming that all K outcomes have nonzero probabilities of occurring). Since c_1 is defined to be 0, the model with K categories has K−2 free parameters c_k in addition to β. This makes sense since K =2 for the usual logistic regression, for which only β needs to be estimated.

The cutpoints c_2,..., c_{K-1} can be estimated using Bayesian or likelihood approaches, simultaneously with the coefficients β

The expressions in (15.6) can be subtracted to get the probabilities of individual outcomes:

$$\begin{aligned}
\Pr(y = k) &= \Pr(y > k-1) - \Pr(y > k) \\
&= \text{logit}^{-1}(X\beta - c_{k-1}) - \text{logit}^{-1}(X\beta - c_k).
\end{aligned}$$

Latent variable interpretation with cutpoints

The ordered categorical model is easiest to understand by generalizing the latent variable formulation (13.5) to K categories:

$$y_i = \begin{cases} 1 & \text{if } z_i < 0 \\ 2 & \text{if } z_i \in (0, c_2) \\ 3 & \text{if } z_i \in (c_2, c_3) \\ & \dots \\ K-1 & \text{if } z_i \in (c_{K-2}, c_{K-1}) \\ K & \text{if } z_i > c_{K-1}, \end{cases}$$

$$z_i = X_i\beta + \epsilon_i, \qquad (15.7)$$

with independent errors εi that have the logistic distribution.

Figure 15.5 illustrates the latent variable model and shows how the distance between any two adjacent cutpoints c_{k-1}, c_k affects the probability that y = k. The lowest and highest categories in (15.7) are unbounded, so if the linear predictor X β is high enough, y will almost certainly take on the highest possible value, and if X β is low enough, y will almost certainly equal the lowest possible value.

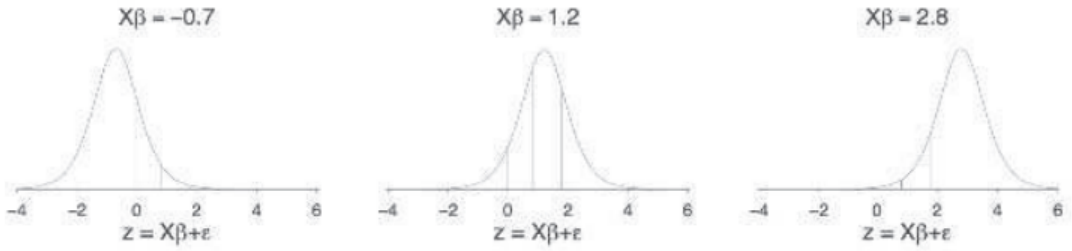

Figure 15.5 *Illustration of cutpoints in an ordered categorical logistic model. In this example, there are $K = 4$ categories and the cutpoints are $c_1 = 0, c_2 = 0.8, c_3 = 1.8$. The three graphs illustrate the distribution of the latent outcome z corresponding to three different values of the linear predictor, $X\beta$. For each, the cutpoints show where the outcome y will equal 1, 2, 3, or 4.*

Alternative approaches to modeling ordered categorical data

Ordered categorical data can be modeled in several ways, including:

- Ordered logit model with K−1 cutpoint parameters.
- The same model in probit form.
- Simple linear regression (possibly preceded by a simple transformation of the outcome values). This can be a good idea if the number of categories is large and if they can be considered equally spaced. This presupposes that a reasonable range of the categories is actually used. For example, if ratings are potentially on a 1 to 10 scale but in practice always equal 9 or 10, then a linear model probably will not work well.
- Nested logistic regressions—for example, a logistic regression model for y = 1 versus y = 2,..., K; then, if y ≥ 2, a logistic regression for y = 2 versus y = 3,..., K; and so on up to a model, if y ≥ K − 1 for y = K −1 versus y = K. Separate logistic (or probit) regressions have

the advantage of more flexibility in fitting data but the disadvantage of losing the simple latent-variable interpretation of the cutpoint model we have described.

- Finally, robit regression, is a competitor to logistic regression that accounts for occasional aberrant data.

Unordered categorical regression

Example, households with unsafe wells had the option to switch to safer wells. But the actual alternatives are more complicated and can be summarized as:

(0) do nothing,

(1) switch to an existing private well,

(2) switch to an existing community well,

(3) install a new well yourself.

If these are coded as 0, 1, 2, 3, then we can model $\Pr(y \geq 1)$, $\Pr(y \geq 2 \mid y \geq 1)$, $\Pr(y = 3 \mid y \geq 2)$. Although the four options could be considered to be ordered in some way, it does not make sense to apply the ordered multinomial logit or probit model, since different factors likely influence the three different decisions.

Rather, it makes more sense to fit separate logit (or probit) models to each of the three components of the decision: (a) Do you switch or do nothing? (b) If you switch, do you switch to an existing well or build a new well yourself? (c) If you switch to an existing well, is it a private or community well?

15.6 Robust regression using the t model

The t distribution instead of the normal

When a regression model can have occasional very large errors, it is generally more appropriate to use a t distribution rather than a normal distribution for the errors.

Regressions estimated using the t model are said to be robust in that the coefficient estimates are less influenced by individual outlying data points.

Robit instead of logit or probit

Logistic regression can run into problems with outliers. Outliers are usually thought of as extreme observations, but in the context of discrete data, an “outlier” is more of an unexpected observation.

Logistic regression can be conveniently "robustified" by generalizing the latent-data formulation:

$$y_i = \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0, \end{cases}$$
$$z_i = X_i\beta + \epsilon_i,$$

to give the latent errors a t distribution:

$$\epsilon_i \sim t_\nu\left(0, \frac{\nu-2}{\nu}\right), \qquad (15.12)$$

with the degrees-of-freedom parameter $\nu > 2$ estimated from the data and the t distribution scaled so that its standard deviation equals 1.

The t model for the ϵ_i's allows the occasional unexpected prediction—a positive value of z for a highly negative value of the linear predictor $X\beta$, or vice versa.

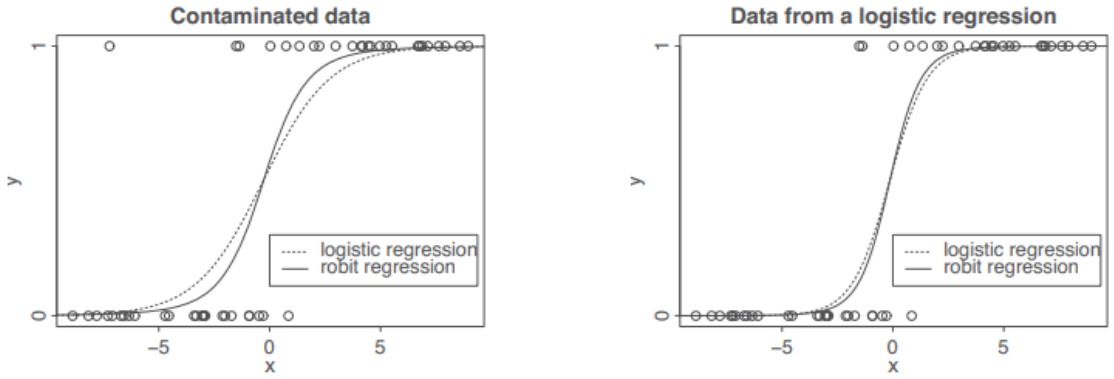

Figure 15.7 *Hypothetical data to be fitted using logistic regression: (a) a dataset with an "outlier" (the unexpected $y = 1$ value near the upper left); (b) data simulated from a logistic regression model, with no outliers. In each plot, the dotted and solid lines show the fitted logit and robit regressions, respectively. In each case, the robit line is steeper—especially for the contaminated data—because it effectively downweights the influence of points that do not appear to fit the model.*

15.7 Constructive choice models

A completely different approach is to model a decision outcome as a balancing of goals or utilities.

We demonstrate this idea using the example of well switching in Bangladesh. To set up a choice model, we must specify a value function, which represents the strength of preference for one

decision over the other—in this case, the preference for switching as compared to not switching. The value function is scaled so that zero represents indifference, positive values correspond to a preference for switching, and negative values result in not switching.

Logistic or probit regression as a choice model in one dimension

Now let us think about switching from first principles as a decision problem. For household i, define

- a_i: the benefit of switching from an unsafe to a safe well,
- $b_i + c_i x_i$: the cost of switching to a new well a distance xi away.

Logit model

Under the utility model, household i will switch if $a_i > b_i + c_i x_i$. However, we do not have direct measurements of the ai's, bi's, and ci's. All we can learn from the data is the probability of switching as a function of x_i; that is,

$$\Pr(\text{switch}) = \Pr(y_i = 1) = \Pr(a_i > b_i + c_i x_i), \qquad (15.13)$$

treating ai, bi, ci as random variables whose distribution is determined by the (unknown) values of these parameters in the population.

Expression (15.13) can be written as,

$$\Pr(y_i = 1) = \Pr\left(\frac{a_i - b_i}{c_i} > x_i\right),$$

a re-expression that is useful in that it puts all the random variables in the same place and reveals that the population relation between y and x depends on the distribution of (a − b)/c in the population.

For convenience, label di = (ai − bi)/ci: the net benefit of switching to a neighboring well, divided by the cost per distance traveled to a new well. If di has a logistic distribution in the population, and if d is independent of x, then Pr(y = 1) will have the form of a logistic regression on x.

If di has a logistic distribution with center μ and scale σ, then $d_i = \mu + \sigma \varepsilon_i$, where ε_i has the unit logistic density. Then

$$\Pr(\text{switch}) = \Pr(d_i > x) = \Pr\left(\frac{d_i - \mu}{\sigma} > \frac{x - \mu}{\sigma}\right)$$
$$= \text{logit}^{-1}\left(\frac{\mu - x}{\sigma}\right) = \text{logit}^{-1}\left(\frac{\mu}{\sigma} - \frac{1}{\sigma}x\right),$$

which is simply a logistic regression with coefficients μ/σ and −1/σ.

Probit model

A similar model is obtained by starting with a normal distribution for the utility parameter: d ∼ N(μ, σ2). In this case,

$$\Pr(\text{switch}) = \Pr(d_i > x) = \Pr\left(\frac{d_i - \mu}{\sigma} > \frac{x - \mu}{\sigma}\right)$$
$$= \Phi\left(\frac{\mu - x}{\sigma}\right) = \Phi\left(\frac{\mu}{\sigma} - \frac{1}{\sigma}x\right),$$

which is simply a probit regression.

Choice models, discrete data regressions, and latent data

Choice models are defined at the level of the individual, as we can see in the well switching example, where each household i has, along with its own data Xi, yi, its own parameters ai, bi, ci that determine its utility function and thus its decision of whether to switch.

Logistic or probit regression as a choice model in multiple dimensions

We can extend the well-switching model to multiple dimensions by considering the arsenic level of the current well as a factor in the decision.

- $a_i * (As)_i$ = the benefit of switching from an unsafe well with arsenic level As_i to a safe well.
- $b_i + c_i x_i$ = the cost of switching to a new well a distance xi away.

Household i should then switch if $a_i * (As)_i > b_i + c_i x_i$ —the decision thus depends on the household's arsenic level (As)i, its distance x_i to the nearest well, and its utility parameters a_i, b_i, c_i. However, a_i, b_i, c_i are not directly observable—all we see are the decisions (yi = 0 or 1) for households, given their arsenic levels As_i and distances xi to the nearest safe well.

Certain distributions of (a, b, c) in the population reduce to the fitted logistic regression, for example, if ai and ci are constants and bi/ai has a logistic distribution that is independent of (As)i and x_i.

Insights from decision models

A choice model can give us some insight even if we do not formally fit it. For example, in fitting logistic regressions, we found that distance worked well as a linear predictor, whereas arsenic level fit better on the logarithmic scale. A simple utility analysis would suggest that both these factors should come in linearly, and the transformation for arsenic suggests that people are (incorrectly) perceiving the risks on a logarithmic scale—seeing the difference between 4 to 8, say, as no worse than the difference between 1 and 2.

15.8 Going beyond generalized linear models

Extending existing models

The usual linear regression model assumes a constant error standard deviation, σ; this assumption is called homoscedasticity. We can also allow the residual standard deviation to vary (heteroscedasticity) and build a model for the error variance, for example, as $y_i \sim N(a + bx_i, e^{c+dx_i})$.

Latent-data models

Another way to model mixed data is through latent data, for example, positing an "underlying" income level z_i—the income that person i would have if he or she were employed—that is observed only if $y_i > 0$. Tobit regression is one such model that is popular in econometrics.