Models for Binary Outcomes and Proportions

Simon Jackman

April 9, 2007

1 Binary Outcomes

This is probably the simplest form of a limited dependent variable (LDV) setup. In the data available for analysis an outcome or attribute is present or it isn't, and this dichotomy is exhaustive. This typically occurs because of fairly natural limits to data collection and our observational powers. For example, someone either votes or doesn't. Someone is employed or isn't. It isn't easy to collect data on, say, the *enthusiasm* with which someone voted, though it might be worthwhile to know that if we were interested in making a prediction about the *probability* of a similar person voting in future elections. Someone may be "on the cusp" of unemployment, but our data generation mechanism is unable to provide that type of information, because like the voting example, the concept is rather vague, difficult and probably very expensive to operationalize. Someone owns a house or doesn't. A pair of countries (a dyad) is at war, or isn't. And so on.

For the purposes of doing quantitative analysis (and without loss of generality) we conventionally code the presence of the outcome or attribute of interest with as $y_i = 1$, and its absence as $y_i = 0$, in a sample of n observations, indexed by $i = 1, \ldots, n$.

We make no additional (or fewer) assumptions about independent variables than for LS regression.

Proportions data come from aggregating these unit-level binary outcomes or attributes into groups; the first applications of probit analysis come from bioassay, modeling the proportion of moths dying in response to a varying levels of exposure to a toxin (Bliss 1935). It will be useful to consider binary responses and proportions data simultaneously; arguments as to the inappropriateness of LS regression and alternative modeling and estimation strategies are similar in both contexts.

2 Why Not Run a Regression? -- the "Linear Probability Model"

If the only thing different here is the peculiar nature of y_i then why not run a regression? There are two points to be made here. When the data we have are aggregations of unit-level binary outcomes the recommends procedure is indeed to run a regression on the data, appropriately transformed (see below). But when **y** consists not of *proportions* of ones and zeros, but the actual ones and zeros *themselves* running a LS regression suffers some serious drawbacks.

Nonetheless, applying LS in this context has a long and venerable history in applied econometrics, and the resulting model is referred to as the "linear probability model" (LPM).

The basic idea here is to proceed with a LS regression,

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u}, \tag{1}$$

and to interpret the resulting conditional expectation $E(y_i|\mathbf{x}_i)$ as a *conditional probability* that the outcome or attribute of interest will occur; i.e.,

$$E(y_i|\mathbf{x}_i) = \mathbf{x}_i\boldsymbol{\beta} = P(y_i = 1|\mathbf{x}_i),$$

making the usual assumptions that **X** is fixed and $E(\mathbf{u}) = \mathbf{0}$ (implying $\hat{\boldsymbol{\beta}}$ is unbiased).

Several problems arise with this model.

2.1 Non-normality of the errors

Fitting (1) by least squares produces residuals that have no known distribution, at least in finite samples (though asymptotically the residuals will tend to be Normal). To see this, consider u_i in the two cases:

$$\begin{aligned} y_i = 1 \quad &\Rightarrow \quad u_i = 1 - \mathbf{x}_i\hat{\boldsymbol{\beta}} \\ y_i = 0 \quad &\Rightarrow \quad u_i = -\mathbf{x}_i\hat{\boldsymbol{\beta}} \end{aligned}$$

Given that u_i can take on only one of two values, it is impossible for it to be Normal; it is in fact binomially distributed (Gujarati 1988, 470).

2.2 Heteroscedasticity

The constraints on u_i (which follow directly from the constraints on y_1) also mean that e_i has a variance that is data dependent. Drawing on results for

the binomial distribution or by a derivation (e.g., Gujarati 1988, 470), it is straightforward to show

$$V(\mathbf{u}) = \mathbf{X\beta}(1 - \mathbf{X\beta}).$$

This is easily fixed via two-stage GLS, weighting the data by $\mathbf{P} = 1/\sqrt{V(\hat{\mathbf{u}})}$ (Greene 1990, 663; Gujarati 1988, 470; Goldberger 1964, 248--50 is the usual cite, see Hanushek and Jackson 1977, 181--2).

2.3 Non-probabilities

The "real" problem with the LPM is that there is no guarantee it will produce conditional probabilities that in fact *are* probabilities. This is a serious drawback for any model purporting to be a "probability model." As it stands, there is no reason why the conditional probabilities produced by the LPM (simply $\mathbf{X}\hat{\boldsymbol{\beta}}_{LS}$) have to lie between 0 and 1, or even have positive variances.

A lot of energy has been expanded trying to constrain the LPM's predictions for y_i to the unit interval (Aldrich and Nelson 1984 review some of these). Basically all these are fairly cumbersome compared to the ease of estimating an alternative model, built from first principles, via MLE. Furthermore, the the "fixes" that one must apply to the LPM are all data-dependent and thus the resulting estimates have *no known sampling properties*. Inference and even interpretation in these circumstances is highly problematic.

3 Probability Models for Binary Data

In view of the problems of the LPM mentioned in section 2.3, we want a model that has the following two attributes:

$$\begin{aligned} \lim_{\mathbf{x}_i\boldsymbol{\beta}\rightarrow+\infty} \Pr[y_i = 1] &= 1 \\ \lim_{\mathbf{x}_i\boldsymbol{\beta}\rightarrow-\infty} \Pr[y_i = 1] &= 0 \end{aligned}$$

This is accomplished by finding a function that maps from plausible values of $E(y_i|X_i) = \mathbf{x}_i\boldsymbol{\beta}$ into the unit interval: i.e., we require a function

$$F(\cdot) : \mathbb{R} \rightarrow [0, 1],$$

or more specifically,

$$\Pr(y_i = 1) = F(\mathbf{x}_i\boldsymbol{\beta}), \quad \forall\, i = 1 \ldots, N.$$

Cumulative probability distribution functions (CDFs) have this property, and seem a fairly natural choice in these settings, although there is no *apriori* reason to restrict attention to CDFs; recent work has tried to create more general models for binary outcomes so as to overcome some of the limitations inherent in choosing a CDF as the function here (e.g., Stukel 1988, Robinson 1988, for reports from statistics and econometrics, respectively).

Two widely used CDFs in this setting are the normal (Gaussian) and the logistic. The functional forms of these CDFs are given below, along with the corresponding function for the LPM:

"Linear Probability Model":

$$F(z) = z$$

"Probit":

$$F(z) = \Phi(z) \equiv \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp[-(t^2/2)]dt.$$

"Logit":

$$F(z) = \Lambda(z) \equiv e^z/(1 + e^z) = 1/(1 + e^{-z})$$

The model for binary data can thus also be written as

$$\begin{aligned} E[y_i] &= E[y_i|y_i = 1] \cdot P[y_i = 1] + E[y_i|y_i = 0] \cdot P[y_i = 0], \\ &= 1[F(\mathbf{x}_i\boldsymbol{\beta})] + 0[1 - F(\mathbf{x}_i\boldsymbol{\beta})], \\ &= F(\mathbf{x}_i\boldsymbol{\beta}). \end{aligned} \tag{2}$$

But note that $Pr(y_i = 1) = F(\mathbf{x}_i\boldsymbol{\beta})$ and so $E[y_i] = F(\mathbf{x}_i\boldsymbol{\beta})$.

3.1 Nomenclature

Technically speaking, the term "logit" refers to the *inverse* of the function $\Lambda(\cdot)$, and similarly the term "probit" refers to the inverse of the function $\Phi(\cdot)$. A model for binary data using the logistic CDF is often (and more precisely) called a logistic regression model, while the term "probit" is used for the model using the normal CDF. The use of the term "probit" is ambiguous and leads to confusion with the model for proportions data that takes "probits" (sometimes called "normits") of the proportions as its dependent variable (see section 7 below).

The term "logit" also arises because it is trivial to show that the logit model is in fact linear in the log-odds of $Pr(y_i = 1)$; that is

$$L_i = \ln\left(\frac{P_i}{1 - P_i}\right) = \mathbf{x}_i\boldsymbol{\beta} \tag{3}$$

The term L_i is the "logit" of the probability $P_i = \Pr(y_i = 1)$. Note that $L_i \in \mathbb{R}$ while the probabilities are constrained to the unit interval. That is, the logit function is just the inverse of the $F()$ we use in this case, or more simply, if

$$P_i = F(\mathbf{x}_i\boldsymbol{\beta}) = \frac{1}{1 + e^{-\mathbf{x}_i\boldsymbol{\beta}}}$$

then

$$L_i = F^{-1}(P_i) = \ln\left(\frac{P_i}{1 - P_i}\right) = \mathbf{x}_i\boldsymbol{\beta}.$$

3.2 Logit vs Probit

The logistic PDF has thicker tails than the Normal, corresponding roughly with a student t-distribution with seven degrees of freedom, assigning marginally greater probability to $y_i = 0$ when $\mathbf{x}_i\boldsymbol{\beta}$ is small than does the probit model. The choice between the two will seldom generate substantively different results, unless the data on y_i are highly skewed, in which case a non-symmetric $F(\cdot)$ may well be worth considering (see Stukel 1988 for a summary). Figure 1 plots the two different PDFs and CDFS. The logistic distribution has a variance of $\pi^2/3$ and so is standardized for comparability with the Normal. This difference in the variances is not consequential in practice since the variance parameters of the $F(\cdot)$ distributions are not identified anyway, and are implicitly set to their default values (this point is elaborated in section 7 below), though it does help explain why probit and logit give different looking parameter estimates.

4 Random Utility Rationale

Another basis for the binary response models often seen in the econometric literature comes from the idea of assigning utilities to the various choices faced by a decision-maker. This idea is typically associated with the econometrician Daniel McFadden (e.g., 1974), but actually comes out of a much older literature in psychology (Thurstone 1927; Luce and Suppes 1965).

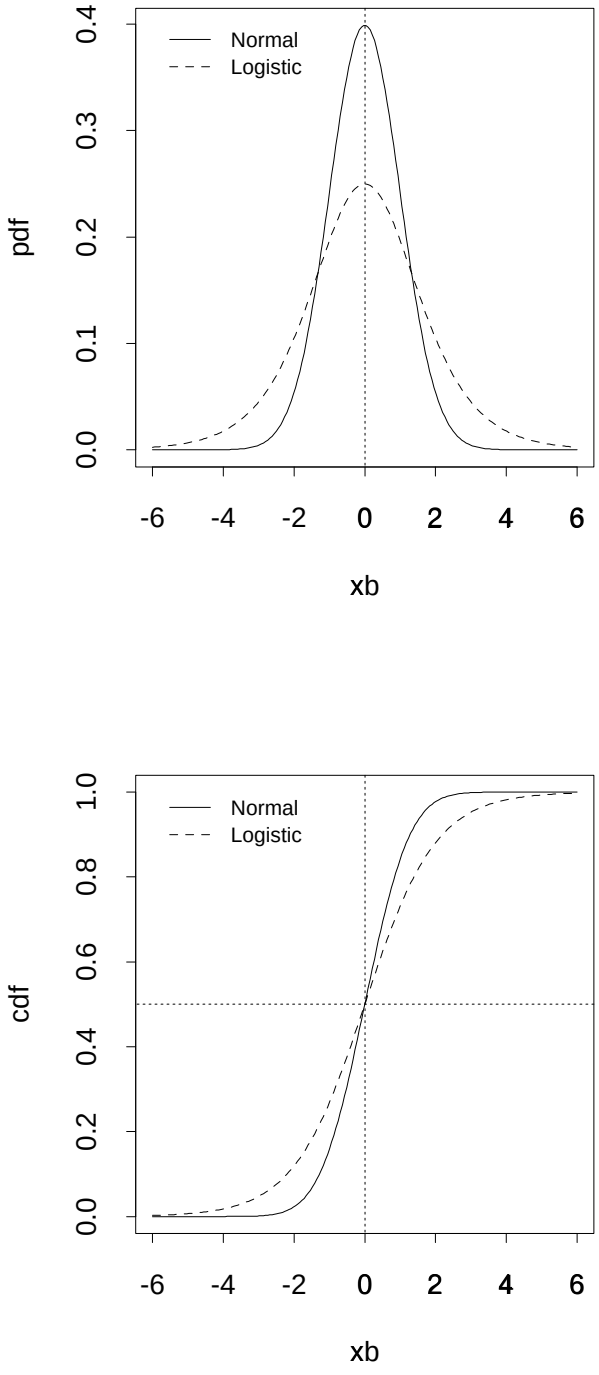

Figure 1: **The Standardized normal and Logistic PDFs and CDFs.** The top panel shows the PDFS; the lower panel the CDFs. Remember that the CDFs are simply the cumulative area under the PDFs. The logistic distribution assigns more probability mass to the tails.

Utility is the unobserved variable underlying the discrete choices observed by the analyst. What makes such a model probabilistic and thus useful for the statistical analysis of binary data is the additional assumption that decision-makers' utilities contain a random component. Generically, if a decision-maker is faced with two options, labelled "0" and "1", then the utility associated with each can be considered a linear function of some observed variables plus a random component:

$$\begin{aligned} U_{i0} &= \mathbf{x}_i\boldsymbol{\beta}_0 + e_{i0} \\ U_{i1} &= \mathbf{x}_i\boldsymbol{\beta}_1 + e_{i1}. \end{aligned}$$

Choice "1" is observed for the ith decision-maker if and only if $U_{i1} > U_{i0}$, and choice "0" otherwise. Note what this decision rule implies in terms of the equations above:

$$\begin{aligned} Pr(y_i = 1) &= Pr(U_{i1} > U_{i0}) \\ &= Pr[(\mathbf{x}_i\boldsymbol{\beta}_1 + e_{i1}) > (\mathbf{x}_i\boldsymbol{\beta}_0 + e_{i0})] \\ &= Pr[(\mathbf{x}_i\boldsymbol{\beta}_1 - \mathbf{x}_i\boldsymbol{\beta}_0) > (e_{i0} - e_{i1})] \\ &= Pr[(e_{i0} - e_{i1}) < \mathbf{x}_i(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)] \\ &= F[\mathbf{x}_i(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)] \end{aligned}$$

This model is very close to the typical setup,

$$Pr(y_i = 1) = F(\mathbf{x}_i\boldsymbol{\beta}),$$

and subject to some assumptions *is* the standard binary response model. As it stands it is impossible to estimate both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$ from the data, and so if we content ourselves to estimate their difference $\boldsymbol{\beta} \equiv \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$ we have the usual model. Also, notice that a specific distributional assumption for the form of $F(\cdot)$ is to do with the distribution of $e_{i0} - e_{i1}$, not each e_{ij} separately. If we assume the *difference* of the errors in the utilities is normally distributed then we have the probit model; if instead we assume the difference of the errors in the utilities is logistically distributed we have the logit model. The probit model in turn implies that the individual error components are normally distributed (since a normal distribution is also the distribution of the sum or difference of two normal distributions) but that the logit model implies that the individual error components each follow a Type-1 extreme value distribution, sometimes called a Gompertz distribution (Aldrich and Nelson 1984, 33) or a log-Weibull distribution (Amemyia 1985, 296) and has CDF

$$\exp[-\exp(-e_j)].$$

The random utility approach is important in understanding models for more than two choices which can not be ordered.

4.1 Latent Variable Approach

For a binary model, we can motivate the logit or probit model with a simpler version of the random utility approach. In this sense random utility might be considered a special case of a more general effort to provide a basis for motivating models for discrete outcomes. At a high level of generality, we can posit the existence of a latent quantity that gives rise to the observed discrete outcomes. Depending on the application it will be useful to give this latent quantity a label such as "probability", "random utility", "tolerance to dose", and so on. Economists sometimes refer to these models for binary outcomes as *index functions*.

We model the unobserved latent variable $\mathbf{y}^*$ with the model

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

with $u_i \overset{\text{iid}}{\sim} N(0, 1)$ for probit or standard logistic for logit. We don't observe y^*, just whether it crossed a threshold or not: i.e.,

$$\begin{aligned} y_i &= 1 \quad \text{if } y_i^* > \kappa \\ y_i &= 0 \quad \text{if } y_i^* \leq \kappa \end{aligned}$$

and so

$$\begin{aligned} \Pr(y_i = 1) &= \Pr(y_i{*} > \kappa) \\ &= \Pr(\mathbf{x}_i\boldsymbol{\beta} + u_i > \kappa) \\ &= \Pr(u_i > \kappa - \mathbf{x}_i\boldsymbol{\beta}) \end{aligned}$$

It is customary to set $\kappa = 0$ or otherwise have non-zero κ absorbed by the intercept term in the model[1] and so

$$\Pr(y_i = 1) = \Pr(u_i > -\mathbf{x}_i\boldsymbol{\beta})$$

Likewise the use of the standard normal or the standardized logistic function imposes unit variance on the $\mathbf{u}$. This is is necessary to identify the $\boldsymbol{\beta}$; the $\boldsymbol{\beta}$ are

[1] The only way to estimate this model *without* an intercept is to assume $\kappa = 0$; in general it is not a good idea to estimate these models without an intercept.

only defined up to a constant multiple of σ, such that both $\boldsymbol{\beta}$ and σ can't be both estimated from the data. Setting $\sigma = 1$ is a convenient way of settling this identification issue. We now have

$$\begin{aligned} P_i &\equiv \Pr[y_i = 1], \\ &= \Pr(u_i > -\mathbf{x}_i\boldsymbol{\beta}) \\ &= \int_{-\mathbf{x}_i\boldsymbol{\beta}}^{\infty} f(u_i)\, du_i\,, \\ &= 1 - \int_{-\infty}^{-\mathbf{x}_i\boldsymbol{\beta}} f(u_i)\, du_i\,, \\ &= 1 - F(-\mathbf{x}_i\boldsymbol{\beta}). \end{aligned}$$

Note that for symmetric distribution functions like the normal or the logistic CDFs,

$$P(y_i = 1) = F(\mathbf{x}_i\boldsymbol{\beta}) = 1 - F(-\mathbf{x}_i\boldsymbol{\beta}),$$

and we have the familiar form of the binary response model

$$P_i \equiv Pr[y_i = 1] = F(\mathbf{x}_i\boldsymbol{\beta}).$$

5 Linearities and Non-linearities

Note immediately that the logit/probit model for the observed y_i is non-linear in the parameters; unlike a standard regression we *do not* have $\partial E(y_i)/\partial x_i = \boldsymbol{\beta}$. Rather, we have

$$\frac{\partial E(Pr[\mathbf{y} = 1])}{\partial \mathbf{X}} = \left\{ \frac{d\, F(\mathbf{X}\boldsymbol{\beta})}{d\,(\mathbf{X}\boldsymbol{\beta})} \frac{d\,(\mathbf{X}\boldsymbol{\beta})}{d\,\mathbf{X}} \right\} = f(\mathbf{X}\boldsymbol{\beta})\boldsymbol{\beta} \tag{4}$$

In the case where $F(\cdot)$ is the normal CDF the marginal effects are $\phi(\mathbf{x}_i\boldsymbol{\beta})\boldsymbol{\beta}$, where $\phi(\cdot)$ is the standard Normal probability density function (PDF). In the case of the logistic CDF we obtain

$$\begin{aligned} \frac{d\,\Lambda(\mathbf{X}\boldsymbol{\beta})}{d\,(\mathbf{X}\boldsymbol{\beta})} &= \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{(1 + e^{\mathbf{X}\boldsymbol{\beta}})^2}, \\ &= \Lambda(\mathbf{X}\boldsymbol{\beta})[1 - \Lambda(\mathbf{X}\boldsymbol{\beta})], \end{aligned}$$

which in turn yields

$$\frac{\partial E[\mathbf{y}]}{\partial \mathbf{X}} = \Lambda(\mathbf{X}\boldsymbol{\beta})[1 - \Lambda(\mathbf{X}\boldsymbol{\beta})]\boldsymbol{\beta}.$$

Note however that logit and probit models *are linear* in the latent variable $\mathbf{y}^*$; for logit, this latent variable is just the log of the odds ratio (as discussed earlier). ***This is a critical feature of interpreting and reporting results from logit and probit models.***

6 Reporting Results

These partial derivatives vary with $\mathbf{X}$, following the "floor" and "ceiling" effects implicit in the use of the function $F(\cdot)$. In practice, analysts usually evaluate these marginal effects holding other variables at their sample means or medians.

Accordingly, one of the first tables that ought to be reported in a write-up of an analysis of a LDV is a listing of descriptive statistics on the independent variables; this is useful for analyst and reader alike in trying to get a sense for the relative effects of the independent variables.

Alternatively, graphical displays of the varying marginal effects are useful; say, plotting $P(y_i = 1)$ over the sample range of an independent variable, given different assumptions about the values of the other independent variables. In section 9.3 I provide a formula for computing the standard errors of these predicted probabilities. It is often the case that some of the independent variables are binary indicators of characteristics of the observations (so-called "dummy variables"). Examples of binary predictors might include gender, race, ethnicity, level of education in an analysis of voter turnout. Taking the mean or median of these binary variables in computing marginal effects will not be as useful as computing (and plotting) the marginal effects sub-group by sub-group. The programming required here can sometimes be tedious, but even "canned" software is getting better at allowing users to perform these sorts of calculations on the data with a set of parameter estimates.

However it is done, investigation of the substantive effects of an independent variable proceeds according to the scheme I employ in the example below. The key notion here is that the relationship between $\mathbf{X}$ and the probabilities is not linear, but mediated by the non-linear $F(\cdot)$ function. A change in $\mathbf{x}_i\boldsymbol{\beta}$ will not have a constant effect on $P[y_i = 1]$, *but will depend on where* $P[y_i = 1]$ *is to begin with*. This "starting probability" can be set by the analyst arbitrarily, or more usefully by generating a predicted probability, $P^* \equiv P[y_i = 1|\mathbf{x}_0, \hat{\boldsymbol{\beta}}]$, from a vector of substantively interesting values on the independent variables ($\mathbf{x}_0$) multiplied by the vector of estimated coefficients $\hat{\boldsymbol{\beta}}$; i.e., $\hat{P} = F(\mathbf{x}_0\hat{\boldsymbol{\beta}})$.

Denote the independent variable of interest as $\mathbf{x}_k$, and some interesting level of change in this variable as $\Delta\mathbf{x}_k$. The probability that $y_i = 1$ after $\mathbf{x}_k$

changes by Δx_k is

$$\hat{P}^* \equiv P[y_i = 1 | \mathbf{x}_0, \Delta\mathbf{x}_k, \hat{\boldsymbol{\beta}}] = F[\mathbf{x}_0'\hat{\boldsymbol{\beta}} + \Delta x_k \hat{\beta}_k] = F[\mathbf{x}^{*'}\hat{\boldsymbol{\beta}}].$$

Some examples will cast light on how to use the results of a logit or a probit in this fashion.

6.1 Example: The 1964 U.S. Presidential Election

Consider the following table of results from an analysis of voting in the 1964 U.S. Presidential election (Hanushek and Jackson 1977, 182, 205):

	Methods of Estimation			
Variable	OLS	LPM	Logit	Probit
Intercept	.087	.074	-2.864	-1.62
	(.021)	(.020)	(.142)	(.074)
Issue Evals	.360	.362	3.060	1.73
(0-1)	(.031)	(.031)	(.198)	(.107)
Party ID	.636	.647	4.176	2.36
(0-1)	(.028)	(.028)	(.044)	(.095)

The dependent variable is scored 1 for a vote for Johnson, 0 for a Goldwater vote (non-voters and voters for other candidates are excluded from the analysis); issue evaluations are collapsed and re-scaled to lie between 0 and 1 (1 being pro-Johnson); and party identification is coded 0, .25, .75 and 1 for strong Republican, weak Republican, weak Democrat, and strong Democrat, respectively.

Consider a voter who is indifferent between Johnson and Goldwater. In this case the voter has a starting probability of .5. This .5 corresponds to $\mathbf{x}_i\boldsymbol{\beta} = 0$, which we can read off by using the inverse of F, $\Phi^{-1}(\cdot)$ for the Normal, and $\Lambda^{-1}(\cdot)$ for the logistic. For brevity I will focus only on the probit case in these examples.

What happens as an indifferent voter starts to align with Johnson on the issues of the election? Specifically, let us imbue this hypothetical indifferent voter with fifty percentage points worth of pro-Johnson issue evaluations. In the notation used above, we have $\Delta\mathbf{x}_k = .5$, which we multiply by the probit coefficient for issue evaluations (1.73) to come up with a $X^{*'}\hat{\boldsymbol{\beta}}$ of $0 + .5(1.73) \approx .87$. Taking $\Phi(.87)$ we obtain the new probability of voting for Johnson, $\hat{P}^* = .81$.

As another example, consider the case of a hypothetical weak Democratic identifier who holds fiercely anti-Johnson issue evaluations. How likely is it that such a person will vote for Johnson? We determine this probability by forming a vector of values on the independent variables capturing this hypothetical voter's characteristics, and multiplying it by the probit parameter estimates;

$$\mathbf{x}^* = (1, 0, .75)\,, \quad \hat{\boldsymbol{\beta}} = (-1.62, 1.73, 2.36)'$$

and the product of these two vectors is just -1.62 + .75 (2.36) = .15. The probability of someone with these characteristics voting for Johnson is $\Phi(.15) \approx .56$, already better than an even-money bet for this weak partisan to vote Johnson, despite the strident divergence from the candidate's issue positions.

But what if this hypothetical weak Democrat should moderate her anti-Johnson issue stances, say, fifty percentage points (i.e., half the possible range of issue evaluations)? To do this we add .5 (1.73) $\approx$ .87 to the .15 quantity we just obtained, which yields a $\mathbf{x}^*\hat{\boldsymbol{\beta}}$ of 1.02. The estimated probability that this voter votes for Johnson is now $\Phi(1.02) = .84$.

The central idea here is that hypothetical movement on the independent variables translates into probability statements via the $F(\cdot)$ function; and, conversely, probabilities can be related back to values of the independent variables via the inverse of the $F(\cdot)$ function, F^{-1}.

7 Proportions Data

Instead of unit-level data we have proportions of $y_i = 1$ in groups $t = 1, \ldots T$. Objections to LS regression apply here, as do the failings of the LPM; we need predicted proportions in the unit interval.

The underlying model is that a (possibly multivariate) "dose" $\mathbf{x}_t$ is applied to each of N_t units (insects, people, econometricians). In each of the $t = 1, \ldots, T$ groups a certain proportion P_t of the group present the outcome or attribute of interest (e.g., death, voting, using public transport). Doses are constant within groups which *effectively* means we have T observations (and therefore the matrix of predictors $\mathbf{X}$ must be of rank less than T).

At the unit-level the level of "stimulation" is related to the dose, $\mathbf{X}$ by

$$s_t = \mathbf{x}_t\boldsymbol{\beta}.$$

If the stimulus exceeds each individuals' unobserved tolerance level x^*_{it} then we observe the outcome or attribute of interest, and for tractability it is typically assumed that the unobserved tolerance levels follow a PDF with

known parameters estimable only up to some scaling factor. The PDF often used is the standard Normal, with mean $\mu = 0$ and variance $\sigma^2 = 1$. The intercept of the model provides an estimate of μ, but σ^2 remains unidentified.

The unit-level model is therefore

$$\begin{aligned} Pr[y_{it} = 1] &= Pr[x^*_{it} < s], \\ &= Pr[x^*_{it} < \mathbf{X}_t\boldsymbol{\beta}], \\ &= F(\mathbf{x}_t\boldsymbol{\beta}), \end{aligned}$$

which is identical to what we have for the binary outcome case estimated with unit-level data.

The data here consist of P_t, the "positive response" proportions, and the $\mathbf{x}_t$, the doses. In each t we have

$$P_t \approx \Phi(\mathbf{x}_t\boldsymbol{\beta}),$$

since the observed P_t is just an estimate of the true group-specific proportion π_t. A model that can be estimated via GLS is simply

$$z_t = \Phi^{-1}(P_t) = \mathbf{x}_t\boldsymbol{\beta} + \mathbf{e}_t$$

where $\Phi^{-1}(\cdot)$ is the inverse of the normal CDF. GLS is necessary since given the argument in section 2.2 it is clear that the $\mathbf{e}_t$ will have heteroscedastic variances. Greene (1993, 654) shows each group-specific observation to have error variance

$$\frac{\Phi_t(1 - \Phi_t)}{N_t\phi_t^2}$$

where $\Phi_t = \Phi_t(\mathbf{x}_t\boldsymbol{\beta})$, and similarly for ϕ_t. The model estimated if the logistic distribution is chosen is

$$z_t = \Lambda^{-1}(P_t) = \ln\left(\frac{P_t}{1 - P_t}\right) = \mathbf{x}_t\boldsymbol{\beta} + \mathbf{e}_t,$$

again subject to the caveat of non-constant error variances.

Maximum likelihood is the usual way to proceed with these types of models, recognizing that "grouped" binary data follow a binomial distribution if the binary events are independent within each group. That is, if $y_{it} \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta_t)$, then the number of successes $r_t = \sum_{i=1}^{n_t} y_{it} \sim \text{Binomial}(\theta_t; n_t)$. We then need a model relating covariates $\mathbf{x}_t$ to the success probability θ_t, for which the logistic model is a convenient choice: i.e., $\theta_t = F(\mathbf{x}_t\boldsymbol{\beta})$ where $F(\cdot) = 1/(1 + \exp(\cdot))$.

8 Estimation

Maximum likelihood provides a convenient and powerful method for estimating the parameters of the logit/probit model. A key assumption is that the data are identically and independently distributed, which allows us to form a likelihood function for the whole data from the product of the likelihoods for each observation:

$$\begin{aligned} P(y_1, y_2, \ldots, y_n) &= P(y_1)P(y_2)\ldots P(y_n) \\ &= \prod_{y_i=1} F(\mathbf{x}_i\boldsymbol{\beta}) \prod_{y_i=0} [1 - F(\mathbf{x}_i\boldsymbol{\beta})] \end{aligned}$$

where the notation $\prod_{y_i=1}$ means "the cumulative product over the observations for which $y_i = 1$", and similarly for $y_i = 0$. This is usually written as

$$\mathcal{L} = \prod_{i=1}^{N} [F(\mathbf{x}_i\boldsymbol{\beta})]^{y_i} [1 - F(\mathbf{x}_i\boldsymbol{\beta})]^{1-y_i}.$$

Each observation thus contributes something to the likelihood, either in the first part when $y_i = 1$, or in the second part when $y_i = 0$ (so $1 - y_i = 1$). As is typical with MLE, it is easier to work with the log-likelihood:

$$\ln \mathcal{L} = \sum_{i=1}^{N} y_i \ln F(\mathbf{x}_i\boldsymbol{\beta}) + (1 - y_i) \ln[1 - F(\mathbf{x}_i\boldsymbol{\beta})].$$

It can be shown that the regularity conditions required for the MLEs of $\boldsymbol{\beta}$ to have the usual consistency and asymptotic normality properties are satisfied (see Amemiya 1985, 270--5; Gouriéroux and Monfort 1981).

Optimizing this function with respect to the unknown vector $\boldsymbol{\beta}$ requires iterative techniques that thankfully take place largely "behind the scenes" for most software packages. The solution is necessarily iterative because the first-order condition

$$\frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \left[\frac{y_i f(\mathbf{x}_i'\hat{\boldsymbol{\beta}})}{F(\mathbf{x}_i'\hat{\boldsymbol{\beta}})} + (1 - y_i) \frac{-f(\mathbf{x}_i'\hat{\boldsymbol{\beta}})}{1 - F(\mathbf{x}_i'\hat{\boldsymbol{\beta}})} \right] \mathbf{x}_i = 0,$$

is non-linear in $\hat{\boldsymbol{\beta}}$ (Greene 1993, 644; Maddala 1983, 256-6), and has no simple, analytical solution for $\hat{\boldsymbol{\beta}}$.

Good software will display results at the end of each iteration so that the user can be sure that convergence of the optimization algorithm is occurring

rapidly and smoothly, as it should if the binary outcomes are not too skewed, and the independent variables are not highly collinear. Both these data-specific factors are almost always the source of any apparent problem in the optimization algorithms used by statistical software.

8.1 Quasi-Complete Separation

MLEs going to plus/minus infinity; predicted probabilities close to zero and 1. Dummy variable for a single micro-level observation in a binary response model.

9 Hypothesis Testing

9.1 Variance-Covariance Matrix

Since the estimates of $\boldsymbol{\beta}$ are MLEs, the asymptotic variance-covariance matrices of the parameters come from inverting estimates of the information matrix. Again, in practice this is largely done "behind the scenes" by statistical software, but Greene (1993, 644--5) notes that there are three common ways to estimate the information matrix:

1. Minus the Hessian of the log-likelihood, evaluated at the MLEs (the Hessian being the matrix of second partial derivatives of $\ln \mathcal{L}$ with respect to $\boldsymbol{\beta}$);

2. Minus the expected value of the Hessian;

3. The outer product of the first derivatives of $\ln \mathcal{L}$ with respect to $\boldsymbol{\beta}$, summed over the observations (the BHHH (1974) estimator).

These differences are not consequential here, but can be in more complicated qualitative dependent variable models. For the relatively simple binary outcome model the second derivatives of the log-likelihood are well-known and programmed directly into most software packages that support estimation of probit and logit models. For more general or complicated qualitative dependent variable setups the Hessian may be difficult to derive, program or compute, so "derivative-free" methods may be employed by the software, typically using a combination of computer-intensive, numerical approximations of the first derivatives in conjunction with the the BHHH method, and the resulting estimates (and their estimated standard errors) can be more sensitive to the

vagaries of model specification, poorly behaved data, etc. All this is to say that things can sometimes go wrong when it comes time to estimate these models, particularly in generating standard errors of the parameters, but that for the standard qualitative dependent variable setups this really shouldn't be too much of a worry.

Standard errors are obtained by taking the square-root of the leading diagonal of the variance-covariance matrix of the parameters. *z*-tests can be applied to the individual coefficients divided by their estimated standard errors, which are valid asymptotically.

9.2 Likelihood Ratio Tests

Likelihood ratio tests are the way to test restrictions of multiple coefficients, instead of the *F* tests used in LS multiple regression; recall from the discussion of MLE that likelihood ratio test statistics are distributed χ^2 with degrees of freedom equal to the number of parameter restrictions. The test statistic is simply

$$-2(\widehat{\ln \mathcal{L}_r} - \widehat{\ln \mathcal{L}})$$

where $\widehat{\ln \mathcal{L}_r}$ is the log-likelihood for a restricted model and $\widehat{\ln \mathcal{L}}$ is the log-likelihood for the full model.

9.3 Predicted Probabilities

The predicted probabilities also have standard errors, and really ought to be reported or graphed when one simulates changes on particular independent variables (i.e., "comparative statics" for a logit/probit model). At the risk of cluttering a graph, one ought to plot alongside predicted probabilities plus and minus the critical number of standard errors required for a given confidence interval (± 1.96 standard errors for a 95% confidence interval with a large data set).

A predicted probability $\hat{P}(y_i = 1|X_i, \hat{\boldsymbol{\beta}}) = F(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$ has asymptotic variance

$$\left(\frac{\partial F(\mathbf{x}_i'\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right)' \mathbf{V} \left(\frac{\partial F(\mathbf{x}_i'\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right),$$

where $\mathbf{V}$ is the asymptotic variance of $\hat{\boldsymbol{\beta}}$.

In equation (4) I showed the derivative of $F(\mathbf{x}_i'\hat{\boldsymbol{\beta}})$ with respect to $\hat{\boldsymbol{\beta}}$ to be $f(\mathbf{x}_i'\hat{\boldsymbol{\beta}})\mathbf{X}$, which substituting into the above expression yields

$$AV[F(\mathbf{x}_i'\hat{\boldsymbol{\beta}})] = [f(\mathbf{x}_i'\hat{\boldsymbol{\beta}})]^2 X_i'\mathbf{V}X_i,$$

(noting that $\mathbf{x}_i'\mathbf{V}\mathbf{x}_i$ is a scalar (a 1 by k vector times the k by k variance-covariance matrix times a k by 1 vector), where *AV* stands for asymptotic variance. Standard errors are just the square root of this quantity. One could substitute a vector of hypothetical values on the independent variables x_i^* into this expression to obtain the standard errors of predicted probabilities under difference scenarios of interest to the analyst, in the fashion of section 6.

The trouble with this method is that it sometimes yields upper or lower confidence bounds that are outside the unit probability interval. These are sometimes merely truncated at 0 or 1.

Another approach, and one I follow, is to calculate confidence intervals in the metric of the underlying continuous variable (i.e., y^*, on the logit log-odds scale or the probit scale), according to the usual formula from least squares regression:

$$var(\hat{Y}_i|\mathbf{x}_0, \mathbf{X}, \hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0) \quad (5)$$

where by assumption, $\sigma^2 = 1$, and then convert $\hat{Y}_i$ and resulting upper and lower confidence bounds on $\hat{Y}_i$ into probabilities using $F(\cdot)$: i.e., the upper-bound on the probability metric is given by $F[\hat{Y}_i + t_{\alpha/2} \cdot se(\hat{Y}_i)]$.

10 Goodness-of-Fit Measures

No straightforward counterpart of the LS regression r^2 exists in this framework. There are many proposals for goodness-of-fit summary measures in the literature. Some of the more popular ones are described below.

10.1 *F* test analogy

When running a LS regression software packages typically dump a F test statistic for the hypothesis that all coefficients are zero. A similar statistic is dumped by many binary outcomes programs, but testing the null hypothesis that all the *slope* coefficients are zero. This statistic is often in the form of a restricted likelihood, obtained by simulating a model with no predictors save for an intercept. This restricted likelihood statistic can be computed as

$$\ln \mathcal{L}_0 = N[P \ln P + (1 - P) \ln(1 - P)]$$

where P is the proportion of observations with $y = 1$ in the sample (Greene 1990, 682). One can then do a likelihood ratio test using the log-likelihood obtained from the full set of predictors. However, seldom is the joint null hypothesis of all predictors being zero a substantively interesting hypothesis.

10.2 r^2 analogy

The quantity

$$1 - \frac{\widehat{\ln \mathcal{L}}}{\ln \mathcal{L}_0},$$

is bounded between 0 and 1 and thus behaves something like r^2 from LS regression, taking on values close to 1 when the model performs well (and there is a big difference between $\widehat{\ln \mathcal{L}}$ and $\ln \mathcal{L}_0$) and values close to 0 when the model performs poorly (McFadden 1974; Maddala 1983, 39--40). This is sometimes referred to as McFadden's pseudo-r^2. Like r^2 in the LS regression context this index has the drawback of lacking a substantively meaningful interpretation.

Another possibility is to take the squared correlation between **y** and the predicted probabilities of $y_i = 1$, $F(\mathbf{X\beta})$. Despite the probabilities being bounded and **y** being binary, one can still calculate the squared correlation between the two, which Maddala (1983, 38; following Goldberger 1973) shows in this case to be

$$r^2 = \frac{V(\hat{p})}{E(\hat{p}) - [E(\hat{p})]^2}.$$

10.3 "Hits and Misses"

One way of assessing the predictive power of a model for a qualitative dependent variables is to cross-tabulate predicted outcomes with actual outcomes, using the rule that if a predicted probability is greater than .5, we ought to count that as model prediction of 1, and 0 otherwise. One then might compare the model's performance against a näive or null model that predicts $y_i = 1, \forall\, i$, if the observed proportion of $y_i = 1 > .5$, or $y_i = 0, \forall\, i$ otherwise. This näive model will predict correctly $100P$ percent of the observations, where P is the proportion of observations with $y_i = 1$, and in many instances stacks the cards against the model being estimated, say, if the data are highly skewed. A model that is highly significant in terms of t-tests and likelihood ratio tests can often fail to improve much over a null model.

As an example, consider Table 1. Here the model predicts 471 + 20 = 491 out of 690 cases (71.2%) of the observations correctly, though the pseudo-r^2 is only .083. The näive model always predicts $y_i = 0$ and thus is correct 487 out of 690 (70.6%). The estimated model, with 16 regressors, picks up only an additional four cases over the null model, and on this criterion might be regarded highly suspect, despite the individual parameter estimates

Table 1: *Tunali's (1986) study of migration and earnings in Turkey (Greene 1990, 683).*

		Predicted		
		$y_i = 0$	$y_i = 1$	*Total*
Actual	$y_i = 0$	471	16	487
	$y_i = 1$	183	20	203
	Total	654	36	690

performing quite well. Also, we see from Table 1 that the model has real difficulty predicting cases with $y_i = 1$; it predicts just 36 ones, 20 of which are "hits", out of 203 in the sample, for a success ratio of just 10%. Still, recall that the null model here predicts *no* ones at all. Tables like these are illustrative at highlighting a model's weaknesses or strengths. In this case a 1 is a relatively rare event, occurring in just 29.4% of the sample, and perhaps an additional regressor or a change in the functional form might be warranted to help explore what is going on in that part of the data. Nonetheless, one can see how even in moderately skewed data it becomes difficult to assess model performance against a seemingly näive competitor.

10.4 ROC

Receiver-operator characteristic curves. Hosmer and Lemeshow, p160.

11 GLM

Many limited dependent variable setups can be written in the McCullagh and Nelder (1989) general linear models (GLM) framework. This is how R implements models for binary responses and counts, and in theory this can be used for models for data with densities are in the exponential family (including regression models for normal data).

GLMs generalize the standard linear model

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i, \tag{6}$$

by positing two functions:

1. a *link* function, $g(\cdot)$, that describes how the mean of $y \equiv E(y) = \mu$ depends on the linear predictors: $g(\mu) = X\beta$
2. a *variance* function that captures how the variance of y depends upon the mean: var(y)=$\phi V(\mu)$, with ϕ constant (See Hastie and Pregibon 1992, 196--7).

The logit model for a special case of a GLM with a logit link $g(\mu) = \log(\mu/(1\text{-}\mu))$, and a binomial variance function, $V(\mu) = \mu(1 - \mu)/n$. Probit uses the inverse of the normal CDF as a link function.

11.1 Exponential Family

$$f(y; \theta, \tau) = \exp\left(\frac{a(y)b(\theta) + c(\theta)}{h(\tau)} + d(y, \tau)\right)$$

where

- τ is a dispersion parameter
- θ is a parameter(s) that relate to the mean function; $\eta = b(\theta)$ is the *natural parameter*.

For example, the binomial probability mass function is

$$f(x) = \binom{n}{x} p^x (1 - p)^{n\text{-}x}$$

for $x \in \{0, 1, 2, \ldots, n\}$. Using the notation introduced above, we have

$$f(x) = \exp\left(x \log\left(\frac{p}{1 - p}\right) + n \log(1 - p)\right)$$

and so

- $p = \theta$
- $\eta = b(\theta) = \log \frac{p}{1\text{-}p}$
- $c(\theta) = n \log(1 - p)$

With covariates $\mathbf{x}_i$ as in a logistic regression model for binary data y we have $p = F(\mathbf{x}_i\boldsymbol{\beta})$, $F(\cdot) = 1/(1 + \exp(\cdot))$ and $n = 1$ (for unit-record, micro-data) so now

$$f(y) = \exp\left(y_i F(\mathbf{x}_i\boldsymbol{\beta}) + \log\left(1 - F(\mathbf{x}_i\boldsymbol{\beta})\right)\right)$$

CHECK THIS.

12 References

Aldrich, John H. and Forrest D. Nelson. 1984. *Linear Probability, Logit, and Probit Models*. Sage University Paper series on Quantitative Applications in the Social Sciences, series np. 07-045. Sage. Newbury Park, California.

Amemiya, Takeshi. 1985. *Advanced Econometrics*. Harvard. Cambridge.

Berndt, E. R., B. H. Hall, R. E. Hall, and J. A. Hausman. 1974. Estimation and Inference in Nonlinear Structural Models. *Annals of Economics and Social Measurement*. 3:653--66.

Bliss, C. I. 1935. The Calculation of the Dosage-Mortality Curve. *Annals of Applied Biology*. 22:134-167.

Goldberger, A. 1964. *Econometric Theory*. Wiley. New York.

Goldberger, A. 1973. Correlations Between Binary Choices and Probabilistic Predictions. *Journal of the American Statistical Association*. 68:84.

Gouriéroux C., and A. Monfort. 1981. Asymptotic Properties of the Maximum Likelihood Estimator in Dichotomous Logit Models. *Journal of Econometrics*. 17:83--97.

Gujarati, Damodar N. 1988. *Basic Econometrics*. 2nd edition. McGraw Hill. New York.

Hastie, Trevor J., and Daryl Pregibon. 1992. Generalized Linear Models. In Chambers, John M. and Trevor J. Hastie. eds. *Statistical Models in S*. Wadsworth and Brooks/Cole. Pacific Grove, California. 195--248.

Hanushek, Eric A., and John E. Jackson. 1977. *Statistical Methods for Social Scientists*. Academic Press. New York.

Hosmer, David W. and Stanley Lemeshow. 2000. *Applied Logistic Regression*. 2nd edition. Wiley. New York.

King, Gary. 1989. *Unifying Political Methodology*. Cambridge University Press. New York.

Luce, R. D. and P. Suppes. 1965. Preference, Utility, and Subjective Probability. In R.D. Luce, R. Bush, and E. Galanter (eds.). *Handbook of Mathematical Psychology*. Volume 3. Wiley. New York.

Maddala, G. S. 1983. *Limited Dependent Variables and Qualitative Variables in Econometrics*. Econometric Society Monograph No. 3. Cambridge University Press. New York.

McCullagh, P. and J.A. Nelder. 1989. *Generalized Linear Models*. 2nd edition. Chapman and Hall. London.

McFadden, Daniel. 1974. The Measurement of Urban Travel Demand. *Journal of Public Economics*. 3:303-28.

Robinson, P. M. 1988. Semiparametric Econometrics: A Survey. *Journal of Applied Econometrics*. 3:35-51.

Stukel, Thérèse A. 1988. Generalized Logistic Models. *Journal of the American Statistical Association*. Theory and Methods. 83:426-431.

Thurstone, L. 1927. A Law of Comparative Judgement. *Psychological Review*. 34:273--86