University of Illinois
Fall 1998

Department of Economics
Roger Koenker

Economics 472

Lecture 16

Binary Dependent Variable Models

Let's begin with a model for an observed proportion, or frequency. We would like to explain variation in the proportion p_i as a function of covariates x_i. We could simply specify that

$$p_i = x_i'\beta + \text{ error}$$

and run it as OLS regression. But this has certain problems. For example, we might find that $\hat{p}_i \notin [0,1]$. So we typically consider transformations

$$g(p_i) = x_i'\beta + \text{ error}$$

where g is usually called the "link" function. A typical example of g is the logit function

$$g(p) = \text{logit}(p) = \log(p/1-p)$$

this corresponds to the logistic df.

The transformation may be seen to induce a certain degree of heteroscedasticity into the model. Suppose each observation $\hat{p}_i$ is based on a moderately large sample of n_i observations with $\hat{p}_i \to p_i$.

We may then use the δ-method to compute the variability of $\text{logit}(\hat{p}_i)$,

$$\begin{aligned} V(g(\hat{p}_i)) &= (g'(p_i))^2 V(\hat{p}_i) \\ g(p) &= \log(p/(1-p)) \\ g'(p) &= \frac{1-p}{p} \cdot \frac{d}{dp}\left(\frac{p}{1-p}\right) = \frac{1}{p(1-p)} \\ V(\hat{p}_i) &= \frac{p_i(1-p_i)}{n_i} \end{aligned}$$

so

$$V(\text{logit}(\hat{p}_i)) = \frac{1}{n_i p_i (1-p_i)}$$

Thus GLS would suggest running the weighted regression of $\text{logit}(\hat{p}_i)$ on x_i with weights $n_i p_i (1-p_i)$. Of course, we could, based on considerations so far, replace $\text{logit}(\hat{p}_i)$ with any other quantile-type transformation from [0,1] to $\mathbb{R}$. For example, we might use $\Phi^{-1}(\hat{p}_i)$ in which case the same logic suggests regressing

$$\Phi^{-1}(\hat{p}_i) \text{ on } x_i \text{ with weights } \frac{n_i \phi^2(\Phi^{-1}(p_i))}{p_i(1-p_i)}$$

An immediate problem presents itself, however, if we would like to apply the foregoing to data in which some of the observed p_i are either 0 or 1.

Since the foregoing approach seems rather *ad hoc* any way based as it is an approximate normality of the $\hat{p}_i$ we might as well leap in the briar patch of mle. But to keep things quite close to the

regression setting we will posit the following latent variable model. We posit the model for the latent (unobserved) variable y_i^* and assume that the observed binary response variable y_i is generated as,

$$\begin{aligned} y_i^* = & \quad x_i'\beta + u_i \\ & y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

so letting the df of u_i be denoted by F,

$$\begin{aligned} P(y=1) &= P(u_i > -x_i'\beta) = 1 - F(-x_i'\beta) \\ P(y=0) &= F(-x_i'\beta) \end{aligned}$$

For F symmetric $F(z) + F(-z) = 1$ so $f(z) = f(-z)$ and we have

$$\begin{aligned} P(y=1) &= F(x_i'\beta) \\ P(y=0) &= 1 - F(x_i'\beta) \end{aligned}$$

and we may write the likelihood of seeing the sample $\{y_i, x_i) : \quad i = 1, \ldots, n\}$ as

$$\begin{aligned} \mathcal{L}(\beta) &= \prod_{i:y_i=0} (1 - F(x_i'\beta)) \prod_{i:y_i=1} F(x_i'\beta) \\ &= \prod_{i=1}^{n} F_i^{y_i} (1 - F_i)^{1-y_i} \end{aligned}$$

Now we need to make some choice of F. There are several popular choices:

(i): Logit $p = F(z) = \frac{e^z}{1+e^z} \Rightarrow \log(p/1-p) = z$ so $Ey_i = p_i = F(x_i\beta) \Rightarrow \text{logit}(p_i) = x_i'\beta$

(ii): Probit $F(z) = \Phi(z) = \int_{-\infty}^{z} \phi(x)dx \;\; \Phi^{-1}(p) = x_i'\beta$

(iii): Cauchy $F(z)\frac{1}{2} + \pi^{-1}\tan^{-1}(z) \;\; F^{-1}(p) = \tan(\pi)(p - \frac{1}{2})) = x_i'\beta.$

(iv): Complementary log log

$$F^{-1}(p) = \log(-\log(1-p)) = x_i'\beta$$

(v): log-log

$$F^{-1}(p) = -\log(-\log(p)) = x_i\beta$$

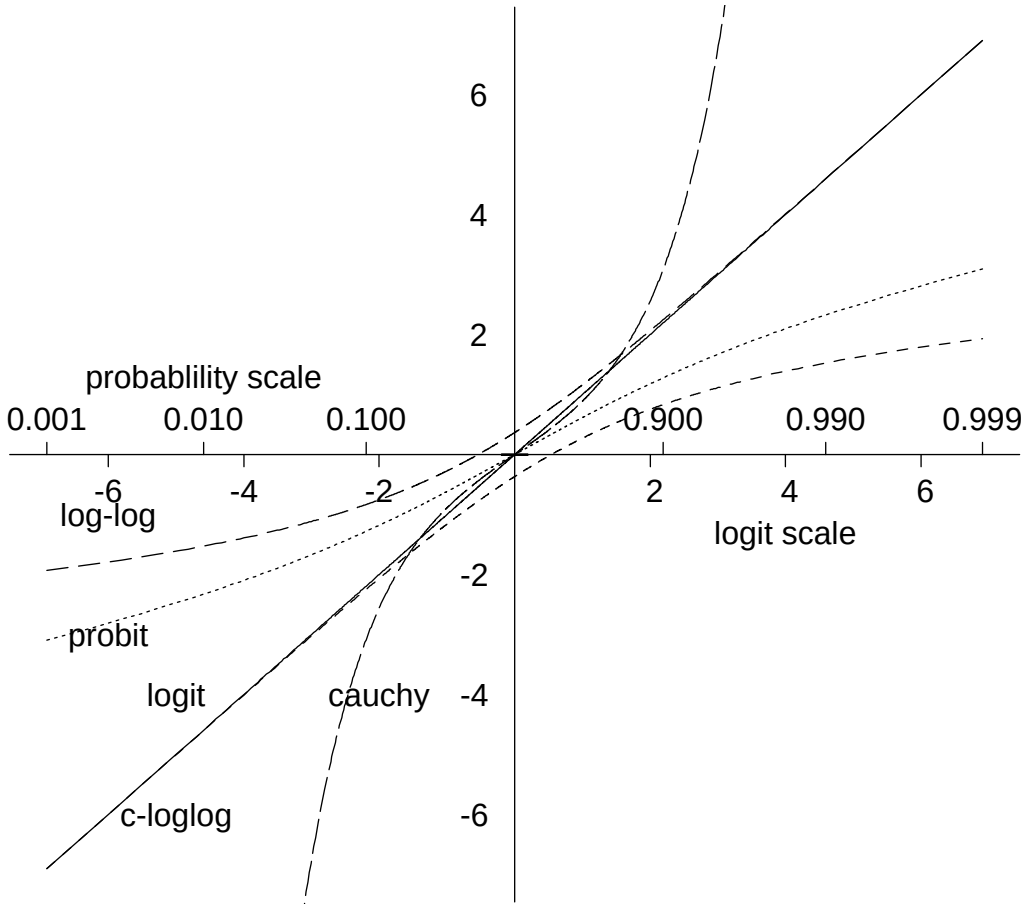

FIGURE 1. Comparison of five link functions: The horizontal axis is on the logistic scale so the logit link appears as the 45 degree line. Symmetry around the y-axis indicates symmetry of the distribution corresponding with the link as in the logit, probit and Cauchy cases. The log-log forms are asymmetric in this respect. Note that while the probit and logit are quite similar the Cauchy link is much more long tailed.

```
\#plots of link functions for binary dep models
postscript("fig1.ps",horizontal=F,width=6.5,height=6.5,
                        font=7,pointsize=12)
eps_.2
u_(1:999)/1000
plot(log(u/(1-u)),log(u/(1-u)),type="l",axes=F, xlab="",ylab="")

tics_c(.001,.01,.1)
tics_c(tics,1-tics)
ytics_0*tics
segments(log(tics/(1-tics)),ytics,log(tics/(1-tics)),ytics+eps)
text(log(tics/(1-tics)),ytics+3*eps,paste(format(round(tics,3))))
text(log(tics[2]/(1-tics[2])),1.3,"probablility scale")

tics_c(2,4,6)
tics_c(tics,-tics)
ytics_0*tics
segments(tics,ytics,tics,ytics-eps)
text(tics,ytics-3*eps,paste(format(round(tics))))
text(tics[2],-1.3,"logit scale")

segments(-eps,ytics,eps,ytics)
text(-3*eps,tics,paste(format(round(tics))))
abline(h=0)
abline(v=0)
lines(log(u/(1-u)),qnorm(u),lty=2)
lines(log(u/(1-u)),log(-log(1-u)),lty=3)
lines(log(u/(1-u)),-log(-log(u)),lty=4)
lines(log(u/(1-u)),tan(Pi*(u-.5)),lty=5)
text(-c(6,6,5,5,2),-c(1,3,4,6,4),c("log-log","probit","logit","c-loglog","cauchy"))
frame()
```

Interpretation of the coefficients

In regression we are used to the idea that

$$\frac{\partial E(y|x)}{\partial x_i} = \beta_i$$

provided we really have a linear model in x_i, but under our symmetry assumption here the situation is slightly more complicated. Now,

$$E(y_i|x_i) = 1 \cdot P(y_i = 1) + 0 \cdot P(y_i = 0) = F(x_i\beta)$$

so now

$$\frac{\partial E(y|x)}{\partial x_j} = f(x'\beta)\beta_j$$

for logit we have

$$F(z) = \frac{e^z}{1 + e^z}$$

so

$$f(z) = F(z)(1 - F(z))$$

while for probit we have

$$f(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

and for Cauchy

$$f(z) = \frac{1}{\pi(1 + z)^2}$$

We can compare these, for example, at $z = 0$ where we get

	factor from f(0)
logit	$1/4$
probit	$1/\sqrt{2\pi}$
Cauchy	$1/\pi$

and roughly speaking the whole $\hat{\beta}$-vector should scale by these factors so e.g.,

$$\frac{1}{4}\beta_j^{\text{logit}} \approx \frac{1}{\sqrt{2\pi}} b_j^{\text{probit}}$$

so

$$\beta_j^{\text{logit}} \approx 1.60\beta_j^{\text{probit}}$$

Diagnostic For the Logistic Link Function

Let $g(p) = \text{logit}(p)$ in the usual one observation per cell logit model, and suppose we've fitted the model

$$\text{logit}(p_i) = X\beta$$

but we'd like to know if there is some more general form for the density which works better. Pregibon suggests, following Box-Cox,

$$g(p) = \frac{p^{\alpha-\delta}-1}{\alpha-\delta} - \frac{(1-p)^{\alpha-\delta}-1}{\alpha+\delta}$$

note as $\alpha, \delta \to 0$ we get

$$\begin{aligned} &= \log p - \log(1-p) \\ &= \log(p/1-p). \end{aligned}$$

$\delta = 0 \Rightarrow$ symmetry, α governs fatness of tails. Expanding g in α, δ we get (with diligence)

$$\begin{aligned} g(p) &= g_0(p) + \alpha g_0^{\alpha}(p) + \delta g_0^{\delta}(p) \\ g_0^{\alpha}(p) &= \frac{1}{2}[\log^2(p) - \log^2(1-p)] \\ g_0^{\delta}(p) &= -\frac{1}{2}[\log^2(p) + \log^2(1-p)] \end{aligned}$$

LM tests of significance of g^{α}, g^{δ}, in an expanded model in which we include $g_0^{\alpha}(\hat{p})$ and $g_0^{\alpha}(\hat{p})$ where these variables are constructed from a preliminary logistic regression, can be used to evaluate the reasonableness of the logit specification.

Semiparametric Methods for Binary Choice Models

It is worthwhile to explore what happens when we relax the assumptions of the prior analysis, in particular the assumptions that a.) we know the form of the df F, and b.) that the u_i's in the latent variable formulation are iid. Recall that in ordinary linear regression we can justify OLS methods with the minimal assumption that u is mean independent of the covariates x, i.e., that $E(u|x) = 0$.

We will see that this condition is *not* sufficient to identify the parameters β in the latent variable form of the binary choice model. The following example is taken from Horowitz (1998). Suppose we have the simple logistic model,

$$y_i^* = x_i\beta + u_i$$

where u_i is iid logistic, i.e., has df

$$F(u) = 1/(1+e^{-u})$$

It is clear that multiplying the latent variable equation through by σ levels observable choices unchanged, so the first observation about identification in this model is that we can only identify β "up to scale". This is essentially the reason we are entitled to impose the assumption that u has a df with known scale. Now let γ be another parameter vector such that $\gamma \neq \sigma\beta$ for any choice of the scalar σ.

It is easy to construct new random variables, say v, *whose dfs will now depend upon* x, and for which

$$(*) \qquad F_{v|x}(x'\gamma) = 1/(1+\exp(-x\beta))$$

and

$$E(v|x) = 0$$

Thus, γ and the v's would generate the same observable probabilities as β and the u's and both would have mean independent errors with respect to x.

The argument is most easily seen by drawing a picture. Suppose we have the original (β, u) model with nice logistic densities at each x, the and a line representing $x\gamma$ and we could imagine recentering the logistic densities so that they were centered with respect to the $x\gamma$ line. Now on the left side of the picture imagine stretching the right tail of the density until the mean matches $x\beta$, similarly we can stretch the left tail an the right side of the picture – as long as the stretching doesn't move mass across the $x\gamma$ line (*) is satisfied.

What this shows is that mean independence is the wrong idea for thinking about binary choice models. What *is* appropriate? The example illustrates that the right concept is *median independence*. As long as

$$\text{median}(u|x) = 0$$

we do get identification under two rather mild conditions.

A simple way to see how to exploit this is to recall that under the general quantile regression model,

$$Q_y(\tau|x)\ x\beta$$

equivariance to monotone transformations implies that for the rather drastic transformation $I(x > 0)$ we have

$$Q_{I(y>0)}(\tau|x) = I(x\beta > 0)$$

but $I(y > 0)$ is just the observable binary variable so this suggests the following estimation strategy

$$\min_{||\beta||=1} \sum \rho_I(y_i - I(x_i\beta > 0))$$

where y_i is the binary variable. This problem is rather tricky computationally but it has a natural interpretation – we want to chose β so that as often as possible $I(x_i\beta > 0)$ predicts correctly.

Manski (1975) introduced this idea under the rather unfortunate name "maximum score" estimator, writing it as

$$\max_{||\beta||=1} \sum (2y_i - 1)(2I(x_i'\beta \geq 0) - 1)$$

In this form we try to maximize the number of matches, rather than minimizing the number of unmatches but the two problems are equivalent. The large sample theory of this estimator is rather complicated, but an interesting aspect of the quantile regression formulation is that it enables us by estimating the model for various values of τ to explore the problem of "heteroscedasticity" in the binary choice model.

Discrete Choice Models – Some Theory

The theory of discrete choice has a long history in both psychology and economics. McFadden's version of the Thurstone (1927) model may be very concisely expressed as follows:

m choices

$y_i^* =$ utility of i^{th} choice

$$y_i = \begin{cases} 1 & \text{if } y_i^* = \max\{y_1^*, \ldots, y_m^*\} \\ 0 & \text{otherwise} \end{cases}$$

Suppose we can express the "utility of the i^{th} choice" as,

$$y_i^* = v(x_i) + u_i$$

where x_i is a vector of attributes of the i^{th} choice, and $\{u_i\}$ are iid draws from some df F. Note that in contrast to classical economic models of choice, here utility has a random component. This randomness has an important role to play, because it allows us to develop simple models with "common tastes" in which not everyone make exactly the same choices.

Thm. If the u_i are iid with $F(u) = F(u_i < u) = e^{-e^{-u}}$, then $P(y_i = 1|x_i) = \frac{e^{v_i}}{\sum e^{v_i}}$ where $v_i \equiv v(x_i)$.

Remark. $F(\cdot)$ is often called the Type 1 extreme value distribution.

Proof. $y_i^* = \max\{\sim\} \Rightarrow u_i + v_i > v_j + v_j$ for all $j \neq i$ or $u_j < u_i + v_i - v_j$. So, conditioning on u_i and then integrating with respect to the marginal density of u_i,

$$P(y_i^* = 1|x_i) = \int \prod F(u_i + v_i - v_j) f(u_i) du_i$$

Note if $F(u)$ takes the hypothesized form, then $f(u) = e^{-e^{-u}} \cdot e^{-u} = e^{-u-e^{-u}}$ so

$$\begin{aligned} \prod_i F(u_i + v_i - v_j) f(u_i) &= \prod_j \exp(-\exp(-u_i - v_i + v_j)) \exp(-u_i - \exp(-u_i)) \\ &= \exp(-u_i - e^{-u_i}(1 + \sum_{j \neq i} \frac{e^{v_j}}{e^{v_i}})) \end{aligned}$$

let

$$\lambda_i = \log(1 + \sum_{j \neq i} e^{v_j}/e^{v_i}) = \log(\sum_{j=1}^{m} e^{v_j}/e^{v_i})$$

so

$$\begin{aligned} P(y_i^* = 1|x_i) &= \int \exp(-u_i - e^{-(u_i - \lambda_i)}) du_i \\ &= e^{-\lambda_i} \int \exp(-\tilde{u}_i - e^{-\tilde{u}_i}) d\tilde{u}_i \qquad \tilde{u}_i = u_i - \lambda_i \\ &= e^{-\lambda_i} \\ &= \frac{e^{v_i}}{\sum_{j=1}^{n} e^{v_j}} \end{aligned}$$

Extensions:

$$\begin{aligned} y_{ij}^* &= \text{utility of } i^{\text{th}} \text{ person for } j^{\text{th}} \text{ choice} \\ y_{ij}^* &+ x_{ij}\beta + z_i\alpha + u_{ij} \end{aligned}$$

Then assuming u_{ij} iid F yields

$$p_{ij} = P(y_{ij} = 1 | x_{ij,z_i}) = \frac{e^{x_{ij}\beta + z_i\alpha}}{\sum e^{x_{ij}\beta + z_i\alpha}}$$

E.g., here x_{ij} is a vector of choice specific individual characteristics like travel time to work by the i^{th} mode of transport, and z_i is a vector of individual characteristics, like income, age, etc.

Critique of IIA – Independence of Irrelevant Alternatives

Luce derived a version of the above model from the assumption that the odds of choosing alternatives i and j shouldn't depend on the characteristics of a 3^{rd} alternative k. Clearly here

$$\frac{P_i}{P_j} = \frac{P(y_i = 1)}{P(y_j = 1)} = \frac{e^{v_i}}{e^{v_j}}$$

which is independent of v_k. One should resist the temptation to relate this to similarly named concepts in the theory of voting. For some purchases this is a desirable feature of choice model, but in other circumstances it might be considered a "bug." Debreu in a famous critique of the IIA property suggested that it might be unreasonable to think that the choice between car and bus transportation would be invariant to the introduction of a new form of bus which differed from the original one only in terms of color. In this red-bus-blue-bus example we would expect that the draws of u_i's for the two bus modes would be highly correlated, not independent. Recently, there has been considerable interest in multinomial probit models of this type in which correlation can be easily incorporated.

References

Pregibon, D. (1980) Goodness of link tests for generalized linear models, *Applied Stat*, 29, 15-24.