University of Illinois
Fall 2012

Department of Economics
Roger Koenker

Economics 508
Lecture 20

Binary Response 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

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

These so-called link functions are illustrated in Figure 1. In this figure the scale is logistic so the logit link appears as a straight line. Probit assumes shomewhat thinner tails than the logit, while Cauchy assumes much fatter tails. The two log-log links asymmetric.

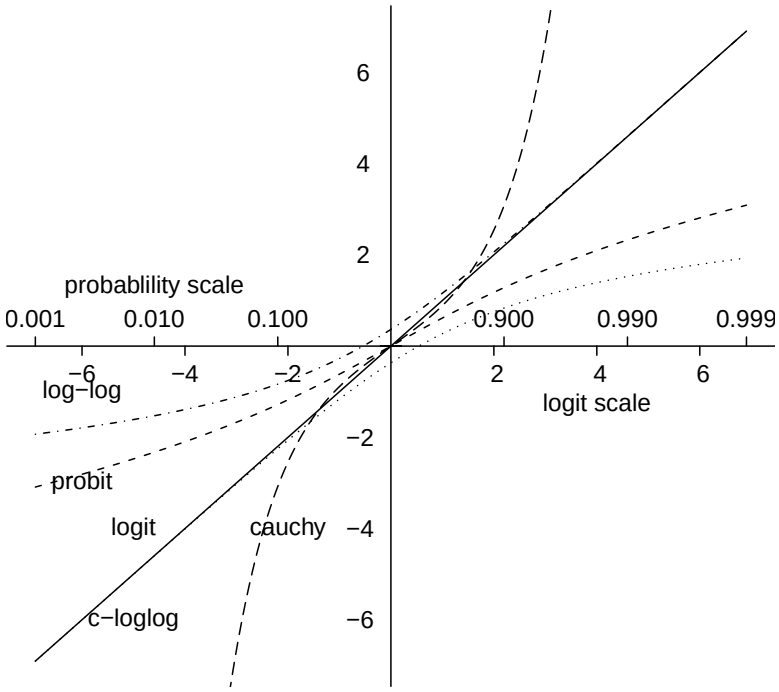

FIGURE 1. Comparison of five link functions: The horizontal axis is on the logistic scale so the logit link appears as the 45 degree line. Symmetry around the y-axis indicates symmetry of the distribution corresponding with the link as in the logit, probit and Cauchy cases. The log-log forms are asymmetric 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
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"))
```

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,

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

(provided of course we have symmetry) 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.,

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

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 (1980) 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.

For a long time it was difficult to know what to suggest to do when this diagnostic failed to vindicate the logit specification, but earlier this year Jungmo Yoon and I took it upon ourselves to explore some new estimation methods for binary response with these links. This work is described in the papers available from

`http://www.econ.uiuc.edu/~roger/research/links/links.html`

Digression on Computation of the MLE

I will try to explain briefly the strategy for computing the MLE in binary response models. The general strategy is based on Newton's method. The fundamental idea is central to much of applied mathematics. Suppose we want to maximize the function G over $x \in \Re^p$, and suppose we have an initial guess, x_0, of the maximizing value. We proceed by defining a quadratic approximation

$$\tilde{G}(x) = G(x_0) + (x - x_0)'\nabla G(x_0) + \frac{1}{2}(x - x_0)'\nabla^2 G(x_0)(x - x_0)$$

Our strategy, that is to say Newton's strategy, will be to maximize $\tilde{G}(x)$ at each step and hope that a sequence of such iterations will bring us to the maximum. To maximize $\tilde{G}(x)$ we differentiate to obtain

$$\nabla G(x_0) + \nabla^2 G(x_0)(x - x_0) = 0$$

solving, we obtain,

$$x = x_0 - [\nabla^2 G(x_0)]^{-1}\nabla G(x_0).$$

Iterating in this manner, that is finding x replacing x_0 by it and continuing leads eventually to a local maximum (or minimum) of G. Suppose the function G is strictly concave so it has a unique maximum, and $\nabla^2 G$ is globally negative definite. Then, the Newton step is always a direction of ascent and since we are always going up we (eventually) get to the maximum. In more complicated situations the algorithm has many variants, but I'll resist the temptation to delve into these here.

In the binary response model the function G is the log likelihood function

$$\ell(\beta) = \sum_{i=1}^{n} y_i \log(F(x_i'\beta)) + (1 - y_i)\log(1 - F(x_i'\beta))$$

So, setting $F_i = F(x_i'\beta)$, and $f_i = f(x_i'\beta)$,

$$\begin{aligned}\nabla\ell(\beta) &= \sum_{i=1}^{n} \frac{y_i}{F_i} f_i x_i - \frac{1-y_i}{1-F_i} f_i x_i \\ &= \sum_{i=1}^{n} \left(\frac{y_i - F_i}{F_i(1-F_i)} \right) f_i x_i\end{aligned}$$

and

$$\nabla^2\ell(\beta) = -\sum \frac{f_i^2}{F_i(1-F_i)} x_i x_i' + (y_i - F_i)[mess]$$

where [mess] denotes several terms that are all multiplied by $(y_i - F_i)$. At this point we invoke a clever trick that is, I believe attributable to R.A. Fisher, called the "method of scoring." In effect, the trick is to simply replace $\nabla^2\ell(\beta)$ in Newton's method with the expectation of $\nabla^2\ell(\beta)$. The beauty of this trick in the present circumstance is that it wipes out the annoying [mess]. This is because $Ey_i = F_i$, of course.

Note that the matrix,

$$E\nabla^2\ell(\beta) = -\sum \left(\frac{f_i^2}{F_i(1-F_i)} \right) x_i x_i'$$

is necessarily negative definite since the weights $f_i^2/(F_i(1-F_i))$ are positive, and the terms $x_i x_i'$ are nonnegative definite.

Now consider the method of scoring step,

$$\hat{\beta}^{(i)} = \hat{\beta}^{(i-1)} + (X'WX)^{-1}X'Wr$$

where $W = \text{ diag}(f_i^2/(F_i(1-F_i)))$ and r is the vector with elements $((y_i - F_i)/f_i$, so it is a rescaled residual. We may interpret the iteration as a modified Newton step in which each step is simply a weighted least squares estimate. To see this it may help to recall that we may write the unweighted least squares estimator as

$$\hat{\beta} = (X'X)^{-1}X'y = (\sum x_i x_i')^{-1} \sum x_i y_i$$

The reader should check the linear algebra of this, if it isn't immediately apparent.

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 leaves 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, and a line representing $x'\gamma$. 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. After the stretching we have conditional median of the latent response lined up on the $x'\gamma$ line, and the conditional means lined up on the $x'\beta$ line. But now we have two different models with the same observable probabilities, but different parameters and of course different underlying error distributions, both with the same conditional mean function.

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 $I(y > 0)$ 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

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

where y_i is the binary variable and the function $\rho_\tau(u) = u(\tau - I(u < 0))$ is the usual quantile regression check function. 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. See Kordas (2000) for further details.

Endogoneity in Binary Response

As we suggested earlier the linear probability model is one way to proceed toward 2SLS methods for binary response. However, this has many difficulties and only one advantage – that it is mindless. I'll briefly sketch an alternative approach that illustrates the advantage of the control variate interpretation of 2SLS. Suppose that we have the model

$$y_{i1}^* = x_i^\top \beta + y_{i2}\gamma + u_i$$

$$y_{i2}^* = z_i^\top \delta + v_i.$$

For the sake of argument, suppose that the pair, (u_i, v_i) are jointly normal, and as usual we get to see only $y_{i1} = I(y_{i1}^* > 0)$. What to do? without loss of generality we can take $\sigma_u^2 = 1$, and we can write

$$u_i = \theta v_i + e_i$$

with $e_i \sim \mathcal{N}(0, 1 - \rho^2)$ where $\rho = \mathrm{Cor}(u_i, v_i)$. Why? Note that $e \perp z$ and $e \perp v$ and therefore $e \perp y_2$. Now $\theta = \eta/\tau^2$ where $\eta = \mathrm{Cov}(u_i, v_i)$ and

$\tau^2 = \mathbb{V}(v)$, and consequently,

$$\begin{aligned}\mathbb{V}(e) &= \mathbb{V}(u) + \theta^2\mathbb{V}(v) - 2\theta\text{Cov}(u,v)\\ &= 1 + \eta^2/\tau^2 - 2\eta^2/\tau^2\\ &= 1 - \rho^2\end{aligned}$$

so we can replace u_i in the first equation by $\theta v_i + e_i$ to get,

$$y_{i1}^* = x_i^\top \beta + y_{i2}\gamma + \theta v_i + e_i,$$

of course we don't know v_i, but we have a *plan* for that. Note that $e_i \sim \mathcal{N}(0, 1-\rho^2)$ so if we did have a v_i in our hand, we would be able to estimate the rescaled coefficients: $\beta/(1-\rho^2$, etc. The plan is the following two step procedure suggested by Vuong:

(1) estimate $\hat{v}_i$ from the second equation by OLS,
(2) estimate our new version of the probit model with $\hat{v}_i$ for v_i.

This is a natural variant of the control variate version of 2SLS. Note however that the naive version of 2SLS in which y_2 is replaced by $\hat{y}_2$ *does not* work for this binary response setting. In the last few years, Chesher and others have developed semiparametric approaches to these problems, albeit with complications that lead us out of the simple world of point identification and into the world of set identified models and inference.

Ratings and Rankings

A standard model for ranking competitive sports teams is the Bradley-Terry or paired comparison logit model. Under this model the probability, π_{ij}, of team i defeating team j is

$$\text{logit}\ (\pi_{ij}) = \alpha_i - \alpha_j$$

Given data on n games occurring among m teams we can construct an $n \times m$ "design" matrix with k^{th} row having i^{th} element 1, j^{th} element -1 and all other elements zero, assuming the k^{th} contest was between teams i and j. Given the estimated ratings we can easily compute an estimate of the probability π_{ij}:

$$\pi_{ij} = \frac{e^{\alpha_i - \alpha_j}}{1 + e^{\alpha_i - \alpha_j}}$$

An elaboration of this model using quantile regression methods is described in Koenker and Bassett(2010) where the application is to forecasting the outcome of the NCAA basketball tournament. The novelty here is that the QR results allow the investigator to make density forecasts for the score of the various games.

Another interesting generalization of this model involves *dynamic* updating of ratings. A leading example is the Elo system of chess ratings. In the

Elo system decent beginners have a rating around 1000 and Grandmasters have ratings in the range 2500-2900.

Simplifying somewhat, if A plays B with initial ratings R_A and R_B so A's probability of a win is π_{AB} then his new rating after a win is

$$R'_A = R_A + K(1 - \pi_{AB})$$

where K is a factor that is set at 16 for masters and 32 for weaker players. If he loses the 1 in the above formula is replaced by 0. In cases of a draw it is replaced by .5.

This provides a nice simple example of a dynamic logit model. Time series models for discrete random variables is a very interesting and challenging topic. A nice reference is MacDonald and Zucchini.

An amusing subliterature on chess ratings has looked at the evolution of chess ratings over the life cycle as a way to explore the inevitable deterioration of mental faculties.

Panel Data with Discrete Response

Panel data with discrete response is an important class of models. I'll only very briefly mention some basic ideas, for a more detailed discussion see Wooldridge's text. Suppose we have a latent variable y^*_{it} arising from the usual panel data model

$$y^*_{it} = x_{it}\beta + \alpha_i + u_{it}$$

and we observe, as usual, $y_{it} = I(y^*_{it} \geq 0)$. Ignoring the possibility of dynamics, possible dependence in u_{it}'s, over time, etc. we can consider both random and fixed effect estimators based on maximum likelihood and penalized maximum likelihood. Computation is somewhat challenging and may require sparse algebra methods when n is large.

As in the classical panel data setting there may be questions about whether random effects treatment induces bias in β due to correlation of α's with x's. An ingenious strategy for dealing with this, introduced by Chamberlain (1980) is to assume that

$$\alpha_i | x_i \sim \mathcal{N}(\alpha_0 + \bar{x}'_i\gamma, \sigma^2_\alpha)$$

then the latent variable model becomes

$$y^*_{it} = \alpha_0 + \bar{x}'_i\gamma + x_{it}\beta + \xi_i + u_{it}$$

where $\xi \perp x_i$. This is attractive since it dramatically reduces the number of estimated parameters. Chamberlain proposes a GMM strategy for estimation:

(i) Estimate $\theta_t = (\alpha_0, \beta', \gamma')'_t$ using $t = 1, \ldots, T$ cross-sectional probits,

(ii) Using GMM combine these estimates using a suitable weighting matrix based on the results of stage (i).

Although it is frequently desirable to estimate dynamic models of this type in which the probability of the event $y_{it} = 1$ depends on past values

of y_{it}, such models are considerably more difficult, and therefore will be neglected here. Again, there is some discussion in Wooldridge.

Multinomial Choice: An Introduction

Obviously binary response is just a special case of a much large class of situations in which we have several discrete categories and would like to model the probabilities of falling into each category. There are many possible variants on these models. I will briefly describe two of these in the remainder of this lecture. Both are addressed to situations in which we have no cardinal scale for the response. (Frequently, we may want to model data like the number of hospital visits or or the number of patents, but this is usually done via count models like poisson regression.)

In some circumstances we have a naturally ordered set of discrete responses such as: strongly disagree, disagree, agree, or strongly agree, which have no cardinal interpretation, but nevertheless have ordinal meaning. In such cases we can specify a latent variable model

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

and in the simplest case we can assume that there are thresholds that y_i^* must cross to put the observed y_i's into each category. Thus, we may assume:

$$y_i = j \quad \text{if} \quad c_{j-1} \leq y_i^* \leq c_j$$

where the cutoffs $c_j : \ j = 0, 1, ..., m$ and we take $c_0 = -\infty$ and $c_m + 1 = \infty$. Given the model, and an iid error assumption

$$P(y_i = j|x_i) = F(c_j - x_i'\beta) - F(c_{j-1} - x_i'\beta)$$

and this immediately yields a likelihood function that we can maximize over the vector β and the vector of cutoffs c. Note that this setup is very simple and it restricts drastically the way in which the covariates can influence the choices. It would be nice to have some semiparametric alternatives to this model, but there is no well established candidate for this at the moment.

We will now turn to the case in which we have non-ordered alternatives.

Random Utility Discrete Choice Models – Some Economic Theory

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

$$\begin{aligned} &m \text{ choices} \\ &y_i^* = \text{utility of } i^{\text{th}} \text{ choice} \\ &y_i = \begin{cases} 1 & \text{if } y_i^* = \max\{y_1^*, \ldots, y_m^*\} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

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:

Suppose

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

Then assuming u_{ij} iid F yields

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

E.g., here x_{ij} is a vector of choice specific individual characteristics like travel time to work by the j^{th} mode of transport, and z_i is a vector of individual characteristics, like income, age, etc. Note that the individual characteristics are assumed to have choice specific coefficient vectors – how age, for example, influences utility for various choices depends on the choice.

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.

Gibbs Sampling for Multivariate probit Model

Markov Chain Monte Carlo (MCMC) has transformed Bayesian statistics from a sleepy theoretical/philosophical hamlet to a bustling modern city. The basic ideas are simple but get complicated quickly. Here is a brief tourist guide. Recall that if

$$\begin{aligned} y_t &= \alpha + \rho y_{t-1} + u_t \\ u_t &\sim \mathcal{N}(0, \sigma^2) \\ |\rho| &< 1 \end{aligned}$$

then starting from any initial condition y_0, as t gets large

$$(*) \qquad y_t \sim \mathcal{N}\left(\frac{\alpha}{1-\rho}, \frac{\sigma^2}{1-\rho^2}\right)$$

Suppose we didn't know this deep fact, but somehow decided we'd try to simulate it: so we get out our iphone, rev up the rand key and get busy – eventually we have a large sample of y's and if, we plot them low and behold they make a histogram like $(*)$ wowy-zowy! Try it you'll like it.

$$y \leftarrow \text{ filter } (2 + \text{ rnorm } (10,000), .8, \text{ method } = \text{“ recursive”}) hist(y)$$

Gibbs sampling – under some conditions – discussed in 574 – we can simulate, by iterating back and forth between conditions.

Ex. (Probit)

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

1. Initialize β
2. Generate u vector $u \sim \mathcal{N}(0, I_n)$
3. Generate y_i^*'s from $N(x_i\beta, 1)$ but truncated on the left at zero if $y_i = 1$, and truncated on the right at zero if $y_i = 0$.
4. Generate $\beta \sim \mathcal{N}(\hat{\beta}, (X'X)^{-1})$
5. Go to 1.

An easy way to think about step 3 is that you know, because you observe the original binary y_i's that we should get a draw from u_i that would agree with the observed value. Thus, we could just generate y^*'s for each observation until we got one that was on the right side of zero and therefore agreed with the observed data.

Do this for a few thousand iterations and then look at the tail of the seqence of β's so generated. The empirical distribution of these generated β's will have the approximate distribution of the mle estimator.

Ex. (Multivariate Probit)

Generalize to SUR model for m latent y_i^*'s with correlation matrix Ω. Now, we need at each iteration to generate a vector of u's with a particular covariance matrix, but this is relatively easy, especially by comparison with computing the orthant probabilities required for the conventional mle. See Chib and Greenberg(1998).

Warning: It is all too easy to write pairs of conditions like this that are *incompatibly* proper, but so that there is no joint distribution with those conditionals.

Discrete Choices with a Profusion of Choices

The WWW has brought us choice problems with a vast number of choices; these are challenging and have also led to significant advances in

numerical linear algebra and statistical methods. I will discuss two highly visible examples.

Google Pagerank[1]. In the early 1990's when the web was young there were several search engines, but they typically produced poor results primarily because they made poor orderings of the matched results. Then Google appeared and mopped the floor, how?

The simple version of the answer is Google's pagerank algorithm. Like any search engine Google created a large hash table of words and phrases from all the public web pages by automated "web-crawling," so it could match search inquiries. But how to order the resulting matching? You would like to rank pages higher if they are more likely to be interesting, and they are more likely to be interesting if they are linked to other interesting pages.

The structure of the web can be viewed as a large directed graph, represented by a matrix G whose ij entry is one if page j links to page i and zero otherwise. This is a large matrix, but very sparse, that is most of its entries are zero. So our question becomes: Given G what is a good index of "interestingness?" There are some obviously bad ideas that spring to mind:

- Count the number of pages that link to page i.
- Sum the ratings of pages that link to page i.

Not all pages are created equal so counting pages is clearly silly, but summing ratings also creates opportunities for pages that have many links to unduly influence rankings. We don't want to create perverse incentives for irrelevant linkages since they are very easy to create. If we modify the second idea slightly so that, $r_i = \sum_{j \in L_i} r_j/n_j$ where L_i is the set of page indices linking to page i and n_j is the number of links made by page j, we can define a new matrix, A with entries,

$$A_{ij} = G_{ij}/n_j.$$

If we rewrite this idea in matrix form, we have, $r = Ar$ or $0 = (A - I)r$, so it appears that we have defined the ratings vector as an eigenvector of the matrix A and claimed that A has an associated eigenvalue of one. How do we know that there *is* such an eigenvector?

A Random Walk through the WWW

Imagine picking a random page on the web and then taking a random walk through the web by picking *randomly* a link from each successive page you visit. If A is "nice" then r_i would be the fraction of time you spent at page i, or the proportion of times you visited page i in the (very) long run. This is the stationary distribution of the Markov chain describing the random walk. Really? What do we mean by "nice?" What can go wrong if A isn't "nice?"

[1]This section is based largely on Candès (2009) and Bryand and Leise (2006)

One thing that can cause problems is obviously pages that don't have any links; they would bring our walk to an abrupt end. We can assume that any page is linked to itself so a page without any other links would leave us stuck at that page, in the usual jargon this would be an absorbing state of the chain. Another problem would be pages that led us around in a circle, or cycle.

Definition A matrix A is Markov if $A_{ij} \geq 0$ and its columns all sum to one, i.e. $\sum_i Aij = 1$ for all j.

Lemma If A is Markov then it has an eigenvalue 1.

Proof A and $A^\top$ have the same eigenvalues – recall that eigenvalues solve the characteristic equation $|A - \lambda I| = 0$, and determinants of transposes are the same as determinants of the untransposed matrix. By the Markov property $A^\top 1 = 1$ so A has an eigenvalue 1.

We now need to ensure that unlinked or otherwise problematic web structures don't cause problems. Google's strategy for this is to replace A by

$$B = (1 - \delta)A + \delta 11^\top / n \quad \delta \in (0, 1),$$

This modifies our random walk strategy so that at each page we either do what we described before with probability $(1 - \delta)$, or with probability δ we take a flier to a completely random page. Google supposedly uses something like $\delta = 0.15$. This ensures that we can get anywhere from anywhere in our random peregrinations.

Theorem [Perron-Frobenius] If B is Markov and $B_{ij} > 0$ for all i, j, then B has largest eigenvalue one and corresponding unique eigenvector with positive components.

Rather than normalizing the eigenvector as usual so that $\|x\| = 1$ we can choose r so that $r_i = x_i / \sum_j x_j$, since the entries are positive. This allows us to interpret r as a discrete probability distribution, the stationary distribution of our random walk.

This is all very well, but how are we going to compute this eigenvector when A has something like a 100 billion rows and columns? An old strategy that seems quite slow and inefficient on smaller problems is about all that we have to rely on for problems of this size. This is the power method:

Initialize $x^{(0)}$
While($\|x^{(i)} - x^{(i-1)}\| > \epsilon$)
  $x^{(i+1)} = Bx^{(i)} / \|Bx^{(i)}\|$

The beauty of this in the present case is that multiplication by B is reasonably quick since B is mostly zeroes, and multiplication by the matrix of ones, $11^\top$ is also quick since we have good hardware for summing. This is done daily using yesterday's x as an initial value, actually it is done essentially continuously on a vast array of distributed machines. Google has a market capitalization of $198B, as of November 15, 2011, so don't let people tell you that eigenvectors aren't worthwhile.

Matrix Completion and the Netflix Competition

The second massive discrete choice problem I want to consider involves matrix completion. It is conveniently illustrated by the recent Netflix prize competition. Netflix rents movies to subscribers, who (sometimes) rate the movies on a scale of 1 to 5. We can think of the resulting ratings as a (large) $m \times n$ matrix with movie row labels and subscriber column labels. The ij entry would be the rating that subscriber j gives to movie i. Obviously, most of the entries of the matrix are missing; our task is to find a good way to fill in these missing entries, that is to find a way to predict ratings for movies that subscribers *have not* seen, or have seen but not rated.

In 2006 Netflix offered a prize of $1M to any individual or group who could improve upon their current algorithm by 10 percent. Contestants were provided with a data matrix with $m = 17,770$ movies and $n = 480,189$ users and roughly 100 million ratings, so the matrix was about 99 percent empty. This so-called training data was supposed to be used to predict ratings in a withheld evaluation data set with predictive performance measured by that old standby – root mean squared error.

Using regression based ideas Netflix default methods had achieved a rmse of 1.054. In 2009 a consolidated team composed of three early competing groups won the prize. Competition for the prize was quite fierce and led to some real conceptual progress in convex optimization.

One formalization of the Netflix problem is the following: let A denote the $m \times n$ unknown full matrix, and $\mathcal{R}$ denote the set of indices for the complete entries of the matrix, denoted y_{ij}. We would like to find A to minimize,

$$f(A) = \sum_{ij \in \mathcal{R}} (y_{ij} - A_{ij})^2$$

A first idea is to consider matrix factorizations like $A = UV^\top$, for some (hopefully) low rank matrices, U and V. As a general matter, this problem is ill-posed, i.e. not identified. There are lots of ways to choose U and V to achieve $f(A) = 0$ so we need some sort of regularization or penalty. One penalty that plays a similar role as the lasso regression penalty for such problems is formulated as,

$$\min_A \{f(A) | A \succeq 0,\ Tr(A) = 1\}$$

where $A \succeq 0$ means that A is positive semi definite, PSD. This formulation can in turn be reformulated as,

$$\min_A \{f(A) + \mu \|A\|_*\}$$

where $\|A\|_*$ is the nuclear norm of A, defined as the sum of the singular values of A.

Recall that an $m \times n$ real matrix A can be factored as,

$$A = UDV^\top$$

where D is a diagonal matrix of singular values, U is a $m \times m$ unitary matrix whose columns are the eigenvectors of $A^\top A$, and V is a unitary $n \times n$ matrix whose columns are the eigenvectors of $AA^\top$. The diagonal elements are square roots of the eigenvalues of $A^\top A$.

Matrix norms play an indispensible role in all of this and there are lots of them with a somewhat perplexing assortment of names and aliases. We will denote the (squared) Frobenius norm, $\|U\|_2^2 = Tr(U^\top U) = \sum\sum U_{ij}^2$, and is just the usual (squared) Euclidean norm of the matrix entries viewed as vector. Similarly, we have $\|U\|_1 = \sum\sum |U_{ij}|$ and $\|U\|_\infty = \max |U_{ij}|$. Likewise we can make matrix norms out of ℓ_p norms of the singular values of the matrix, so the nuclear norm is corresponding ℓ_1 norm of the singular values, the ℓ_∞ norm analogue is called the operator norm and is simply the largest singular value. In this literature this operator norm is sometimes (confusingly) referred to as the 2-norm. I'll denote it below by simply $\|\cdot\|$.

The linkage between the last two formulations rests on the following result.

Lemma $\|A\|_* = \min_{U,V}\{(\|U\|_2^2 + \|V\|_2^2)/2 \mid UV^\top = A\}$

The proof employs some convex analysis that I won't go into here, details can be found in Recht, Fazel and Parrilo (2010), but essentially boils down the duality of the nuclear and operator norms. Fortunately, there are good ways to solve such problems even those formulated on a very large scale. These methods can be viewed as gradient descent algorithms, uniqueness of solutions are ensured by the convexity of the problem – PSD matrices constituting a convex cone.

Recently Candès, Li, Ma, and Wright (2009) have used closely related techniques to create robust form of principle component analysis. In their setup we are given a large data matrix A and we would like to decompose it into the sum of a low rank component and a sparse component:

$$A = L + S.$$

In classical PCA we do something similar, except that instead of the sparse component we assume that S is a dense matrix modeled as Gaussian noise.

This is formulated as,

$$\min_L \{\|A - L\| \mid \text{rank}(L) \leq k\}$$

So we are trying to minimize the maximal singular value of the noise matrix computed after (optimally) choosing a rank k version of L. When S is sparse it seems better to try to encourage it to be sparse in the optimization by some sort of lasso like penalty. The formulation in Candès et al is:

$$\min_L \{\|L\|_* + \mu\|A - L\|_1\}$$

Again, we have a convex optimization problem that can be solved by methods like those used for matrix completion. These methods seem to offer a wide scope for applications to many problems that have proven to be resiliant to other approaches. Econometrics has tended to ignore developments in multivariate statistics, thus we sometimes have to rely on the Soren Johansen's of the world to clue us into the utility of these methods. This seems (somehow) unfortunate.

References

Bryan, K. and T. Leise, (2006) The $25,000,000,000 Eigenvector: The Linear Algebra behind Google. *SIAM Review*, 48, 569–581.

Chamberlain, G. (1980) Multivariate regression models for panel data, *J. of Econometrics*, 18, 5-46.

Candès, Li, Ma, and Wright (2009) Robust PCA, Technical Report, Stanford Statistics.

Candès, I. (2009) Stanford ACM 106a Lecture Slides, available from: `http://www-stat.stanford.edu/~candes/acm106a/Handouts/google.pdf`.

Chib, S. and E. Greenberg (1998) Analysis of Multivariate Probit Models, *Biometrika*, 85, 347–361.

Horowitz, J. (1998) *Semiparametric Methods in Econometrics*, Springer-Verlag, 1998.

Koenker, R. and Bassett Jr, G.W., (2010), March Madness, Quantile Regression Bracketology, and the Hayek Hypothesis, *Journal of Business and Economic Statistics*, 28, 26–35,

Kordas, G. (2000) *Quantile Regression for Binary Response*, UIUC Phd Thesis.

Manski, C. (1975) Maximum score estimation of the stochastic utility model of choice. *J. of Eonometrics 3* (1975), 205–228.

MacDonald, I.L. and Zucchini, W., 1997, *Hidden Markov and other models for discrete-valued time series*, Chapman & Hall.

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

Recht, B., M. Fazel, and P.A. Parrilo (2010) *SIAM Review*, 52, 471-501.