PS 271B: Quantitative Methods II
Lecture Notes
(Part 4: Categorical and Ordinal Data)

Langche Zeng
zeng@ucsd.edu
Models for Categorical and Ordinal Data

1. Binary Data: Logit and Probit

$Y_i$ can take 0 or 1. Bernoulli. $Y_i$ and $Y_j$ independent for $i \neq j$.

$Pr(Y_i = 1) = \pi,$

$Pr(Y_i = 0) = 1 - \pi \rightarrow$

$Pr(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$

Recall that $E(Y_i) = \pi_i$

*Logit*: Assumes $\pi_i = \frac{1}{1+e^{-x_i\beta}}$; *Probit*: Assumes $\pi_i = \Phi(x_i\beta)$, where $\Phi()$ is the normal cumulative density function.
A deeper, utility theory motivation, for the general case of mixed type of independent variables:

There are two alternatives, \( \{0,1\} \), for decision maker \( i \) to choose from (abstain/vote; peace/war, etc.) with utility functions

\[
    u_{ij} = x_i \beta_j + w_{ij} \gamma + \epsilon_{ij}, \quad (j = 0, 1; i = 1, 2, \ldots, n)
\]

where \( x_i \) are individual specific and do not vary w.r.t. alternatives, and \( w_{ij} \) are alternative specific specific covariates.

The decision maker maximizes utility, and chooses alternative 1 over 0 if

\[
    u_{i1} > u_{i0}, \text{ or } u_{i1} - u_{i0} > 0, \text{ or } x_i (\beta_1 - \beta_0) + (w_{i1} - w_{i0}) \gamma + (\epsilon_{i1} - \epsilon_{i0}) > 0, \text{ or } \\
    \epsilon^*_i > -(x_i \beta^* + w_i^* \gamma) = -z_i \theta
\]

where “*” denotes differenced version (so alternative specific variables enter the model as differenced values. e.g., diff in thermometer scores for two candidates.)
Thus, a model for \( Pr(Y_i = 1) \) is obtained if we assume a distribution for the “disturbances” \( \epsilon_i \) (and in turn \( \epsilon_i^* \)). Logit model results from the so-called “extreme value distributions”, and Probit results from assuming normal distributions. The two are nearly identical. For symmetrical distributions, \( Pr(\epsilon_i > -x_i \beta) = Pr(\epsilon_i < x_i \beta) \) (thus \( \Phi(x_i \beta) \) for the Probit.)

Some start with \( Y_i^* = x_i(\beta_1 - \beta_0) + (w_{i1} - w_{i0})\gamma + (\epsilon_{i1} - \epsilon_{i0}) = z_i \theta + \epsilon_i \) and call this the index or latent function for the observed \( Y_i \).

Draw logit/probit curve.

—Likelihood function for Logit:

\[
P(y|\pi) = \prod_i \pi_i^{y_i}(1 - \pi_i)^{1-y_i} \quad \longrightarrow \\
\ln L = \sum_i \{y_i \ln \pi_i + (1 - y_i) \ln(1 - \pi_i)\}
\]
\[
= \sum \{-y_i \ln(1 + e^{-x_i \beta}) - (1 - y_i) \ln(1 + e^{x_i \beta})\} \\
= - \sum \ln(1 + e^{(1-2y_i)x_i \beta})
\]

(Verify: what does the last expression become when \(y_i = 1\) or \(y_i = 0\)?)

MLE: numerical methods.

Probit likelihood can be similarly derived using the probit expression for \(\pi_i\). Note that probit error variance is 1, while that for logit is \(\pi^2/3\). Thus, dividing the logit latent equation by the sd (about 1.81) will make the variances equal. the estimated logit coefs are therefore roughly 1.8 that of the probit. the pr is not affected.

—Interpretation:
1) Marginal effects plots: Graph $Pr(Y = 1|X)$ as a function of one (or two) $X$’s, holding other $X$’s at some meaningful values.

2) fitted values: $Pr(Y = 1|X^*)$, $X^*$: some hypothetical profile, typically that of a “representative” individual.

3) First difference: $Pr(Y = 1|X_1) - Pr(Y = 1|X_0)$. dummy var: 0 to 1; continuous: typically 1 sd increase.

4) Risk ratio: $Pr(Y = 1|X_1)/Pr(Y = 1|X_0)$

5) Partial derivatives/marginal effects:

$$\frac{\partial Pr(Y=1|X)}{\partial X_j} = \beta_j \pi (1 - \pi)$$

where $\pi = Pr(Y = 1|X)$. (what value of $\pi$ makes the derivative the largest? motivation for scobit.)

“odds ratio”: unintuitive.
All the estimated quantities of interest involve *uncertainties*, as they are functions of the estimated parameters.

Uncertainty info can be obtained either analytically (e.g., the delta method–Taylor expand the function of coeffs, then take variance of the first order approximation) or through simulation. Simulation involves drawing $M$ random numbers from the distribution of the parameters (obtained either from asymptotic normal approximation, or as Bayesian posterior samples, or from Bootstrapping), which translate into $M$ values for a given quantity of interest. Use these $M$ values we can do histogram, mean/sd, CI’s, etc. for the quantity. (Zelig uses normal approximation or bootstrapping or posterior sampling.)

Note the difference between expected value and individual predicted value. The former does not involve fundamental uncertainty, the latter does. In logit, the former is a probability, the latter is binary, 0/1.
Do demo(vertci); marginal.effects.R (and plot.ci)

–Model evaluation

1) Residuals and influence

Various residual and influence measures can be obtained after model fitting, similar to the linear model case. The deviance residual is based on the individual point’s contribution to the deviance ($-2 \ln L$). The Pearson residual compares the observed $y$ (0 or 1) with the predicted probability, $\pi$, taking into consideration the s.d. of $y$: $r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$

These residuals can be standardized by the s.d. of $r_i$ itself. Graphing these residuals against observation number may reveal outliers. relevant R functions: residuals(), rstudent(), rstandard().

Influence.measures() computes an array of influence measures. Other func-
tions such as cooks.distance(), dfbetas(), hatvalues() provides direct access to various components. One way of measuring the influence of a data point is by the change in the estimated parameters when that point is omitted. (This in principle requires estimating the model $N$ times, but analytic approximations can be computed after estimating it once.) This is what dfbeta() does. cooks.distance() summarizes the effects on the entire vector of coefficients. hat value represents the potential of the point to influence the fit.

```r
inf<-influence.measures(z.out)
summary(inf)
head(inf$is.inf)
which(inf$is.inf[,10])
summary(inf$infmat[,10])
which(inf$infmat[,10]>0.025)
turnout[1595,]
which(apply(inf$is.inf,1,any))
plot(rstudent(z.out)~hatvalues(z.out))
```
plot(1:nrow(turnout), cooks.distance(z.out))
summary(dfbeta(z.out))
plot(density(residuals(z.out)))

See Faraway 6.4 for more details.

2) Scalar measures of goodness of fit

Single number summary of fit, in the spirit of $R^2$, or adjusted $R^2$. Various measures proposed, none having the clear interpretation of the linear model $R^2$, and no clear evidence of which ones are superior to others.

Information measures:

$$ AIC = -2 \ln \hat{L} + 2P $$

AIC is part of summary() output. $-2 \ln \hat{L}$ is the deviance of the model (deviance()), measuring how far the fit is from the best possible (for which
\(-2 \ln \hat{L} = 0\). So AIC is goodness of fit with penalty for model complexity (as reflected in \(P\), number of parameters.) AIC the smaller the better.

BIC (Bayesian information criterion) is similar in spirit:

\[
BIC = -2 \ln \hat{L} + P \ln N
\]

Can show that \(2 \ln \left[ \frac{Pr(D|M_2)}{Pr(D|M_1)} \right] \approx BIC_1 - BIC_2 \) where the Bayes factor, \(\frac{Pr(D|M_2)}{Pr(D|M_1)}\) is the posterior odds of the models, if we have no prior preference for one model against another. BIC smaller the better.

3) Goodness of fit should not be the criterion, at least not the only criterion, for evaluating models. Out of sample performance is what really matters.

A model could overfit and miss the underlying structure.
The possibility of over-fitting increases with model complexity. Generally, fitting gets better and better as the complexity increases, but out of sample performance worsens after certain point. [draw figure]

Make it a standard practice to reserve some of your data for out of sample evaluation.

Dividing data into in-sample/out-of-sample is a special case of “cross validation”, which involves dividing the data into $k$ parts, each time leave one part out in estimation and evaluate the model on the observations left out, then average the results over the $k$ passes. the extreme form is “leave one out”. Computationally intensive.

4) ROC curve:

If misclassifying a 1 is $C'$ times more costly than a 0 (e.g., war/peace), then decision theory says predict $Y = 1$ if $\hat{\pi} > 1/(1 + C')$. When $C' = 1$, the cut
point is the usual .5.

If unsure what value of $C$ to use, could draw the ROC curve, which plots % of 0’s correctly predicted against the % of 1’s correctly predicted, resulting from all sorts of cut point values in $[0, 1]$ (reflecting all sorts of $C$ values.) A better model has a larger area under the ROC curve. Ideally dominates the other model everywhere. (demo(roc) in Zelig; rocr package)

5) Calibration: Sort estimated probabilities into bins, say $[0, 0.1)$, $[0.1,0.2)$, ... compute the mean predicted Pr in each bin, and the fraction of 1’s actually observed in the corresponding observations. Plot the pairs. Systematic deviation from the 45 degree line indicates problem.
2. Ordinal data: Ordered Logit and Probit

Start with the model for the latent $Y^*$: $Y^* = x\beta + \epsilon$

Suppose the observed $Y$ can take $J$ ordered values, and suppose the observation mechanism is

$$Pr(Y = j) = Pr(\tau_{j-1} \leq Y^* < \tau_j), \ j = 1, 2, \ldots, J; \tau_0 = -\infty, \text{ and } \tau_J = \infty$$

Define $Y_{ij} = 1$ if $Y_i = j$, 0 otherwise, then the likelihood function is:

$$P(Y) = \prod_{i=1}^{n} \prod_{j=1}^{J} Pr(Y_{ij} = 1)^{y_{ij}}$$

(1)

where

$$Pr(Y_{ij} = 1) = Pr(Y_i = j) = Pr(\tau_{j-1} \leq Y_i^* < \tau_j)$$

$$= Pr(\tau_{j-1} \leq x_i\beta + \epsilon_i < \tau_j)$$
Assuming $\epsilon_i$ follows normal distribution, this becomes:

$$Pr(Y_{ij} = 1) = \Phi(\tau_j - x_i\beta) - \Phi(\tau_{j-1} - x_i\beta)$$

which is the **Ordered Probit** model. Ordered logit results from assuming the logistic distribution for the errors. Interpretation is done using these probabilities (which sum to 1, of course). In general, the sign of the marginal effects is unambiguous only for the smallest and the largest $Y$ values!

Inserting this into (1), we have the likelihood function as a function of the data $(y_{ij}, x_i)$ and the parameters $(\beta$, plus the $J - 2$ $\tau$’s—normalization gets rid of another $\tau$). MLE for the parameters are obtained by maximizing the (log) likelihood.

**Zelig:**

```r
z.out <- zelig(as.factor(Y) ~ X1 + X2, model = 'oprobit',
```
data = mydata)

(if \( Y \) is discrete integer with values reflecting the right order; otherwise first create an ordered factor, see Zelig manual on ordered probit for an example.)

Interpretation can be similarly done as for binary probabilities (fitted values, first difference, marginal effect plot, etc.)

* demo(ologit)

\textit{Parallel regression assumption}: Consider carefully the data observation mechanism assumed in the ordered probit/logit models: there is one underlying \( y^* \) whose values give rise to the observed (ordered) categorical values on \( y \). This assumption implies, in addition to the order probit/logit probability expression, also expressions for a group of cumulative probabilities: \( pr(Y <= j), j = 1, 2, \ldots, J - 1 \), which are given by:

\[
Pr(Y <= j) = Pr(Y^* < \tau_j) = Pr(x\beta + \epsilon < \tau_j) = Pr(\epsilon < \tau_j - x\beta) \quad (3)
\]
which means that if you run a set of binary logit/probit models for $Z_j = (Y <= j)$, the $\beta$ parameters in all of them should be identical, with only the constants differing from one another.

(see propodds() in library VGAM for model fitting and testing the assumption.)

When parallel regression assumption does not hold, consider models for unordered data. (Below.)
3. Multinomial Choice Models: Multinomial Logit and Probit

Example: Vote choice with more than two candidates.

There are \( I \) decision makers, each individual \( i \) choosing among \( J_i \) (for notational simplicity, \( J \) hereafter) alternatives. Utility of alternative \( j \) to \( i \) is \( u_{ij} = v_{ij} + \epsilon_{ij} \), where \( v_{ij} \) is a deterministic component, and \( \epsilon_{ij} \) is an unobserved, stochastic disturbance. \( v_{ij} \) is usually specified as a linear function of observed independent variables, \( x_i \beta_j \) (multinomial logit), or \( w_{ij} \gamma \) (conditional logit or McFadden’s choice model), or \( x_i \beta_j + w_{ij} \gamma \) (mixed). mlogit in zelig estimates the multinomial logit model. for the other two with alternative specific covariates, stata command asclogit is handy.

Under stochastic utility maximization, \( i \) chooses \( j \) if and only if \( u_{ij} > u_{ik} \) for all \( k \neq j \). So the probability of \( i \) choosing \( j \) is:

\[
p_{ij} = Pr\{u_{ij} > u_{ik} \ \forall k \neq j\}
\]
\[
Pr\{v_{ij} + \epsilon_{ij} > v_{ik} + \epsilon_{ik}, \forall k \neq j\} = Pr\{\epsilon_{ik} < v_{ij} + \epsilon_{ij} - v_{ik}, \forall k \neq j\} = \int_{-\infty}^{+\infty} \int_{-\infty}^{v_{ij} + \epsilon_{ij} - v_{i1}} \cdots \int_{-\infty}^{v_{ij} + \epsilon_{ij} - v_{iJ}} f(\epsilon_{i1}, \ldots, \epsilon_{iJ}) d\epsilon_{iJ} \ldots d\epsilon_{i1} d\epsilon_{ij} \quad (4)
\]

where \(f(\epsilon_{i1}, \ldots, \epsilon_{iJ})\) is the joint density of the \(\epsilon'_{ij}\)s. Let \(F(\epsilon_{i1}, \ldots, \epsilon_{iJ})\) be the cumulative distribution function of the disturbances. Equation 4 can be equivalently expressed as:

\[
p_{ij} = \int_{-\infty}^{+\infty} F_j(v_{ij} + \epsilon_{ij} - v_{i1}, \ldots, \epsilon_{ij}, \ldots, v_{ij} + \epsilon_{ij} - v_{iJ}) d\epsilon_{ij} \quad (5)
\]

where \(F_j\) is the partial derivative of \(F\) with respect to its \(j^{th}\) argument.

From (4) or (5) a particular choice model is obtained by specifying the joint distribution of the disturbances.

Criteria for choosing \(f(.)\) include functional flexibility (allow general patterns of heterogeneity and interdependence of the disturbances) and computational
The Multinomial Probit model is obtained by assuming that $\epsilon_{ij}$’s follow a multivariate normal distribution. This allows relatively general covariance structure. However, the multivariate normal cumulative distribution function $F(.)$ (and its partial derivatives) has no closed form solutions, so for probit choice probabilities expression (5) is a simplification of (4) only in notation. The computation involves dealing with multiple integrals (through numerical methods or by simulation). Costly when the choice set contains more than a few alternatives.

The Multinomial Logit model results from assuming that $\epsilon_{ij}$ have i.i.d. extreme value distribution: $F(\epsilon_{ij}) = e^{-e^{-\epsilon_{ij}}}$. Unlike the normal distribution function, the cumulative distribution function $F(.)$ here has a closed form,
and logit choice probabilities \( p_{ij} \) can be derived from equation (5) as:

\[
p_{ij} = \frac{e^{v_{ij}}}{\sum_{k=1}^{J} e^{v_{ik}}}, \quad j = 1, 2, \ldots, J
\]  

(6)

with \( \sum_j p_{ij} = 1 \).

Likelihood function formed similar to ordered probit (1). Note that only \( J - 1 \) sets of parameters can be estimated (any alternative \( k \) can serve as the “base” category. If none of the variables matter for a particular comparison, the two categories can be combined.)

The mlogit probabilities have intuitive interpretations: the higher the relative utility provided by an alternative, the higher the probability of that alternative being chosen. mlogit is a generalization of binary logit, of course. (show it.) mprobit similarly generalizes binary probit.

Zelig:
z.out <- zelig(as.factor(Y) ~ X1 + X2, model = "mlogit",
data = mydata)

“setx()” and “sim()” used the standard way. Quantities of interest evolve around the $J$ probabilities (first differences, risk ratios, etc.) If $J = 3$, could do ternary plot. (See Zelig manual, example for mlogit. Mexican 3-party election. Substantive question: if voters had thought the ruling PRI party was weakening, who’d have won??)

Run “demo(mlogit)”.

Generalizations. The mlogit model is computationally simple. However, $\epsilon_{ij}$’s are assumed to be independent and identically distributed. $\rightarrow$ Homoscedastic errors. Restrictions also include the IIA (independence from irrelevant alternatives) property: the ratio of the choice probabilities of any two alternatives not affected by the presence of other alternatives (show.) IIA not plausible when some alternatives are close substitutes.
The *Nested logit model* generalizes the mlogit and relaxes the IIA restriction: “similar” alternatives grouped into subsets, error terms within the subsets can be correlated. The GEV (generalized extreme value) model is a further generalization, the choice probabilities taking the form:

\[
p_{ij} = e^{v_{ij}} G_j(e^{v_i1}, \ldots, e^{v_iJ}) / G(e^{v_i1}, \ldots, e^{v_iJ})
\]

(7)

where \( G(Y_1, \ldots, Y_J) \) is a non-negative generating function (satisfying a set of technical conditions that ensure the joint distribution and the resulting marginal distributions are well defined), and \( G_j \) is the partial derivative of \( G \) with respect to its \( j^{th} \) argument.

The standard mlogit model results when

\[
G(Y_1, \ldots, Y_J) = \sum_{k=1}^{J} Y_k. \quad \text{(verify)}
\]
The nested logit model results from $G$ taking the form:

$$G(Y_1, \ldots, Y_J) = \sum_{k=1}^{K} \left( \sum_{i \in I_k} Y_i^{-1/\sigma_k} \right) \sigma_k$$

(8)

where $I_k \subset \{1, \ldots, J\}$, $\bigcup_{k=1}^{K} I_k = \{1, \ldots, J\}$, and $0 < \sigma_k \leq 1$. Thus the $J$ alternatives are grouped into $K$ subsets. The parameter $\sigma_k$ is interpreted as an index of the similarities of the alternatives within subset $I_k$. When $\sigma_k \equiv 1$, the nested logit model reduces to the mlogit model. (stata: nlogit.)

However, the entire GEV class of models still imposes homoscedasticity. The generalized GEV class (Zeng 2000) allows for heteroscedasticity, with the heteroscedastic logit/mlogit model being a special case. When heteroscedasticity is across individuals only (not alternatives), the choice probabilities retain a simple/intuitive form:

$$p_{ij} = e^{v_{ij} \theta_i} / \sum_{k=1}^{J} e^{v_{ik} \theta_i}$$

(9)
where $\theta_i$’s are inversely related to the standard error of the error terms. Intuitively, meaning more weight is given to individuals whose utility functions have smaller error variances. With heteroscedasticity across alternatives also, the model is:

$$p_{ij} = \int_{-\infty}^{\infty} \theta_{ij} e^{-\epsilon \theta_{ij}} \exp\{-\sum_{k=1}^{J} e^{-(v_{ij}+\epsilon-v_{ik})\theta_{ik}}\} d\epsilon$$

(10)

This involves only one dimensional integration, regardless of the number of alternatives $J$ (unlike mprobit). $\theta_i$ or $\theta_{ij}$ can be specified as functions of covariates with unknown parameters to be estimated along with other parameters in the model, and in the $\theta_{ij} = \theta_i$ case the model can be estimated with programs for standard models with simple transformation of variables. Alternatively, $\theta$’s can be given their own distributions, resulting in a class of random parameter logit models.