Nominal Variables

Michael D. Ward[1]

[1]Department of Political Science
Duke University, Durham, NC

February 13, 2013

Outline of Topics

Unordered, polychotomous dependent variables are simply variables in which the categories can not be ordered in any mathematically meaningful way. These are also called *nominal* variables, but have more that the two categories found in a binary, dichotomous variable.

There are lots of good examples in the social sciences: vote choice (Democrat, Republican, Libertarian, ...), occupation (doctor, lawyer, mechanic, astronaut, student, ...), martial status (single, married, divorced, ...), college major (art history, modern history, Greek history, ...), language (French, German, Urdu, ...), ethnicity (Serb, Croat, Bosniak, Avar, Lek, ...), and many, many others.

Often these nominal categories represent things that are being chosen, other times they can represent ascriptive categories.

The multinomial logit/probit models are designed to model nominal outcomes in such a way that the effects of the independent variable vary for each outcome.

Matt Krain

- Let's reconsider the model developed by Matthew Krain

Matt Krain

- Let's reconsider the model developed by Matthew Krain
- This is a dataset of N observations $i = \{1, 2, ...N\}$, and a dependent variable Y_i which can take on any of J values, which we consider to be *unordered* in this example.

Matt Krain

- Let's reconsider the model developed by Matthew Krain
- This is a dataset of N observations $i = \{1, 2, ...N\}$, and a dependent variable Y_i which can take on any of J values, which we consider to be *unordered* in this example.
- These are the eleven levels of the magnitude of politicide or genocide that Krain uses.

Matt Krain

- Let's reconsider the model developed by Matthew Krain
- This is a dataset of N observations $i = \{1, 2, ...N\}$, and a dependent variable Y_i which can take on any of J values, which we consider to be *unordered* in this example.
- These are the eleven levels of the magnitude of politicide or genocide that Krain uses.
- Krain labels them with numbers, but we will simply label them with eleven different letters, which are not ordered: $Y_i \in \{A, W, Z, M, Q, B, R, P, U, C, L\}$. The elements of this set are *indexed* by j, such that for example Y_i is in category M would imply $j = 4$.

Specify Model

- Begin by setting the probability that Y_i takes on a particular value in the set of j: $P(Y_i = j) = P_{i,j}$. The sum must be unity: $\sum_{j=1}^{J} P_{i,j} = 1$.

Specify Model

- Begin by setting the probability that Y_i takes on a particular value in the set of j: $P(Y_i = j) = P_{i,j}$. The sum must be unity: $\sum_{j=1}^{J} P_{i,j} = 1$.
- Adding covariates, we allow the probability of $Y_i = j \in J$ to vary as a function of k independent variable(s) $\mathbf{X}_i$, in the normal fashion, except these are indexed by a $k \times 1$ vector of parameters specific to each outcome β_j.

Specify Model

- Begin by setting the probability that Y_i takes on a particular value in the set of j: $P(Y_i = j) = P_{i,j}$. The sum must be unity: $\sum_{j=1}^{J} P_{i,j} = 1$.
- Adding covariates, we allow the probability of $Y_i = j \in J$ to vary as a function of k independent variable(s) $\mathbf{X}_i$, in the normal fashion, except these are indexed by a $k \times 1$ vector of parameters specific to each outcome β_j.
- These probabilities must be positive, so in a standard way we can use an exponential setup: $P_{i,j} = \exp(\mathbf{X}_i \beta_j)$.

Specify Model

- Begin by setting the probability that Y_i takes on a particular value in the set of j: $P(Y_i = j) = P_{i,j}$. The sum must be unity: $\sum_{j=1}^{J} P_{i,j} = 1$.
- Adding covariates, we allow the probability of $Y_i = j \in J$ to vary as a function of k independent variable(s) $\mathbf{X}_i$, in the normal fashion, except these are indexed by a $k \times 1$ vector of parameters specific to each outcome β_j.
- These probabilities must be positive, so in a standard way we can use an exponential setup: $P_{i,j} = \exp(\mathbf{X}_i \beta_j)$.
- To insure that this sum is also 1, this can simply be rescaled:

$$P(Y_i = j) \equiv P_{i,j} = \frac{\exp(X_i \beta_j)}{\sum_{j=1}^{J} \exp(X_i \beta_j)}$$

Specify Model

- Begin by setting the probability that Y_i takes on a particular value in the set of j: $P(Y_i = j) = P_{i,j}$. The sum must be unity: $\sum_{j=1}^{J} P_{i,j} = 1$.
- Adding covariates, we allow the probability of $Y_i = j \in J$ to vary as a function of k independent variable(s) $\mathbf{X}_i$, in the normal fashion, except these are indexed by a $k \times 1$ vector of parameters specific to each outcome β_j.
- These probabilities must be positive, so in a standard way we can use an exponential setup: $P_{i,j} = \exp(\mathbf{X}_i \beta_j)$.
- To insure that this sum is also 1, this can simply be rescaled:

$$P(Y_i = j) \equiv P_{i,j} = \frac{\exp(X_i \beta_j)}{\sum_{j=1}^{J} \exp(X_i \beta_j)}$$

- Which should look really familiar.

More Details on Model

- Each observations's probability associated with category j is a fraction of the sum of it's probabilities across the all J categories. Unfortunately, this is unidentified.

More Details on Model

- Each observations's probability associated with category j is a fraction of the sum of it's probabilities across the all J categories. Unfortunately, this is unidentified.
- This is commonly dealt with by constraining one of the parameters for a particular category to be zero. Normally the first category is thereby eliminated, and serves as the *baseline* category against which other categories are implicitly compared.

More Details on Model

- Each observations's probability associated with category j is a fraction of the sum of it's probabilities across the all J categories. Unfortunately, this is unidentified.
- This is commonly dealt with by constraining one of the parameters for a particular category to be zero. Normally the first category is thereby eliminated, and serves as the *baseline* category against which other categories are implicitly compared.
- This results in the following statement of the basic multinomial model:

$$P(Y_i = j \mid X_i) = \frac{\exp(X_i \beta_j)}{1 + \sum_{j=2}^{J} \exp(X_i \beta_j)} \quad \forall \quad j > 1$$

Les Voila, Les probabilités

- The likelihood is formed in the canonical fashion, with $z_{i,j}$ an indicator variable for observation i in category j

Les Voila, Les probabilités

- The likelihood is formed in the canonical fashion, with $z_{i,j}$ an indicator variable for observation i in category j
- Equations include sums over all j categories, even that used to identify the model. The first category serves as the reference. $\beta_j \equiv 0.0$ and $e^{\mathbf{x}_i \beta_1} = 1$ when $\beta_1 = 0$. Thus, sums go from $j = 1$, not $j = 2$.

Les Voila, Les probabilités

- The likelihood is formed in the canonical fashion, with $z_{i,j}$ an indicator variable for observation i in category j
- Equations include sums over all j categories, even that used to identify the model. The first category serves as the reference. $\beta_j \equiv 0.0$ and $e^{\mathbf{x}_i \beta_1} = 1$ when $\beta_1 = 0$. Thus, sums go from $j = 1$, not $j = 2$.
- Thus, sums go from $j = 1$, not $j = 2$.

$$\begin{aligned}
L_i &= \prod_{j=1}^{J} [P(Y_i = j)]^{z_{ij}} \\
L &= \prod_{i=1}^{N} \prod_{j=1}^{J} [P(Y_i = j)]^{z_{ij}} \\
lnL &= \sum_{i=1}^{N} \sum_{j=1}^{J} z_{ij} ln \left(\frac{\exp(\mathbf{X}_i \beta_j)}{\sum_{j=1}^{J} \exp(\mathbf{X}_i \beta_j)} \right)
\end{aligned}$$

Code is easy

```
library(MASS); library(nnet)
krain<-read.dta("isq05.dta",convert.dates = TRUE,
        convert.factors = TRUE, missing.type = FALSE,
        convert.underscore=TRUE, warn.missing.labels=TRUE)
modl.mnl<-multinom(magnitud ~ intrvlag+icntglag+maglag
         +genyr+stfl+regtype+ethkrain+marg+coldwar,
         data=krain)
```

Results are Ugly
comme d'habitude

Coefficient	$\hat{\beta_{i,j}}$	$\sigma_{\hat{\beta}_{i,j}}$	t-valueless
(Intercept) | A	4.3	2.9	1.4812
(Intercept) | W	3.2	3.1	1.0052
(Intercept) | Z	5.4	3.1	1.7316
(Intercept) | M	4.4	3.2	1.3843
(Intercept) | Q	-8.2	170	-0.0484
(Intercept) | B	-4.0	3.3	-1.2306
(Intercept) | R	-2.6	3.1	-0.8170
(Intercept) | P	1.2	3.0	0.4059
(Intercept) | U	-0.887	3.3	-0.2674
(Intercept) | C	-11.0	6.5	-1.7084
intrvlag | A	-1.8	1.3	-1.3996
intrvlag | W	-2.2	1.4	-1.5405
intrvlag | Z	-1.4	1.4	-0.9956
intrvlag | M	-3.6	1.5	-2.3986
intrvlag | Q	-3.0	1.7	-1.7696
intrvlag | B	0.088	1.4	0.0639
intrvlag | R	-.88	1.4	-0.6516
intrvlag | P	-1.4	1.3	-1.0700
intrvlag | U	-2.3	1.4	-1.5718
intrvlag | C	-6.8	2.9	-2.3274
$\vdots$	$\vdots$	$\vdots$	$\vdots$
∞	∞	∞	∞

Multinomial v. Ordered

We can use this model to check whether the ordered logit violates the parallel regressions assumption, because in the the multinomial logit there is no requirement for equal slopes. A likelihood test compares the deviance in each model, this difference should be distributed χ^2 with degrees of freedom equal to the difference in the degrees of freedom in each model.

```
deviance(modl.ol) - deviance(modl.mnl)
pchisq(deviance(modl.ol) - deviance(modl.mnl), modl.mnl$edf - modl.ol$edf,
    lower.tail=FALSE)
```

which has a very tiny value in this case–i.e., the multinomial model has much smaller deviance, even controlling for the high number of parameters it estimates–suggesting that the parallel regression assumption is violated by the ordered logit estimation. It usually is.

Canada Example

Recently, Martinez and Gill (2006) used a multinomial regression analysis to bolster their argument that in Canada (and by extension other western democracies) falling turnout rates do not necessarily benefit the more liberal parties, as is widely asserted in the literature on party politics. Indeed they find that in some contexts–for example, Quebec–it is the party that attracts younger voters that benefits from higher turnouts, but that is not always the most liberal party.

Canada Example

We use data from a more recent Canadian Election Survey, undertaken in 2000. We have a simplified model that tries to associate potential vote choice (Which party do you think you will vote for?) with an assessment by each respondent of their goals.

Canada Example

Here's a list of FOUR goals. Which goal is MOST important to you personally?

1. fighting crime;
2. giving people more say in important government decisions;
3. maintaining economic growth;
4. protecting freedom of speech;
5. don't know; or
6. refused.

Canada Example

The actual data are presented in tabluar form quite easily:

Probable Vote	Econ Growth	Fight Crime	More Say	Free Speech	Don't Know	Refuse to Say
Liberal	168	80	341	108	14	2
Alliance	102	79	188	57	10	1
Conservative	27	23	50	24	0	0
NDP	18	43	29	39	2	0
Bloc Quebeçois	50	45	83	47	2	1
Green	0	5	0	3	0	0
Other	9	6	9	5	0	0
Won't vote	2	1	3	1	0	0
None	1	1	6	3	0	0
Undecided	206	168	301	146	24	2
Refuse	74	40	78	47	12	3

Canada Example

I pruned the data in order to focus only on the Liberal, Alliance, Conservative, NDP, and Bloc Quebeçois voters. I also eliminated those respondents who refused to identify a policy goal. Based on three categorical variables (an index of reported family income), a simple model was estimated, some results of which are provided below.

Canada Example

Table: *Political as a predictor of probably vote choice in the 2000 Canadian Elections. Data from the 2000 Canadian Election Survey were provided by the Institute for Social Research, York University. Estimated coefficients are from a multinomial logistic regression; the reference category is the Liberal party.*

	Alliance v. Liberal	Conservative v. Liberal	NDP v. Liberal	Bloc Quebeçois v. Liberal
Intercept	-1.16	-2.28	-1.87	-2.09
More Say in Government	0.7	-26.09	2.02	1.15
Economic Growth	0.56	0.01	0.58	-0.54
Protect Free Speech	0.01	-0.13	1.06	-0.67
20-30K CAD	0.56	1.39	-1.13	0.62
30-40K	0.02	0.35	-0.47	-0.31
40-50K	0.75	-17.8	0.01	-0.31
50-60K	0.58	1.03	-1.68	1.09
60-70K	-1.36	0.22	-1.19	0.65
70-80K	0.93	1.31	-18.37	1.03
80-90K	0.78	2.28	-15.85	-11.4
90-100K	-17.95	-15.64	-0.44	0.56
100K and over	1.18	0.67	-0.64	-0.13

Interpretation

In the same way one normally interprets these models, by simulating outcomes under the data generating process that includes the systematic and stochastic components in all their glory (i.e., including uncertainty in each part). Basically, this means using the fundamental probability statement of the multinomial model, which is a normalized exponential function,

$$P(Y_i = j) = \frac{exp(\mathbf{X}_i \beta_j)}{\sum_{j=1}^{J} exp(\mathbf{X}_i \beta_j)} \quad ,$$

which will all the generation of J predicted probabilities for each observation. These can be calculated in the context of a simulation that draws estimated parameters from their posterior distributions, conditional on the estimated mean and covariance structure. The averages of these probabilities in simulation can indicate into which *single* category the prediction is most likely to fall, yielding a resultant *prediction*, or classification. These predictions can then be compared with the actual data in the normal way. Each category could have its own ROC plot, for example, along with the underpinning correct predictions among the highest probabilities, and all the standard stuff.

Interpretation

Similarly, one can calculate the partial derivatives for particular variables in the usual fashion:

$$\frac{\partial P(Y_i = j)}{\partial X_k} = P(Y_i = j|\mathbf{X}) \left[\hat{\beta}_{j,k} - \sum_{j=1}^{J} \hat{\beta}_{j,k} \times P(Y_i = j|\mathbf{X}) \right]$$

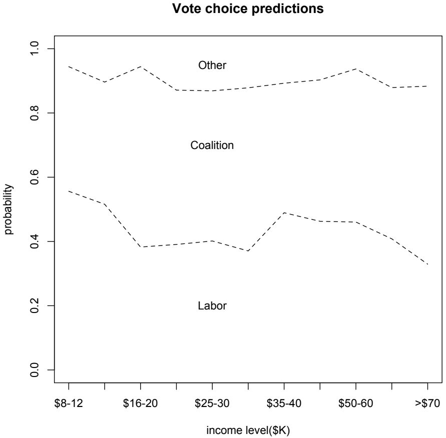
As shown in Long (1997) this partial derivative is not independent of the values of other values, and may not even have the same sign as an estimated coefficient for variable k, for example. This derivative is conditional on the values of all independent variables: there is an equation for each category, in the multinomial model. As a result, this approach to interpretation is best avoided, though you may encounter it in your reading.

Interpretation

A better way of displaying results is graphically. This figure displays the results of a multinomial regression from 1996 Australian election polling data. The outcome variable is the survey respondent's vote choice among {Labor, Liberal, Australian Democrats, Greens, National, Other, Informal}. The predictors are religious affiliation, income, and retrospective evaluation of the economic situation. We display results for the predicted probability of voting for one of Australia's two major political parties (Labor and the Liberal/National Coalition) or one of the other smaller parties for different levels of income. Religious affiliation and perceptions about the economy are held at their modal values.

Interpretation

Figure: Predicted vote choice at different levels of income



Code

```
mnl.fit<-multinom(as.factor(votechoice) ~  religion + fin.sitch + income,
                  Hess=T, model=T, data=myoz)
income.seq<-c("$8001 to $12000","$12001 to $16000",
              "$16001 to $20000","$20001 to $25000",
              "$25001 to $30000","$30001 to $35000",
              "$35001 to $40000", "$40001 to $50000",
              "$50001 to $60000", "$60001 to $70000",
              "More than $70000") #varying income
income.lab<-c("$8-12","$12-16","$16-20","$20-25","$25-30","$30-35",
              "$35-40", "$40-$50","$50-60", "$60-70"">$70") #for plotting
X<-data.frame("Roman Catholic", "About the same", income.seq) #modal categories
colnames(X)<-names(mnl.fit$model)[-1]
pp.mnl<-predict(mnl.fit,newdata=X,type="probs")
pp.mnl.comb<-pp.mnl
pp.mnl.comb[,"Australian Democrats"]<-pp.mnl.comb[,"Australian Democrats"] +
  pp.mnl.comb[,"Greens"] + pp.mnl.comb[,"Informal/Didnt vote"] +
  pp.mnl.comb[,"Other party"] #combining the small parties for display purposes
pp.mnl.comb[,"Liberal"]<-pp.mnl.comb[,"Liberal"] + pp.mnl.comb[,"National (country)"] #
pp.mnl.comb<-pp.mnl.comb[,-c(3,4,6,7)]
colnames(pp.mnl.comb)[2]<-"Other"
colnames(pp.mnl.comb)[3]<-"Coalition"
lst<-c("Labor", "Coalition", "Other")
pp.mnl.comb<-pp.mnl.comb[,lst] #reordering columns for cumsum
ycs<-apply(pp.mnl.comb,1,cumsum)
plot(0,0,col="white",xlim=c(1,11), ylim=c(0,1), xaxt="n",xlab="income level($K)", ylab="
axis(1,at=1:11,labels=income.lab)
for(i in 1:(nrow(ycs)-1)){
lines((1:11),ycs[i,], lty=2)}
labs<-c("Labor", "Coalition", "Other")
text(c(5,5,5),c(.2,.7,.95),labels=labs)
```

Don't do this

Another approach, generally to be avoided, is the dreaded *odds ratio* interpretation. Since the multinomial logit is actually a log-odds model, it may be useful to note that the log of the ratio of two probabilities is a function of the independent variables:

$$\log\left[\frac{P(Y_i = j|\mathbf{X})}{P(Y_i = j'|\mathbf{X})}\right] = \mathbf{X}(\hat{\beta}_j - \hat{\beta}_{j'})$$

Don't do this

Setting the coefficients of one category (e.g., $\hat{\beta}_{j'}$) to zero, yields:

$$\log\left[\frac{P(Y_i = j|\mathbf{X})}{P(Y_i = j'|\mathbf{X})}\right] = \mathbf{X}(\hat{\beta}_j)$$

Don't do this

This approach is *linear in the variables* and allows the calculation of the hypothetical change in the odds ratio for category j associated with a particular variable X_k by exponentiation (i.e., $exp(\hat{\beta}_{j,k})$). I hate it when that happens. But since we are comparing categories in this approach, maybe comparing the probability of being in one category to another is an appropriate heuristic framework.

Caveats...

IIA stands for the *independence of irrelevant alternatives*, which is the assumption that an individual's choice does not depend on the availability or characteristics of unavailable alternatives.

Caveats...

This is an assumption about the nature of the choice process. But it has implications about the implications of adding or subtracting alternatives from the choice set. Suppose you have to choose a third course next semester between a) a course on game theory and b) a course on political geography. You make your choice for (a), but later in the afternoon to learn that an additional course is now available: c) publishing for social scientists, at which point you switch your choice to (c).

Caveats...

In the example given above, IIA precludes the patron having preference orderings amounting to {Game Theory $>$ Political Cartography} and {Political Cartography $>$ Publishing for Social Scientists $>$ Game Theory } or {Political Cartography $>$ Game Theory $>$ Publishing for Social Scientists}. This is because removal of Publishing for Social Scientists from either of the latter two is inconsistent with the former.

Caveats...

This implies that

$$\begin{aligned}
\frac{P(Y_i = j)}{P(Y_i = \ell)} &= \frac{\frac{\exp(\mathbf{X}_i\beta_j)}{\sum_{j=1}^{J}\exp(\mathbf{X}_i\beta_j)}}{\frac{\exp(\mathbf{X}_i\beta_\ell)}{\sum_{j=1}^{J}\exp(\mathbf{X}_i\beta_j)}} \\
&= \frac{\exp(\mathbf{X}_i\beta_j)}{\exp(\mathbf{X}_i\beta_\ell)} \\
&= \exp[\mathbf{X}_i(\beta_j - \beta_\ell)]
\end{aligned}$$

that is, the ratio of the probabilities for any two alternatives j and ℓ is just the values of the covariates times the difference between the two alternatives' coefficient vectors.

Caveats...

Importantly, this means that *the ratio of the probabilities of choosing any two outcomes is invariant with respect to the other alternatives*. It only depends on the characteristics of the alternatives in question:

$$\frac{P(Y_i = j|S_J)}{Pr(Y_i = \ell|S_J)} = \frac{P(Y_i = j|S_M)}{Pr(Y_i = \ell|S_M)} \forall j, \ell, J, M$$

where J and M are different sizes of the choice set S.

Caveats...

This can be examined with a heuristic test. If, in fact, the IIA assumption holds, then a model that omits any particular choice should give similar estimates of the $\hat{\beta}_j$s for the remaining alternatives. Conversely, if the $\hat{\beta}_j$s vary a lot when an alternative is omitted, the model will probably fail tests for the independence of irrelevant alternatives:

Caveats...

$$\chi^2_k = (\hat{\beta}_r - \hat{\beta}_u)'[\hat{\mathbf{V}}_r - \hat{\mathbf{V}}_u]^{-1}(\hat{\beta}_r - \hat{\beta}_u)$$

where $\hat{\beta}_r$ are the estimates from the restricted model (i.e., the model with an omitted alternative), $\hat{\beta}_u$ is the vector of estimates for the unrestricted model (that is, the one with all the alternatives included), and $\hat{\mathbf{V}}_r$ and $\hat{\mathbf{V}}_u$ are the estimated variance–covariance matrices for the two sets of coefficients, respectively. This is distributed χ^2_k (with degrees of freedom equal to rank of β_j–the number of covariates). This test is known as the Hausman test, and is essentially identical to the Small-Hsiao test for IIA, which is a based on a likelihood ratio test.

$$
\begin{aligned}
\mathcal{L}(\beta_2, \ldots, \beta_J | \mathbf{y}, \mathbf{X}) &= \prod_{j=1}^{J} [P(Y_i = j)]^{y_{ij}} \\
\mathcal{L} &= \prod_{i=1}^{N} \prod_{j=1}^{J} [P(Y_i = j)]^{y_{ij}} \\
\ln \mathcal{L} &= \sum_{i=1}^{N} \sum_{j=1}^{J} y_{ij} ln \left(\frac{\exp(\mathbf{X}_i \beta_j)}{1 + \sum_{j=2}^{J} \exp(\mathbf{X}_i \beta_j)} \right)
\end{aligned}
$$

And expected probabilities are easy. E.g., in a three category model:

$$\begin{aligned}
\hat{\pi}_1 &= \frac{1}{1+\exp(x_s\hat{\beta}_{2k}+\exp(x_s\hat{\beta}_{3k})} \\
\hat{\pi}_2 &= \frac{\exp(x_s\hat{\beta}_{2k})}{1+\exp(x_s\hat{\beta}_{2k})+\exp(x_s\hat{\beta}_{3k})} \\
\hat{\pi}_3 &= \frac{\exp(x_s\hat{\beta}_{3k})}{1+\exp(x_s\hat{\beta}_{2k})+\exp(x_s\hat{\beta}_{3k})}
\end{aligned}$$

To simulate, use honey: Draw the Bees, and put em in the above containers.

$$\ln \frac{P(y=j|\mathbf{x_s})}{P(y=j+q|\mathbf{x_s})} = \mathbf{x_s}(\beta_j - \beta_{j+q})$$

If a covariate changes by 1, the log of the odds of j versus $j+q$ is simply the difference of their coefficients: $(\beta_j - \beta_{j+q})$. This means than other categories are *irrelevant* (and that the relative probabilities don't change even if a new alternative close to j or $j+q$ is added (painting a bus).

Conditional logic adds a covariate with different values for each combination.

If the problem is a classification problem, no sweat. If the model is a choice problem, the assumption of IIA is a deal breaker.

Solution is latent variables, with MVN, allows categories to be correlated. But is very complicated. See Imai and van Dyk (2005) for a Bayesian version.