PPOL 503-03, PPOL 503-04, Fall 2016
Course notes #9: Example of Multinomial Logit Model in Stata

Example 1: Multinomial Logit with One Explanatory Variable

We have data on the type of health insurance available to 616 psychologically depressed subjects in the United States (Tarlov et al. 1989; Wells et al. 1989). The insurance is categorized as either an indemnity plan (that is, regular fee-for-service insurance, which may have a deductible or coinsurance rate) or a prepaid plan (a fixed up-front payment allowing subsequent unlimited use as provided, for instance, by an HMO). The third possibility is that the subject has no insurance whatsoever. We wish to explore the demographic factors associated with each subject's insurance choice. One of the demographic factors in our data is the race of the participant, coded as white or nonwhite:

. tabulate insure nonwhite, chi2 col

INSURE	NONWHITE = 0	NONWHITE = 1	TOTAL
Indemnity (Y = 1)	251 50.71%	43 35.54%	294 47.73%
Prepaid (Y =2)	208 42.0%	69 57.0%	277 44.97%
Uninsure (Y = 3)	36 7.27%	9 7.44%	45 7.31%
TOTAL	495 100.0%	121 100.0%	616 100.00

Pearson chi2(2) = 9.5599 Pr = 0.008

When we fit a multinomial logit model, we can tell mlogit which outcome to use as the base outcome, or we can let mlogit choose. To fit a model of insure on nonwhite, letting mlogit choose the base outcome, we type

. mlogit insure nonwhite

Iteration 0: log likelihood = -556.59502
Iteration 1: log likelihood = -551.78935
Iteration 2: log likelihood = -551.78348
Iteration 3: log likelihood = -551.78348

Multinomial logistic regression — Number of obs = 616
LR chi2(2) = 9.62
Prob > chi2 = 0.0081
Log likelihood = -551.78348 — Pseudo R2 = 0.0086

Insure	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Indemnity (base outcome)						
Prepaid						
nonwhite	.6608212	.2157321	3.06	0.002	.2379942	1.083648
_cons	-.1879149	.0937644	-2.00	0.045	-.3716896	-.0041401
Uninsure						
nonwhite	.3779586	.407589	0.93	0.354	-.4209011	1.176818
_cons	-1.941934	.1782185	-10.90	0.000	-2.291236	-1.592632

These results agree with the column percentages presented by tabulate because the mlogit model is fully saturated. That is, there are enough terms in the model to fully explain the column percentage in each cell. The model chi-squared and the tabulate chi-squared are in almost perfect agreement; both test that the column percentages of insure are the same for both values of nonwhite.

Example 2: Specifying the base outcome

By specifying the baseoutcome() option, we can control which outcome of the dependent variable

is treated as the base. Left to its own, mlogit chose to make outcome 1, indemnity, the base outcome.
To make outcome 2, prepaid, the base, we would type
. mlogit insure nonwhite, base(2)

Iteration 0: log likelihood = -556.59502
Iteration 1: log likelihood = -551.78935
Iteration 2: log likelihood = -551.78348
Iteration 3: log likelihood = -551.78348
Multinomial logistic regression Number of obs = 616
LR chi2(2) = 9.62
Prob > chi2 = 0.0081
Log likelihood = -551.78348 Pseudo R2 = 0.0086

insure	Coef.	Std. Err.	Z	P>|z|	[95% Conf. Interval]
Indemnity					
nonwhite	-.6608212	.2157321	-3.06	0.002	-1.083648 -.2379942
_cons	.1879149	.0937644	2.00	0.045	.0041401 .3716896
Prepaid (base outcome)					
Uninsure					
nonwhite	-.2828627	.3977302	-0.71	0.477	-1.0624 .4966742
_cons	-1.754019	.1805145	-9.72	0.000	-2.107821 -1.400217

The baseoutcome() option requires that we specify the numeric value of the outcome, so we could not type base(Prepaid).

Example 3: Model with continuous and multiple categorical variables

One of the advantages of mlogit over tabulate is that we can include continuous variables and multiple categorical variables in the model. In examining the data on insurance choice, we decide
that we want to control for age, gender, and site of study (the study was conducted in three sites):

```
. mlogit insure age male nonwhite i.site

Iteration 0: log likelihood = -555.85446
Iteration 1: log likelihood = -534.67443
Iteration 2: log likelihood = -534.36284
Iteration 3: log likelihood = -534.36165
Iteration 4: log likelihood = -534.36165
Multinomial logistic regression                 Number of obs = 615
                                                LR chi2(10) = 42.99
                                                Prob > chi2 = 0.0000
Log likelihood = -534.36165                     Pseudo R2 = 0.0387

------------------------------------------------------------------------------
      insure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Indemnity (base outcome)
-------------+----------------------------------------------------------------
Prepaid      |
         age |   -.011745   .0061946    -1.90   0.058    -.0238862    .0003962
        male |    .5616934   .2027465     2.77   0.006     .1643175    .9590693
    nonwhite |    .9747768   .2363213     4.12   0.000     .5115955    1.437958
        site |
          2  |    .1130359   .2101903     0.54   0.591    -.2989296    .5250013
          3  |   -.5879879   .2279351    -2.58   0.010    -1.034733   -.1412433
       _cons |    .2697127   .3284422     0.82   0.412    -.3740222    .9134476
-------------+----------------------------------------------------------------
Uninsure     |
         age |  -.0077961   .0114418    -0.68   0.496    -.0302217    .0146294
        male |    .4518496   .3674867     1.23   0.219     -.268411    1.17211
    nonwhite |    .2170589   .4256361     0.51   0.610    -.6171725    1.05129
        site |
          2  |   -1.211563   .4705127    -2.57   0.010    -2.133751   -.2893747
          3  |   -.2078123   .3662926    -0.57   0.570    -.9257327    .510108
       _cons |   -1.286943   .5923219    -2.17   0.030    -2.447872   -.1260134
------------------------------------------------------------------------------
```

These results suggest that the inclination of nonwhites to choose prepaid care is even stronger than it was without controlling age, gender and site. We also see that subjects in site 2 are less likely to be uninsured.

Example 4: Specifying constraints to test hypotheses

We can use constraints to test hypotheses, among other things. In our insurance-choice model, let's test the hypothesis that there is no distinction between having indemnity insurance and being
uninsured. Indemnity-style insurance was the omitted outcome, so we type

```
. test [Uninsure]
( 1) [Uninsure]age = 0
( 2) [Uninsure]male = 0
( 3) [Uninsure]nonwhite = 0
( 4) [Uninsure]1b.site = 0
( 5) [Uninsure]2.site = 0
( 6) [Uninsure]3.site = 0
chi2( 5) = 9.31
Prob > chi2 = 0.0973
```

If indemnity had not been the omitted outcome, we would have typed test [Uninsure=Indemnity]. The results produced by test are an approximation based on the estimated covariance matrix of the coefficients. Because the probability of being uninsured is low, the log likelihood may be nonlinear for the uninsured. Conventional statistical wisdom is not to trust the asymptotic answer under these circumstances but to perform a likelihood-ratio test instead.

To use Stata's lrtest (likelihood-ratio test) command, we must fit both the unconstrained and constrained models. The unconstrained model is the one we have previously fit. Following the instruction in [R] lrtest, we first store the unconstrained model results:

```
. estimates store unconstrained
```

To fit the constrained model, we must refit our model with all the coefficients except the constant set
to 0 in the Uninsure equation. We define the constraint and then refit:

```
. constraint 1 [Uninsure]
. mlogit insure age male nonwhite i.site, constraints(1)
Iteration 0: log likelihood = -555.85446
Iteration 1: log likelihood = -539.80523
Iteration 2: log likelihood = -539.75644
Iteration 3: log likelihood = -539.75643
```

Multinomial logistic regression — Number of obs = 615
Wald chi2(5) = 29.70
Log likelihood = -539.75643 — Prob > chi2 = 0.0000
(1) [Uninsure]o.age = 0
(2) [Uninsure]o.male = 0
(3) [Uninsure]o.nonwhite = 0
(4) [Uninsure]2o.site = 0
(5) [Uninsure]3o.site = 0

insure	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Indemnity (base outcome)						
Prepaid						
Age	-.0107025	.0060039	-1.78	0.075	-.0224699	.0010649
male	.4963616	.1939683	2.56	0.010	.1161907	.8765324
nonwhite	.9421369	.2252094	4.18	0.000	.5007346	1.383539
site						
2	.2530912	.2029465	1.25	0.212	-.1446767	.6508591
3	-.5521773	.2187237	-2.52	0.012	-.9808678	.1234869
_cons	.1792752	.3171372	0.57	0.572	-.4423023	.8008527
Uninsure						
age	0 (omitted)					
male	0 (omitted)					
nonwhite	0 (omitted)					
site						
2	0 (omitted)					
3	0 (omitted)					
_cons	-1.87351	.1601099	-11.70	0.000	-2.18732	-1.5597

We can now perform the likelihood-ratio test:
. lrtest unconstrained .
Likelihood-ratio test LR chi2(5) = 10.79
(Assumption: . nested in unconstrained) Prob > chi2 = 0.0557
The likelihood-ratio chi-squared is 10.79 with 5 degrees of freedom—just slightly greater than the magic p = 0.05 level—so we should not call this difference significant. We fail to reject the null hypothesis that having indemnity insurance is distinct from being uninsured.