Long & Freese, 2nd Edition

9 More topics

In this final chapter, we discuss some disparate topics that were not covered in the preceding chapters. We begin by considering complications on the right-hand side of the model: nonlinearities, interactions, and nominal or ordinal variables coded as a set of dummy variables. Although the same principles of interpretation apply in these cases, several tricks are necessary for computing the appropriate quantities. We then examine ways you can use basic Stata programming to simplify data analysis. Next we discuss briefly what is required if you want to modify SPost to work with other estimation commands. The final section discusses a menagerie of Stata "tricks" that we find useful for working more efficiently in Stata.

9.1 Ordinal and nominal independent variables

When an independent variable is categorical, it should be entered into the model as a set of binary indicator variables. Although our example uses an ordinal variable, the discussion applies equally to nominal independent variables, with one exception.

9.1.1 Coding a categorical independent variable as a set of dummy variables

A categorical independent variable with J categories can be included in a regression model as a set of $J-1$ dummy variables. Here we use a binary logit model to analyze factors affecting whether a scientist has published. The outcome is a dummy variable, `hasarts`, that is equal to 1 if the scientist has one or more publications and equals 0 otherwise. In our analysis in the last chapter, we included the independent variable `ment`, which we treated as continuous. But suppose instead that the data were from a survey in which the mentor was asked to indicate whether he or she had 0 articles (none), 1–3 articles (few), 4–9 (some), 10–20 (many), or more than 20 articles (lots). The resulting variable, which we call `mentord`, has the following frequency distribution:[1]

1. Details on creating `mentord` from the data in `couart2.dta` are located in `st9ch9other.do`, which is part of the `spost9_do` package. For details, when you are in Stata and online, type `search spost9_do, net`.

```
. tab mentord, missing
```

Ordinal measure of mentor's articles	Freq.	Percent	Cum.
None	90	9.84	9.84
Few	201	21.97	31.80
Some	324	35.41	67.21
Many	213	23.28	90.49
Lots	87	9.51	100.00
Total	915	100.00	

We can convert `mentord` into a set of dummy variables using a series of `generate` commands. Because the dummy variables are used to indicate in which category an observation belongs, they are often referred to as *indicator variables*. First, we construct `none` to indicate that the mentor had no publications:

```
. gen none = (mentord == 0) if mentord < .
```

Expressions in Stata equal 1 if true and 0 if false. Accordingly, `gen none = (mentord==0)` creates `none` equal to 1 for scientists whose mentors had no publications and equal to 0 otherwise. Although we have no missing values for `mentord`, it is a good habit to always add an `if` condition so that missing values continue to be missing. This is done by adding `if mentord < .` to the command (remember that missing values are treated by Stata as larger than all nonmissing values when evaluating expressions). We use `tab` to verify that `none` was constructed correctly:

```
. tab none mentord, missing
```

	Ordinal measure of mentor's articles					
none	None	Few	Some	Many	Lots	Total
0	0	201	324	213	87	825
1	90	0	0	0	0	90
Total	90	201	324	213	87	915

Likewise, we create indicator variables for the other categories of `mentord`:

```
. gen few = (mentord == 1) if mentord < .
. gen some = (mentord == 2) if mentord < .
. gen many = (mentord == 3) if mentord < .
. gen lots = (mentord == 4) if mentord < .
```

Note You can also construct indicator variables using `xi` or `tabulate`'s `gen()` option. For more information, type `help xi` or `help tabulate`.

9.1.2 Estimation and interpretation with categorical independent variables

Since `mentord` has $J = 5$ categories, we must include $J - 1 = 4$ indicator variables as independent variables in our model. To see why one of the indicators must be dropped, consider our example. If you know that `none`, `few`, `some`, and `many` are all 0, it must be that `lots` equals 1 because a person has to be in one of the five categories. Another way to think of this is to note that `none` + `few` + `some` + `many` + `lots` = 1 so that including all J categories would lead to perfect collinearity. If you include all five indicator variables, Stata automatically drops one of them. For example,

```
. logit hasarts fem mar kid5 phd none few some many lots, nolog
note: lots dropped due to collinearity
Logistic regression                               Number of obs   =        915
  (output omitted)
```

The category that is excluded is the *reference category*, as the coefficients for the included indicators are interpreted relative to the excluded category, which serves as a point of reference. Which category you exclude is arbitrary, but with an ordinal independent variable, it is generally easier to interpret the results when you exclude an extreme category. For nominal categories, it is often useful to exclude the most important category. For example, we fit a binary logit, excluding the indicator variable `none`:

```
. logit hasarts fem mar kid5 phd few some many lots, nolog
Logistic regression                               Number of obs   =        915
                                                  LR chi2(8)      =      73.80
                                                  Prob > chi2     =     0.0000
Log likelihood = -522.46467                       Pseudo R2       =     0.0660
```

hasarts	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
fem	-.2579293	.1601187	-1.61	0.107	-.5717562	.0558976
mar	.3300817	.1822141	1.81	0.070	-.0270514	.6872147
kid5	-.2795751	.1118578	-2.50	0.012	-.4988123	-.0603379
phd	.0121703	.0802725	0.15	0.879	-.145161	.1695017
few	.3859147	.2586461	1.49	0.136	-.1210223	.8928517
some	.9602176	.2490493	3.86	0.000	.4720889	1.448346
many	1.463606	.2829625	5.17	0.000	.9090099	2.018203
lots	2.335227	.4368715	5.35	0.000	1.478975	3.19148
_cons	-.0521187	.3361977	-0.16	0.877	-.7110542	.6068167

Logit models can be interpreted in terms of factor changes in the odds, which we compute using `listcoef`:

```
. listcoef
logit (N=915): Factor Change in Odds
  Odds of: Arts vs NoArts
```

hasarts	b	z	P>|z|	e^b	e^bStdX	SDofX
fem	-0.25793	-1.611	0.107	0.7726	0.8793	0.4987
mar	0.33008	1.812	0.070	1.3911	1.1690	0.4732
kid5	-0.27958	-2.499	0.012	0.7561	0.8075	0.7649
phd	0.01217	0.152	0.879	1.0122	1.0121	0.9842
few	0.38591	1.492	0.136	1.4710	1.1734	0.4143
some	0.96022	3.856	0.000	2.6123	1.5832	0.4785
many	1.46361	5.172	0.000	4.3215	1.8568	0.4228
lots	2.33523	5.345	0.000	10.3318	1.9845	0.2935

The effect of an indicator variable can be interpreted just as we interpreted dummy variables in chapter 4, but with comparisons being relative to the reference category. For example, the odds ratio of 10.33 for `lots` can be interpreted as

> The odds of a scientist publishing are 10.3 times larger if his or her mentor had lots of publications compared with no publications, holding other variables constant.

Or, you can say equivalently:

> If a scientist's mentor has lots of publications as opposed to no publications, the odds of a scientist publishing are 10.3 times larger, holding other variables constant.

The odds ratios for the other indicators can be interpreted in the same way.

9.1.3 Tests with categorical independent variables

The basic ideas and commands for tests that involve categorical independent variables are the same as those used in prior chapters. But, because the tests involve some special considerations, we will review them here.

Testing the effect of membership in one category versus the reference category

When a set of indicator variables is included in a regression, a test of the significance of the coefficient for any indicator variable is a test of whether being in that category compared with being in the reference category affects the outcome. For example, the coefficient for **few** can be used to test whether having a mentor with few publications compared with having a mentor with no publications significantly affects the scientist's publishing. Here $z = 1.492$ and $p = 0.136$, so we conclude that the effect of having a mentor with a few publications compared with none is not significant using a two-tailed test ($z = 1.492$, $p = 0.14$).

Often the significance of an indicator variable is reported without mentioning the reference category. For example, the test of many could be reported as follows:

> Having a mentor with many publications significantly affects a scientist's productivity ($z = 5.17$, $p < .01$).

Here the comparison is implicitly being made to mentors with no publications. Such interpretations should be used only if you are confident that the implicit comparison will be apparent to the reader.

Testing the effect of membership in two nonreference categories

What if neither of the categories that we wish to compare is the reference category? A simple solution is to refit the model with a different reference category. For example, to test the effect of having a mentor with some articles compared with a mentor with many publications, we can refit the model using some as the reference category:

```
. logit hasarts fem mar kid5 phd none few many lots, nolog
Logistic regression                               Number of obs   =        915
                                                  LR chi2(8)      =      73.80
                                                  Prob > chi2     =     0.0000
Log likelihood = -522.46467                       Pseudo R2       =     0.0660
```

hasarts	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
fem	-.2579293	.1601187	-1.61	0.107	-.5717562	.0558976
mar	.3300817	.1822141	1.81	0.070	-.0270514	.6872147
kid5	-.2795751	.1118578	-2.50	0.012	-.4988123	-.0603379
phd	.0121703	.0802726	0.15	0.879	-.145161	.1695017
none	-.9602176	.2490498	-3.86	0.000	-1.448346	-.4720889
few	-.5743029	.1897376	-3.03	0.002	-.9461818	-.2024241
many	.5033886	.2143001	2.35	0.019	.0833682	.9234091
lots	1.37501	.3945447	3.49	0.000	.6017161	2.148303
_cons	.9080989	.3182603	2.85	0.004	.2843202	1.531878

The z-statistics for the mentor indicator variables are now tests comparing a given category with that of the mentor having some publications.

(*Continued on next page*)

Advanced: lincom For the model that excludes some, the estimated coefficient for many equals the difference between the coefficients for many and some in the earlier model that excluded none. This suggests that instead of refitting the model, we could have used lincom to fit $\beta_{\texttt{many}} - \beta_{\texttt{some}}$:

```
. lincom many-some
 ( 1) - some + many = 0
```

hasarts	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	.5033886	.2143001	2.35	0.019	.0833682	.9234091

The result is identical to that obtained by refitting the model with a different base outcome.

Testing that a categorical independent variable has no effect

For an omnibus test of a categorical variable, our null hypothesis is that the coefficients for all the indicator variables are zero. In our model where none is the excluded variable, the hypothesis to test is

$$H_0\colon\ \beta_{\texttt{few}} = \beta_{\texttt{some}} = \beta_{\texttt{lots}} = \beta_{\texttt{many}} = 0$$

This hypothesis can be tested with an LR test by comparing the model with the four indicators with the model that drops the four indicator variables:

```
. logit hasarts fem mar kid5 phd few some many lots, nolog
  (output omitted)
. estimates store fmodel
. logit hasarts fem mar kid5 phd, nolog
  (output omitted)
. lrtest fmodel
Likelihood-ratio test                          LR chi2(4)  =     58.32
(Assumption: . nested in fmodel)               Prob > chi2 =    0.0000
```

The key result is the change as few changes from 0->1 (which, because it is a dummy variable, is also the change from min->max). A Wald test can also be used, although the LR test is generally preferred:

```
. logit hasarts fem mar kid5 phd few some many lots, nolog
  (output omitted)
. test few some many lots
 ( 1)  few = 0
 ( 2)  some = 0
 ( 3)  many = 0
 ( 4)  lots = 0
           chi2(  4) =   51.60
         Prob > chi2 =    0.0000
```

which leads to the same conclusion as the LR test.

Exactly the same results would be obtained for either test if we had used a different reference category and tested, for example,

$$H_0\colon\ \beta_{\texttt{none}} = \beta_{\texttt{few}} = \beta_{\texttt{lots}} = \beta_{\texttt{many}} = 0$$

Testing whether treating an ordinal variable as interval loses information

Ordinal independent variables are often treated as interval in regression models. For example, rather than include the four indicator variables that were created from mentord, we might simply include only mentord in our model:

```
. logit hasarts fem mar kid5 phd mentord, nolog
Logistic regression                            Number of obs   =        915
                                               LR chi2(5)      =      72.73
                                               Prob > chi2     =     0.0000
Log likelihood = -522.99932                    Pseudo R2       =     0.0650
```

hasarts	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
fem	-.266308	.1598617	-1.67	0.096	-.5796312	.0470153
mar	.3329119	.1823256	1.83	0.068	-.0244397	.6902635
kid5	-.2812119	.1118409	-2.51	0.012	-.500416	-.0620078
phd	.0100783	.0802174	0.13	0.900	-.147145	.1673016
mentord	.5429222	.0747143	7.27	0.000	.3964848	.6893595
_cons	-.1553251	.3050814	-0.51	0.611	-.7532736	.4426234

The advantage of this approach is that interpretation is simpler, but to take advantage of this simplicity you must make the strong assumption that successive categories of the ordinal independent variable are equally spaced. For example, it implies that an increase from no publications by the mentor to a few publications involves an increase of the same amount of productivity as an increase from a few to some, from some to many, and from many to lots of publications.

Accordingly, before treating an ordinal independent variable as if it were interval, you should test whether this leads to a loss of information about the association between the independent and dependent variable. A likelihood-ratio test can be computed by comparing the model with only mentord to the model that includes both the ordinal variable (mentord) and all but two of the indicator variables. Below, we add some, many, and lots, but including any three of the indicators leads to the same results. If the categories of the ordinal variables are equally spaced, the coefficients of the $J-2$ indicator variables should all be 0. For example,

```
. logit hasarts fem mar kid5 phd mentord some many lots, nolog
  (output omitted)
. estimates store fmodel
. logit hasarts fem mar kid5 phd mentord, nolog
  (output omitted)
. lrtest fmodel
Likelihood-ratio test                          LR chi2(3)  =      1.07
(Assumption: . nested in fmodel)               Prob > chi2 =    0.7845
```

We conclude that the indicator variables do not add more information to the model ($LRX^2 = 1.07$, $df = 3$, $p = .78$). If the test were significant, we would have evidence that the categories of mentord are not evenly spaced, and so we should not treat mentord as interval. A Wald test can also be computed, leading to the same conclusion:

```
. logit hasarts fem mar kid5 phd mentord some many lots, nolog
  (output omitted)
. test some many lots
 ( 1)  some = 0
 ( 2)  many = 0
 ( 3)  lots = 0
           chi2(  3) =    1.03
         Prob > chi2 =    0.7950
```

9.1.4 Discrete change for categorical independent variables

There are a few tricks that you must be aware of when computing discrete change for categorical independent variables. To show how this is done, we will compute the change in the probability of publishing for those with a mentor with few publications compared with a mentor with no publications. There are two ways to compute this discrete change. The first way is easier, but the second is more flexible.

Computing discrete change with prchange

The easy way is to use prchange, where we set all the indicator variables to 0:

```
. logit hasarts fem mar kid5 phd few some many lots, nolog
  (output omitted)
. prchange few, x(some=0 many=0 lots=0)
logit: Changes in Probabilities for hasarts
      min->max      0->1      -+1/2     -+sd/2   MargEfct
few     0.0957     0.0957     0.0962     0.0399     0.0965
          NoArts     Arts
Pr(y|x)   0.4920   0.5080

          fem        mar        kid5       phd        few        some       many       lots
   x=   .460109    .662295    .495082    3.10311    .219672          0          0          0
sd(x)=  .498679    .473186    .76488     .984249    .414251    .478501    .422839    .293489
```

We conclude that having a mentor with a few publications compared with none increases a scientist's probability of publishing by .10, holding all other variables at their mean.

> The effect of the mentor's productivity is significant at the .01 level ($LRX^2 = 58.32$, $df = 4$, $p < .01$).

Even though we say "holding all other variables at their mean", which is clear within the context of reporting substantive results, the key to getting the right answer from prchange is holding all the indicator variables at 0, not at their mean. It does not make sense to change few from 0 to 1 when some, many, and lots are at their means.

mputing discrete change with prvalue

A second approach to computing discrete change is to use two calls to prvalue. The advantage of this approach is that it works in situations where prchange does not. For example, how does the predicted probability change if we compare a mentor with a few publications to a mentor with some publications, holding all other variables constant? This involves computing probabilities as we move from few=1 and some=0 to few=0 and some=1. We cannot compute this with prchange because two variables are changing at the same time. Instead, we use two calls to prvalue:[2]

```
. quietly prvalue, x(few=1 some=0 many=0 lots=0) save
. prvalue, x(few=0 some=1 many=0 lots=0) diff
logit: Change in Predictions for hasarts
Confidence intervals by delta method
                    Current      Saved     Change    95% CI for Change
Pr(y=Arts|x):        0.7125     0.5825     0.1300   [ 0.0452,   0.2147]
Pr(y=NoArts|x):      0.2875     0.4175    -0.1300   [-0.2147,  -0.0452]

            fem        mar        kid5       phd        few        some
Current=  .46010929  .66229508  .49508197  3.1031093          0          1
 Saved=   .46010929  .66229508  .49508197  3.1031093          1          0
  Diff=           0          0          0          0         -1          1

            many       lots
Current=       0          0
 Saved=        0          0
  Diff=        0          0
```

Because we have used the save and diff options, the difference in the predicted probability (i.e., the discrete change) is reported. When we use the save and diff options, we usually add quietly to the first prvalue because all the information is listed by the second prvalue.