The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- CompuMath Citation Index®
- Current Contents/Social and Behavioral Sciences®
- RePEc: Research Papers in Economics
- Science Citation Index Expanded (also known as SciSearch®)
- Scopus™
- Social Sciences Citation Index®

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata, Mata, NetCourse, and Stata Press are registered trademarks of StataCorp LP.

The Stata Journal (2012)
12, Number 1, pp. 162–164

Stata tip 106: With or without reference

Maarten L. Buis
Department of Sociology
Tübingen University
Tübingen, Germany
maarten.buis@uni-tuebingen.de

A convenient way to define a set of indicator variables (often called dummy variables) is to use Stata's factor-variable notation (see [U] **11.4.3 Factor variables**). In that case, the default is to leave one category out, the so-called reference category. However, the factor-variable notation also allows you to include an indicator variable for the reference category. This can provide a useful alternative representation of the same model. The estimation and interpretation of these models are best explained using examples, like the ones below.

```
. sysuse auto
(1978 Automobile Data)
. summarize weight if foreign == 0, meanonly
. generate c_weight = (weight - r(min))/2000
. label var c_weight "weight centered at lightest domestic car (short tons)"
. quietly regress price i.foreign c_weight
. estimates store a1
. quietly regress price ibn.foreign c_weight, noconstant
. estimates store b1
. estimates table a1 b1, b(%9.3g)
```

Variable	a1	b1
foreign		
0	(base)	1034
1	3637	4671
c_weight	6641	6641
_cons	1034	

In this example, the average price of "domestic" (U.S.) cars is compared with the average price of "foreign" cars while controlling for the weight of the car. Model `a1` uses the default method of using indicator variables. The results are interpreted as follows: the lightest domestic car costs on average $1,034, and an equally light foreign car costs on average $3,637 more. Model `b1` includes both an indicator variable for foreign cars and an indicator variable for domestic cars. These results are interpreted as follows: the lightest domestic car costs on average $1,034, and an equally light foreign car costs on average $4,671.

It is useful to note three things about these results. First, these models are completely equivalent; they are just different ways of saying the same thing. Model `a1` emphasizes the comparison of the categories, while model `b1` emphasizes the levels in

 st0250

each category. Second, the two indicator variables in model `b1` contain information that was present in the indicator variable and the constant in model `a1`. Thus in model `b1`, there is no information left to put in the constant. As a consequence, you must leave the constant out of model `b1`, which was done by adding the `noconstant` option. Third, it helps to center all variables in the model on some meaningful value. In this example, I centered the weight on the lightest domestic car. If I had not done that, then the prices in models `a1` and `b1` would refer to cars weighing 0 tons.

This trick can also be useful when you have interactions, as is shown in the example below. Model `a2` uses the default parameterization, which leaves out the reference category for both `foreign` and `good`. Model `b2` includes an indicator variable for the reference category of `foreign` but leaves the reference category out for `good`. Model `c2` contains indicator variables for all reference categories.

```
. generate byte good = rep78 > 3 if rep78 < .
(5 missing values generated)
. quietly regress price i.foreign##i.good c_weight
. estimates store a2
. quietly regress price i.foreign ibn.foreign#i.good c_weight, noconstant
. estimates store b2
. quietly regress price ibn.foreign#ibn.good c_weight, noconstant
. estimates store c2
. estimates table a2 b2 c2, b(%9.3g)
```

Variable	a2	b2	c2
foreign			
0	(base)	974	
1	3150	4124	
good			
1	-251		
foreign#good			
0 0	(base)	(base)	974
0 1	(base)	-251	723
1 0	(base)	(base)	4124
1 1	708	457	4581
c_weight	6711	6711	6711
_cons	974		

Consider models `a2` and `b2`. Model `a2` says that a bad, light domestic car will cost \$974, while a similar foreign car will cost \$3,150 more. Model `b2` says that the bad, light domestic car costs \$974, while a similar foreign car will cost \$4,124. Model `a2` says that good cars are \$251 cheaper if they are domestic cars, while the effect of being a good car increases by \$708 if the car is foreign. Model `b2` says that the effect of being a good car is −\$251 for domestic cars and \$457 for foreign cars.

Consider models `b2` and `c2`. Model `c2` says that bad, light domestic cars cost \$974, while good, light domestic cars cost \$723. Model `b2` says that bad, light domestic cars cost $974, while good domestic cars cost $251 less. Model c2 says that bad, light foreign cars cost $4,124, while good, light foreign cars cost $4,581. Model b2 says that bad, light foreign cars cost $4,124, while good foreign cars cost $457 more.

This trick is not limited to linear regression but can be applied to any model. For example, assume we are worried about the right-skewed nature of price and think that a log transformation would be better, but we want to continue making statements in terms of the average price and not in terms of the average log price. In that case, we can use glm with the link(log) option (see [R] **glm** and Cox et al. [2008]) or poisson (see [R] **poisson** and Wooldridge [2010]). An important difference with linear regression is that one interprets the exponentiated parameters, and these are interpreted in multiplicative terms rather than additive terms. Consider the example below. Model a3 says that a light domestic car will cost on average $2,102, while a similar foreign car will cost 2.145 times as much. Model b3 says that a light domestic car will cost on average $2,102, while a similar foreign car will cost on average $4,509.

```
. quietly glm price i.foreign c_weight, link(log) eform
. estimates store a3
. quietly glm price ibn.foreign c_weight, noconstant link(log) eform
. estimates store b3
. estimates table a3 b3, b(%9.4g) eform
```

Variable	a3	b3
foreign		
0	(base)	2102
1	2.145	4509
c_weight	3.516	3.516
_cons	2102	

References

Cox, N. J., J. Warburton, A. Armstrong, and V. J. Holliday. 2008. Fitting concentration and load rating curves with generalized linear models. *Earth Surface Processes and Landforms* 33: 25–39.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.