The Stata Journal (*forthcoming*) 12, Number 1, pp. 1–3

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 ([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 one 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 is best explained using examples, like the ones below.

```
. sysuse auto, clear
(1978 Automobile Data)

.
. sum weight if foreign == 0, meanonly

. gen c_weight = (weight - r(min))/2000

. label var c_weight "weight centered at lightest domestic car (short tons)"

.
. qui reg price i.foreign c_weight

. est store a1

. qui reg price ibn.foreign c_weight, nocons

. est store b1

. est tab 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' (US) cars is compared with the average price of 'foreign' cars while controlling for the weight of the car. Model a1 uses the default way of using indicator variables. The results are interpreted as: the lightest domestic car costs on average 1,034 US$ and an equally light foreign car costs on average 3,637 US$ more. Model b1 includes both an indicator variable for foreign cars and an indicator variable for domestic cars. These results are interpreted as: the lightest domestic car costs on average 1,034 US$ and an equally light foreign car costs on average 4,671 US$.

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 each category. Second, the two indicator variables in model b1 contain all information that was present in the indicator variable and the constant in model a1. So in model b1 there is no information left to put in the constant. As a consequence, one 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 weighting 0 tons.

This trick can also be useful when one has interactions, as is shown in the example below. Model a2 uses the default parametrization, 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.

```
. gen byte good = rep78 >3 if rep78 < .
(5 missing values generated)

.
. qui reg price i.foreign##i.good c_weight

. est store a2

. qui reg price i.foreign ibn.foreign#i.good c_weight, nocons

. est store b2

. qui reg price ibn.foreign#ibn.good c_weight, nocons

. est store c2

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

Consider models b2 and c2. Model c2 says that bad light domestic cars cost 974 US$ while good light domestic cars cost 723 US$. Model b2 says that bad light domestic cars cost 974 US$ while good domestic cars cost 251 US$ less. Model c2 says that bad light foreign cars cost 4,124 US$ while good light foreign cars cost 4,581 US$. Model b2 says that bad light foreign cars cost 4,124 US$ while good foreign cars cost 457 US$ 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 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 [R] **glm** with the `link(log)` option (Cox et al. 2008) or [R] **poisson** (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 US$ 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 US$ while a similar foreign car will cost on average 4,509 US$.

```
. qui glm price i.foreign c_weight, link(log) eform

. est store a3

. qui glm price ibn.foreign c_weight, nocons link(log) eform

. est store b3

. 
. est tab 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.* Cambridge, MA: MIT Press.