Stata tip 1: the eform() option of regress

Roger Newson
King’s College London, UK
roger.newson@kcl.ac.uk
http://www.kcl-phs.org.uk/rogernewson

Did you know about the eform() option of regress? It is very useful for calculating confidence intervals for geometric means and their ratios. These are frequently used with skewed Y-variables, such as house prices and serum viral loads in HIV patients, as approximations for medians and their ratios. In Stata, I usually do this by using the regress command on the logs of the Y-values, with the eform() and noconstant options. For instance, in the auto data, we might compare prices between non-US and US cars as follows:

```
. sysuse auto,clear
(1978 Automobile Data)
. gene logprice=log(price)
. gene byte baseline=1
. regress logprice foreign baseline,noconst eform(GM/Ratio) robust
```

```
  Regression with robust standard errors
Number of obs = 74
F( 2, 72) =18043.56
Prob > F = 0.0000
R-squared = 0.9980
Root MSE = .39332

            | Robust
  logprice   | GM/Ratio Std. Err. t P>|t| [95% Conf. Interval]
-------------|--------------------------------------------------------------
     foreign | 1.07697 .103165 0.77 0.441 .8897576 1.303573
     baseline| 5533.565 310.8747 153.41 0.000 4947.289 6189.316
```

We see from the baseline parameter that US-made cars had a geometric mean price of 5534 dollars (95% CI from 4947 to 6189 dollars), and we see from the foreign parameter that non-US cars were 108% as expensive (95% CI, 89% to 130% as expensive). An important point is that, if you want to see the baseline geometric mean, then you must define the constant variable baseline and enter it into the model with the noconst option. Stata usually suppresses the display of the intercept when we specify the eform() option, and this trick will fool Stata into thinking that there is no intercept for it to hide. (The same trick can be used with logit using the or option, if you want to see the baseline odds as well as the odds ratios.)

I find that my non-statistical colleagues understand regression models for log-transformed data a lot better this way than any other way. Continuous X-variables can also be included, in which case the parameter for each X-variable is a ratio of Y-values per unit change in X, assuming an exponential relationship. (Or assuming a power relationship, if X is itself log-transformed.)

© 2001 Stata Corporation