The Stata Journal (2017)
17, Number 3, pp. 774–778

Stata tip 128: Marginal effects in log-transformed models: A trade application

Luca J. Uberti
Department of Politics
University of Otago
Dunedin, New Zealand
luca.jacopo.uberti@gmail.com

Since the introduction of the `margins` command in Stata 11, the empirical literature has increasingly used marginal effects, predictive margins, and adjusted predictions in postestimation analysis. Marginal effects are particularly useful for the interpretation of parameter estimates after `logit`, `probit`, `poisson`, and other nonlinear regression models. If the covariate of interest is in logs, however, obtaining meaningful results from `margins, dydx()` is not straightforward. In this article, I first illustrate these difficulties in the context of estimation with `poisson`. I then suggest that a researcher should always compute the derivative of interest and code it manually with `margins`'s `expression()` option. Lastly, I illustrate these problems using the gravity equation from the trade literature.

I focus on `poisson` for two reasons. First, most examples of `margins` tend to focus on `logit` and `probit`. Second, it is increasingly common for researchers to fit Poisson models with log-transformed covariates. Going beyond traditional applications with count data, some recent literature has shown that `poisson` should be the preferred estimator for constant-elasticity models such as the gravity equation or the Cobb–Douglas production function (Manning and Mullahy 2001; Santos Silva and Tenreyro 2006). In these applications, at least some of the variables on the right-hand side are log transformed.

Let's assume a Poisson model with the following conditional expectation, or "response":

$$E(y_i|x_i) = e^{a+bx_i}$$

y_i and x_i are random variables, b is a vector of estimated parameters, and a is a constant. In Stata language, the "effect" of a covariate x_i is defined as the derivative of the response with respect to x_i:[1]

$$\frac{dE}{dx_i} = be^{a+bx_i}$$

The marginal effect, or average marginal effect (AME), is then the predictive margin (or simply "margin") of this effect. Margins are obtained by averaging or integrating a response over all n observations (or conditions) in the sample (StataCorp 2015, 1404):

$$\text{AME}(x_i) = \frac{1}{n}\sum_{i}^{n} be^{a+bx_i}$$

1. The derivative of the response is itself a response.

 st0497

In the absence of log transformations, a researcher can obtain this statistic by simply typing `margins, dydx(xi)` after fitting the model (`poisson yi xi`). Things are more complicated, however, when the variable of interest is in logs. If the conditional expectation is specified as $E(y_i|x_i) = e^{a+b\ln x_i}$, the researcher must create a new Stata variable (`gen ln_xi = log(xi)`) before fitting the log-transformed model by typing `poisson yi ln_xi`. The AME then becomes

$$\text{AME}(x_i) = \frac{e^a}{n}\sum_i^n \frac{d}{dx_i}(e^{b\ln x_i}) = \frac{e^a}{n}\sum_i^n bx_i^{b-1} \tag{1}$$

Because `xi` no longer appears in the estimation command, this AME cannot be recovered by typing `margins, dydx(xi)`, which is the case after running the model without log transformations. The command `margins, dydx(ln_xi)`, which Stata does accept, instead computes

$$\text{AME}(\ln x_i) = \frac{e^a}{n}\sum_i^n \frac{d}{d\ln x_i}(e^{b\ln x_i}) = \frac{e^a}{n}\sum_i^n bx_i^{b}$$

This statistic, however, is not the AME of x_i, nor is it particularly useful for purposes of interpretation.[2]

How can researchers compute the correct AME after fitting models with log transformations? One possible solution is the `margins` command's `expression()` option. In what follows, I illustrate how this option can help researchers interpret the results of constant elasticity models such as the gravity equation. Using trade data from Santos Silva and Tenreyro (2006), I fit a gravity model of trade with form $T_{ij} = gY_i^{\alpha_1}Y_j^{\alpha_2}\tau_{ij}^{\beta}$, where T_{ij} is bilateral trade, Y_i and Y_j are an exporter's and importer's gross domestic product (GDP), respectively, τ_{ij} are bilateral trade costs, and $g = e^G$ is a constant. Trade costs (transportation and tariff costs) are operationalized using geographical distance and an indicator variable that equals one if i and j are members of a free trade agreement. Log-linearizing the gravity equation and exponentiating through, I obtain an expression that may be estimated by `poisson`,

$$T_{ij} = \exp(G + \alpha_1\texttt{lypex}_i + \alpha_2\texttt{lypim}_j + \beta_1\texttt{ldist}_{ij} + \beta_2\texttt{comfrt}_{ij}) \times \eta_{ij}$$

where `lypex` and `lypim` are, respectively, the log of an exporter's and importer's GDP, `ldist` is the log of geographical distance, `comfrt` is the free trade agreement dummy, and η_{ij} is a multiplicative error term that is uncorrelated with the regressors. The Stata output is displayed below:

2. The same problem arises in an ordinary least-squares context when fitting log-linear or linear-log models.

```
. copy "http://personal.lse.ac.uk/tenreyro/regressors.zip" any_name.zip, public
. unzipfile any_name.zip
    inflating: Log of Gravity.do
    inflating: coding.xls
    inflating: Log of Gravity.dta
    inflating: countrycodes.xls
    inflating: Log of Gravity.xls
successfully unzipped any_name.zip to current directory
. use "Log of Gravity.dta"
(FROM: Santos Silva, J & Tenreyro, S 2006 The log of gravity, RESTAT 88, 641-658)
. poisson trade lypex lypim ldist comfrt, vce(robust) nolog
Poisson regression                              Number of obs     =     18,360
                                                Wald chi2(4)      =    5202.50
                                                Prob > chi2       =     0.0000
Log pseudolikelihood = -1.202e+09               Pseudo R2         =     0.8944

------------------------------------------------------------------------------
             |               Robust
       trade |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lypex |   .8031383   .0238621    33.66   0.000     .7563694    .8499072
       lypim |   .7982473   .0260628    30.63   0.000     .7471652    .8493294
       ldist |  -.6612203   .0532762   -12.41   0.000    -.7656396   -.5568009
      comfrt |   .3659858   .1738048     2.11   0.035     .0253347    .7066369
       _cons |   -23.2916   1.206315   -19.31   0.000    -25.65594   -20.92727
------------------------------------------------------------------------------
```

The coefficient on `lypex` implies a 0.8% increase in bilateral trade following a 1% increase in country i's GDP. This relation holds at any value of the covariates. Though mathematically simple, this interpretation does not always provide substantively useful insights. The policymaker, for instance, may wish to know exactly by how much we might expect bilateral trade to increase following a spell of rapid economic growth and increased consumer demand in i. To answer this question, we cannot simply type `margins, dydx(lypex)` for the reasons explained above.

The challenge is computing the AME of Y_i:

$$\text{AME}(Y_i) = \frac{1}{n}\sum_{i}^{n}\left(e^G \alpha_1 Y_i^{(\alpha_1 - 1)} Y_j^{\alpha_2} \tau_{ij}^{\beta}\right)$$

Because $\text{AME}(Y_i)$ is a nonlinear combination of both the parameters and the data, the calculation may be accomplished with the `predictnl` command:

```
. predictnl AME = (1/18360)*sum(exp(_b[_cons])*_b[lypex]*(exp(lypex)^(_b[lypex]-1))
> *(exp(lypim)^_b[lypim])*(exp(ldist)^_b[ldist])*exp(_b[comfrt]*comfrt))
. list AME in 18360/18360

       +----------+
       |      AME |
       |----------|
18360. | 1.56e-06 |
       +----------+
```

This implies that on average, an increase of 1 billion U.S. dollars in i's GDP will lead to an increase in bilateral trade worth 156 million U.S. dollars. We might also wish to know the standard errors of this prediction—and thus whether the trade effects of GDP growth are statistically significant. In principle, the **predictnl** command allows the researcher to specify the **se()** and **p()** options to compute the standard errors and *p*-values of the prediction. In practice, however, the algorithm of the **predictnl** command may fail to converge. While users may increase the number of iterations with the **iterate()** option, convergence may still not be achieved in large datasets, which is the case in the example above.

A more tractable solution is **margins, expression()**,[3] which allows the explicit specification of a formula for the (derivative of the) response to be margined:

```
. margins, expression(exp(_b[_cons])*_b[lypex]*(exp(lypex)^(_b[lypex]-1)
> *(exp(lypim)^_b[lypim])*(exp(ldist)^_b[ldist])*exp(_b[comfrt]*comfrt))
Warning: expression() does not contain predict() or xb().

Predictive margins                              Number of obs     =     18,360
Model VCE    : Robust

Expression
                : exp(_b[_cons])*_b[lypex]*(exp(lypex)^(_b[lypex]-1))*
       > (exp(lypim)^_b[lypim])*(exp(ldist)^_b[ldist])*
       > exp(_b[comfrt]*comfrt)
```

	Margin	Delta-method Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	1.56e-06	1.24e-07	12.57	0.000	1.32e-06	1.81e-06

In contrast to the syntax for **predictnl**, here we do not need to average the effect (the derivative of the response) by including **(1/18360)*sum()** in the argument of **expression()**. After all, averaging over (some of or all) the covariates is precisely what "taking margins" means.

In sum, the researcher should exercise care when using **margins** to compute the marginal effect of a log-transformed covariate because the **margins** command does not provide an option to exponentiate a logged variable and retrieve the original variable in levels.[4] Thus researchers have no alternative but to calculate the derivative of the response manually and code it into the argument of the **margins** command's **expression()** option. Because manually coding the derivative of the response may quickly become unwieldy in the presence of a large number of variables, the Stata developers should consider updating the **margins** command and including options to exponentiate the variables appearing in the previous regression command. A possible solution might be to allow the argument of **dydx()** to include functions of covariates. For example, it should be possible to compute the AME given by (1) by typing **margins, dydx(exp(ln_xi))**.

3. Another advantage of using margins, expression() over predictnl is that the former, but not the latter, integrates over unobserved components after models with random effects, for example, mepoisson (see [ME] **mepoisson**). This was a new feature in Stata 14.
4. By contrast, the margins command does provide users with options to log-transform either the variate (eydx) or the covariate of interest (dyex), or both (eyex).

1 References

Manning, W. G., and J. Mullahy. 2001. Estimating log models: To transform or not to transform? *Journal of Health Economics* 20: 461–494.

Santos Silva, J. M. C., and S. Tenreyro. 2006. The log of gravity. *Review of Economics and Statistics* 88: 641–658.

StataCorp. 2015. *Stata 14 Base Reference Manual*. College Station, TX: Stata Press.