The Stata Journal (*forthcoming*) 12, Number 1, pp. 1–2

Stata tip 107: the baseline is now reported

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

For a long time Stata has had the capability to report exponentiated coefficients. Examples are the `or` option in [R] **logit** and [R] **ologit**, the `irr` option in [R] **poisson**, [R] **zip**, and [R] **nbreg**, and the `hr` and `tr` options in [ST] **streg** (also see: Newson 2003; Buis 2010). These exponentiated coefficients can be interpreted as odds ratios, incidence rate ratios, hazard ratios or time ratios. However, until Stata 12 the baseline odds, incidence rate, hazard or time — that is, the exponentiated constant — was not reported. That was unfortunate as this baseline can be helpful for evaluating the size of the effects and provides a convenient way of discussing the exact interpretation of the coefficients. As of Stata 12 this omission has been redressed. The usefulness of the baseline value and a couple of caveats are illustrated using the example below.

```
. sysuse nlsw88, clear
(NLSW, 1988 extract)
. gen c_grade = grade - 12
(2 missing values generated)
. gen high_occ = occupation < 3 if occupation < .
(9 missing values generated)
. logit union c_grade i.high_occ, or nolog
Logistic regression                             Number of obs   =       1867
                                                LR chi2(2)      =      49.44
                                                Prob > chi2     =     0.0000
Log likelihood = -1016.5579                     Pseudo R2       =     0.0237
```

union	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
c_grade	1.123325	.0248694	5.25	0.000	1.075624	1.173141
1.high_occ	.4651723	.0644307	-5.53	0.000	.3545803	.6102575
_cons	.3358115	.02213	-16.56	0.000	.2951218	.3821112

Odds ratios have a bad reputation for being hard to interpret. Part of the problem is that many people are not used to working with odds. Researchers rarely frequent racetracks or betting shops. Starting the results section of an article with interpreting the baseline odds is a nice way to remind the readers of the right interpretation. This trick works well because it fits naturally within the normal format of an academic article. In this case we expect to find 0.34 union members for every non-member within the group of respondents that has 12 years of education (`c_grade` = 0) and a lower occupation (`high_occ` = 0). The odds ratios tell us that that odds increases by a factor of 1.12 or 12% ($(1.12 - 1) \times 100\% = 12\%$) for every additional year of education, while the odds decreases 53% ($(0.47 - 1) \times 100\% = -53\%$) when the respondent has a high

occupation. Reporting the baseline odds in the results section of a paper allows one to translate the abstract concept of odds to the concrete situation that is being studied, in this case translate 'the number of successes per failure' to 'the number of union members per non-member'.[1]

A 53% decrease in the odds of being a union members sounds like a large effect. However, we can get a better understanding of the size of this effect by comparing it with the baseline odds. In this case the odds changes from 0.34 union members per non-member for respondents with lower occupations to 0.16 ($0.47 \times 0.34 = 0.16$) union members per non-member, which is a substantively meaningful change. But what if being a union member was very rare? For example, assume that the baseline odds was 0.001 union member for every non-union member. In that case the odds would change from 0.001 to 0.00047 union members per non-member when a respondent obtained a high occupation, which does not sound nearly as impressive as a change of -53%. So the baseline value can play an important role in evaluating how large an effect is.

There are however a couple of things one needs to consider when interpreting these baseline values. First, the baseline value is the value when all explanatory variables are 0. So, in order to get a meaningful baseline value one needs to make sure the value 0 is meaningful for all explanatory variables. In the example above I did so by centering the variable `grade` at 12 years of education (obtaining high school). Second, the practice of reporting p-values or assigning stars to significant parameters needs a bit of thought in the case of baseline values. Stata automatically reports the results of the test of the null-hypothesis that the coefficient is 0 and thus the exponentiated coefficient is 1. So in the example that would mean that the null hypothesis for the baseline odds is that there is 1 union member for every non-member, that is, the probability of being a union member is 50%.

References

Agresti, A. 2007. *An Introduction to Categorical Data Analysis*. 2nd ed. New York, NY: Wiley.

Buis, M. 2010. Stata tip 87: Interpretation of interactions in non-linear models. *The Stata Journal* 10: 305–308.

Fienberg, S. E. 2007. *The Analysis of Cross-Classified Categorical Data*. 2nd ed. New York, NY: Springer.

Newson, R. 2003. Stata tip 1: The eform() option of regress. *The Stata Journal* 3: 445.

1. More complete discussions of odds and odds ratios can be found in the textbooks by Fienberg (2007) and Agresti (2007).