The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

The Stata Journal

Editor
H. Joseph Newton
Department of Statistics
Texas A & M University
College Station, Texas 77843
979-845-3142; FAX 979-845-3144
jnewton@stata-journal.com

Editor
Nicholas J. Cox
Department of Geography
Durham University
South Road
Durham City DH1 3LE UK
n.j.cox@stata-journal.com

Associate Editors

Christopher F. Baum, Boston College
Rino Bellocco, Karolinska Institutet, Sweden and Univ. degli Studi di Milano-Bicocca, Italy
A. Colin Cameron, University of California–Davis
David Clayton, Cambridge Inst. for Medical Research
Mario A. Cleves, Univ. of Arkansas for Medical Sciences
William D. Dupont, Vanderbilt University
Charles Franklin, University of Wisconsin–Madison
Joanne M. Garrett, University of North Carolina
Allan Gregory, Queen's University
James Hardin, University of South Carolina
Ben Jann, ETH Zürich, Switzerland
Stephen Jenkins, University of Essex
Ulrich Kohler, WZB, Berlin
Jens Lauritsen, Odense University Hospital
Stanley Lemeshow, Ohio State University
J. Scott Long, Indiana University
Thomas Lumley, University of Washington–Seattle
Roger Newson, Imperial College, London
Marcello Pagano, Harvard School of Public Health
Sophia Rabe-Hesketh, University of California–Berkeley
J. Patrick Royston, MRC Clinical Trials Unit, London
Philip Ryan, University of Adelaide
Mark E. Schaffer, Heriot-Watt University, Edinburgh
Jeroen Weesie, Utrecht University
Nicholas J. G. Winter, University of Virginia
Jeffrey Wooldridge, Michigan State University

Stata Press Production Manager Lisa Gilmore
Stata Press Copy Editor Gabe Waggoner

Copyright Statement: The Stata Journal and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp LP. The contents of the supporting files (programs, datasets, and help files) may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the Stata Journal.

The articles appearing in the Stata Journal may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the Stata Journal.

Written permission must be obtained from StataCorp if you wish to make electronic copies of the insertions. This precludes placing electronic copies of the Stata Journal, in whole or in part, on publicly accessible web sites, fileservers, or other locations where the copy may be accessed by anyone other than the subscriber.

Users of any of the software, ideas, data, or other materials published in the Stata Journal or the supporting files understand that such use is made without warranty of any kind, by either the Stata Journal, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the Stata Journal is to promote free communication among Stata users.

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

The Stata Journal (2007)
7, Number 2, pp. 221–226

predict and adjust with logistic regression

Maarten L. Buis
Department of Social Research Methodology
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
m.buis@fsw.vu.nl

Abstract. Within Stata there are two ways of getting average predicted values for different groups after an estimation command: `adjust` and `predict`. After OLS regression (`regress`), these two ways give the same answer. However, after logistic regression, the average predicted probabilities differ. This article discusses where that difference comes from and the consequent subtle difference in interpretation.

Keywords: st0127, adjust, predict, logistic regression

1 Introduction

A useful way of interpreting the results from a regression model is to compare predicted values from different groups. Within Stata both `adjust` and `predict` can be used after an estimation command to set up values at which predictions are desired and then display those predictions in a tabular form. In a Stata frequently asked question, Poi (2002) showed the following example:

```
. sysuse auto
(1978 Automobile Data)
. regress mpg weight length foreign
  (output omitted)
. adjust, by(rep78)
```

```
    Dependent variable: mpg     Command: regress
  Variables left as is: weight, length, foreign
```

Repair Record 1978	xb
1	21.3651
2	19.3989
3	19.9118
4	21.86
5	24.9181

Key: xb = Linear Prediction

In this example, `adjust` shows the average predicted mileage for different states of repair. To show that this is exactly what `adjust` does, Poi actually computes the

© 2007 StataCorp LP

predicted mileage for each observation with `predict` and then shows that the averages for each state of repair corresponds exactly to the output from `adjust`.

```
. predict yhat, xb
. tabstat yhat, statistics(mean) by(rep78)
Summary for variables: yhat
     by categories of: rep78 (Repair Record 1978)
```

rep78	mean
1	21.36511
2	19.39887
3	19.91184
4	21.86001
5	24.91809
Total	21.20081

However, when the same procedure is applied to predicted probabilities from logistic regression, the average predicted probabilities no longer match the output from `adjust`. The aim of this article is to explain where that difference comes from and to discuss the resulting difference in interpretation of the results from `adjust` and `predict`.

```
. use http://www.stata-press.com/data/r9/lbw, clear
(Hosmer & Lemeshow data)
. gen black = race==2
. gen other = race==3
. logit low age lwt black other smoke
  (output omitted)
. predict p
(option p assumed; Pr(low))
. tabstat p, statistics(mean) by(ht)
Summary for variables: p
     by categories of: ht (has history of hypertension)
```

ht	mean
0	.3154036
1	.2644634
Total	.3121693

```
. adjust, pr by(ht)
------------------------------------------------------------------------------
     Dependent variable: low     Command: logit
   Variables left as is: age, lwt, smoke, black, other
------------------------------------------------------------------------------

----------------------
has        |
history    |
of         |
hypertens  |
ion        |         pr
-----------+----------
         0 |    .291936
         1 |    .251055
----------------------
     Key:  pr  =  Probability
```

2 Computing predicted probabilities that involve a nonlinear transformation

The key in understanding this difference is noticing that getting predicted probabilities from logistic regression requires a nonlinear transformation. In the example, `logit` modeled the probability of getting a child with low birthweight according to (1).

$$\Pr(\texttt{low} = 1) = \frac{e^{xb}}{1 + e^{xb}} \tag{1}$$

whereby xb is usually called the linear predictor and is given by

$$xb = \beta_0 + \beta_1\texttt{age} + \beta_2\texttt{lwt} + \beta_3\texttt{black} + \beta_4\texttt{other} + \beta_5\texttt{smoke}$$

Once the model is fitted, computing the predicted probabilities involves two steps. First, the predicted values for the linear predictor are calculated. Next the linear predictor is transformed to the probability metric by using (2). Predicted values are identified by a $\widehat{\ }$ on top of their symbol.

$$\widehat{\Pr} = \frac{e^{\widehat{xb}}}{1 + e^{\widehat{xb}}} \tag{2}$$

The difference between `predict` and `adjust` is that `predict` first applies the transformation to the linear predictor and then computes the mean, whereas `adjust` first computes the mean of the linear predictor and then applies the transformation (see [R] **adjust**). To see why this matters, first look at a special case where it does not matter, such as when xb is distributed symmetrically around 0 (figure 1). It shows that the transformation "squeezes" the values of xb on the unit interval. Furthermore, it squeezes

values further away from zero harder than it does values closer to zero. So in the transformed metric the smallest value became less extreme because it got squeezed a lot. Remember that extreme values influence the mean more than less extreme values. So, the lowest value exerts less influence on the mean in the transformed probability metric than in the original linear predictor metric. However, the change in mean due to the loss of influence of the lowest value was exactly balanced by the change in mean due to the loss of influence from the largest value, since the linear predictor was symmetrically distributed around zero.

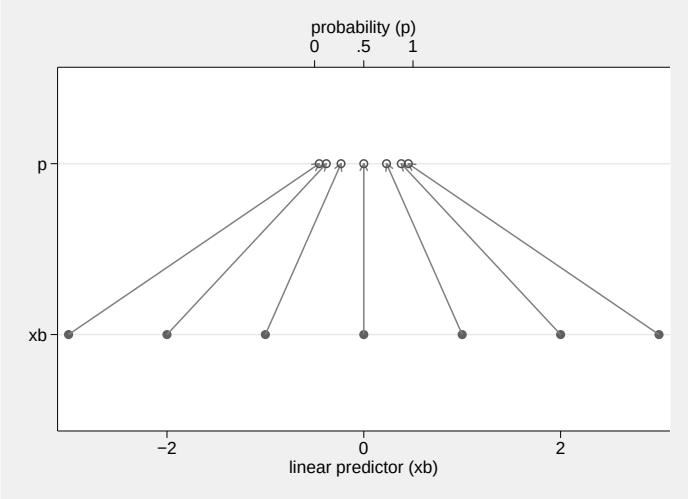

Figure 1: Logit transformation if xb is symmetric around 0

The likelihood that a real model on real data will yield a distribution of linear predictors that are symmetric around zero is extremely small. Figure 2 shows what happens if the distribution is asymmetric around 0. The loss in influence for the largest values is not balanced by the loss of influence for the smallest values. As a consequence, the largest values exert more influence on the mean in the original linear predictor metric than in the transformed probability metric. So, for figure 2, those who first compute the mean and then transform (i.e., use `adjust`) will find a larger probability than those who first transform and then compute the mean (i.e., use `predict`).

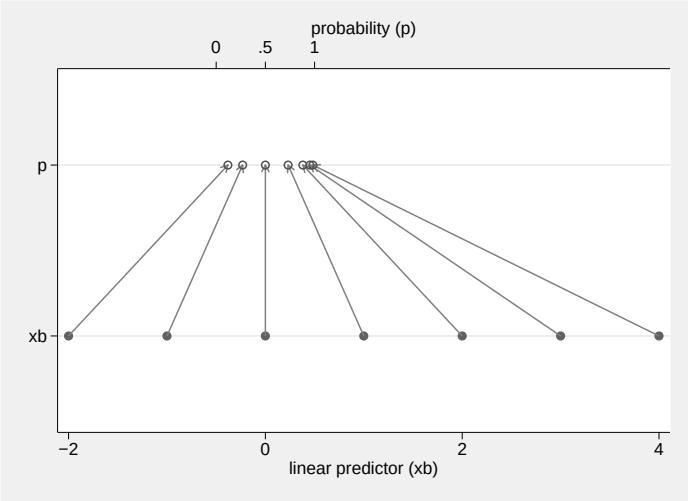

Figure 2: Logit transformation if xb is asymmetric around 0

This point is not unique to getting predicted probabilities after logistic regression. Any transformation of the linear predictor that squeezes some parts more than others will show this behavior. To be exact, whether the mean is computed before or after the transformation will matter for any nonlinear transformation of the linear predictor, i.e., any transformation other than adding, subtracting, multiplying, and dividing. Calculating predicted values for models like binomial probit regression, multinomial logistic regression, ordered logistic regression, and any generalized linear model with a link function other than the identity function will involve a nonlinear transformation, so the same argument applies. Similarly, computing marginal effects in these models will typically involve a nonlinear transformation of the linear predictor, so again the same argument applies. Bartus (2005) has discussed this latter point.

3 What does this difference mean?

To make sense of this difference, it is helpful to see that the average linear predictor is the linear predictor for someone with average values on its explanatory variables. Equations (3)–(6) show why.

(*Continued on next page*)

$$\begin{aligned}
\overline{xb}_k &= \frac{\sum_k \beta_0 + \beta_1 x_1 + \beta_2 x_2}{N_k} && (3)\\
&= \frac{\sum_k \beta_0}{N_k} + \frac{\sum_k \beta_1 x_1}{N_k} + \frac{\sum_k \beta_2 x_2}{N_k} && (4)\\
&= \frac{N_k \beta_0}{N_k} + \beta_1 \frac{\sum_k x_1}{N_k} + \beta_2 \frac{\sum_k x_2}{N_k} && (5)\\
&= \beta_0 + \beta_1 \overline{x}_{1k} + \beta_2 \overline{x}_{2k} && (6)
\end{aligned}$$

Say that we have two explanatory variables, x_1 and x_2, and we want to compute the mean linear predictor for group k. Equation (3) is just the definition of that mean. Equation (4) shows that that fraction can be broken up. Equation (5) is based on the fact that the β's are constant, so they can be moved outside the summation sign. And finally (6) is again based on the definition of the mean. $\overline{x}_{1k}$ and $\overline{x}_{2k}$ are the means for group k only, not the overall means. `adjust` does have facilities to fix (some of) the explanatory variables at their grand mean, or other values, but I do not discuss that here.

There is therefore a subtle difference in interpretation between the results of `predict` and `adjust`. If we return to our logistic regression example and look at someone with hypertension, then `predict` will give us the average predicted probability for someone with hypertension, whereas `adjust` will give us the predicted probability for someone with average values on `age`, `lwt`, `black`, `other`, and `smoke` for someone with hypertension. It is the difference between a typical predicted probability for someone within a group and the predicted probability for someone with typical values on the explanatory variables for someone within that group.

4 Acknowledgments

I thank Fred Wolfe, whose question on Statalist provided the initial spark that resulted in this article, and an anonymous referee for useful comments.

5 References

Bartus, T. 2005. Estimation of marginal effects using margeff. *Stata Journal* 5: 309–329.

Poi, B. P. 2002. FAQ: Predict and adjust.
http://www.stata.com/support/faqs/stat/adjust.html.

About the author

Maarten L. Buis is a PhD student at the Department of Social Research Methodology of the Vrije Universiteit Amsterdam, where he studies long-term trends in inequality of educational opportunity between children of different socioeconomic background. He is also a frequent contributor to Statalist.