The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$115	1-year subscription	$145
2-year subscription	$210	2-year subscription	$270
3-year subscription	$285	3-year subscription	$375
1-year student subscription	$ 85	1-year student subscription	$115
1-year institutional subscription	$345	1-year institutional subscription	$375
2-year institutional subscription	$625	2-year institutional subscription	$685
3-year institutional subscription	$875	3-year institutional subscription	$965
Electronic only		**Electronic only**	
1-year subscription	$ 85	1-year subscription	$ 85
2-year subscription	$155	2-year subscription	$155
3-year subscription	$215	3-year subscription	$215
1-year student subscription	$ 55	1-year student subscription	$ 55

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.


The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, STATA, Stata Press, Mata, MATA, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2015)
15, Number 2, pp. 599–604

Stata tip 125: Binned residual plots for assessing the fit of regression models for binary outcomes

Jessica Kasza
Monash University
Melbourne, Australia
jessica.kasza@monash.edu

Plots based on residuals, such as plots of residual-versus-fitted values, are now standard after fitting linear regression models. These plots are used to assess the validity of assumptions, to identify features not captured by the model, and to find problematic data points or clusters. However, such plots are typically not very useful for regression models for binary outcomes because of the discrete nature of residuals from these models. Binned residual plots, as recommended by Gelman and Hill (2007), can be used to assess both the overall fit of regression models for binary outcomes (for example, logistic or probit models) and the inclusion of continuous variables. I demonstrate the construction of such plots in Stata. These binned residual plots are related to those produced by the `rbinplot` command from the `modeldiag` package (Cox [2004], updated in Cox [2010]), with the addition of approximate confidence limits.

To construct a binned residual plot to assess the overall fit of a logistic regression model, one orders predicted probabilities from smallest to largest and calculates residuals. Data are split into bins containing equal numbers of observations (a recommended number of bins is the square root of the number of observations), and the average residual is plotted against the average predicted probability for each bin. For each bin, approximate 95% confidence limits are $\pm 2\sqrt{p(1-p)/n}$, estimated using the standard deviation of each bin's residuals.

If the model is correct, about 95% of the points are expected to lie within the confidence limits. As is the case for a residual-versus-fitted plot used for linear regression, departures from random scatter are indicative that the fitted model does not accurately describe the data. To assess the fit of a continuous covariate, one orders observations and constructs bins in terms of that covariate instead of in terms of the predicted probabilities. The average residual is then plotted against the average covariate in each bin.

To demonstrate the construction of these plots, I simulate an example dataset consisting of 5,000 observations, where the binary outcome (or response variable) is dependent on two continuous covariates and the square of one of these covariates:

```
. set obs 5000
obs was 0, now 5000
. set seed 86206
. generate x1 = rnormal()
. generate x2 = rnormal()
. generate prob_y = exp(-1+x1+x2+x1^2)/(1+exp(-1+x1+x2+x1^2))
. generate y = rbinomial(1, prob_y)
```


To demonstrate the usefulness of binned residual plots, I omit the squared term from the logistic regression model for the binary outcome.

```
. quietly logit y x1 x2
. predict pred_y, pr
. generate resid = y - pred_y
```

After I obtain the predicted probabilities for each observation, the construction of the binned residual plot proceeds as follows:

```
. sort pred_y
. generate myids = _n if pred_y < .
. local nbins = floor(sqrt(5000))
. egen binno = cut(myids) if pred_y < . , group(`nbins´) icodes
. egen avefit = mean(pred_y), by(binno)
. egen myaveres = mean(resid), by(binno)
. egen mysd = sd(resid), by(binno)
. egen mytag = tag(binno)
. bysort binno: egen binsize = count(pred_y)
. generate uplim = 2*mysd/sqrt(binsize)
. generate dwlim = -2*mysd/sqrt(binsize)
. graph twoway (scatter myaveres avefit if inrange(myaveres, dwlim, uplim)
> & mytag == 1, msymbol(oh))
> (line uplim avefit, clcolor(gray) lstyle(solid))
> (line  dwlim avefit, clcolor(gray) lstyle(solid))
> (scatter myaveres avefit if !inrange(myaveres, dwlim, uplim) & mytag == 1,
> mcolor(black)),
> xtitle(Average predicted mortality probability) ytitle(Average residual)
> legend(off) scheme(sj)
```

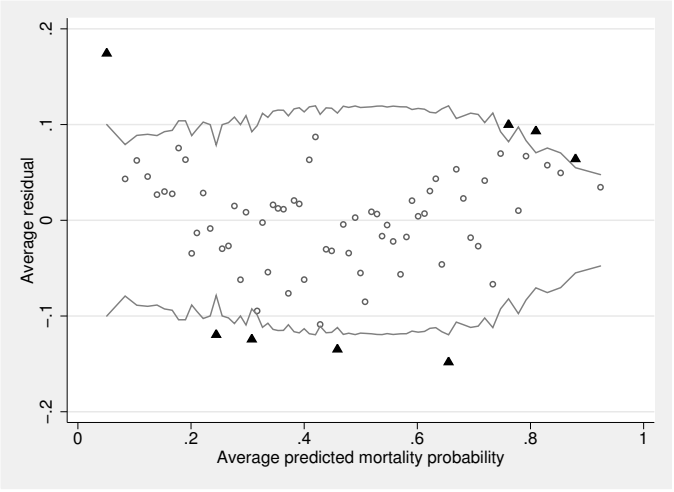

Figure 1. Binned residual plot to assess the overall fit of the model

Figure 1 displays the produced binned residual plot. There is some curvature to the pattern of binned residuals, although this is not particularly extreme. However, assessing the fit of the model with respect to `x1` does indicate problems.

```
. keep y x1 x2 pred_y resid
. sort x1
. generate myids = _n if pred_y < .
. local nbins = floor(sqrt(5000))
. egen binno = cut(myids) if pred_y < . , group(`nbins´) icodes
. egen avex1 = mean(x1), by(binno)
. egen myaveres = mean(resid), by(binno)
. egen mysd = sd(resid), by(binno)
. egen mytag = tag(binno)
. bysort binno: egen binsize = count(pred_y)
. generate uplim = 2*mysd/sqrt(binsize)
. generate dwlim = -2*mysd/sqrt(binsize)
. graph twoway (scatter myaveres avex1  if inrange(myaveres, dwlim, uplim) &
> mytag == 1, msymbol(oh))
> (line uplim avex1, clcolor(gray) lstyle(solid))
> (line dwlim avex1, clcolor(gray) lstyle(solid))
> (scatter myaveres avex1 if !inrange(myaveres, dwlim, uplim) & mytag == 1,
> mcolor(black)),
> xtitle(Average x1) ytitle(Average residual) legend(off) scheme(sj)
```

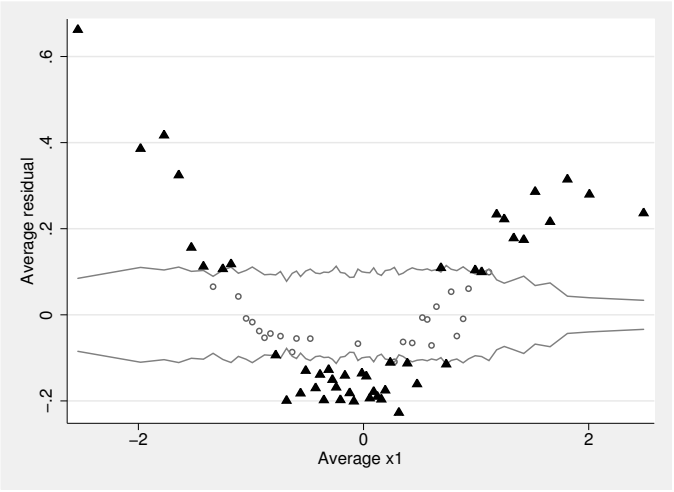

Figure 2. Binned residual plot to assess the fit of `x1`

The systematic pattern in figure 2 indicates that nonlinear terms for `x1` should be included in the logistic regression model. We fit the model including a quadratic term for `x1` as follows:

```
. keep y x1 x2
. qui logit y x1 x2 c.x1#c.x1 c.x2#c.x2
. predict pred_y
(option pr assumed; Pr(y))
. generate resid = y - pred_y
```

For this model, binned residual plots are constructed as above and displayed in figure 3. As is expected, a few points lie outside the confidence limits, but there are no systematic patterns in the plots.

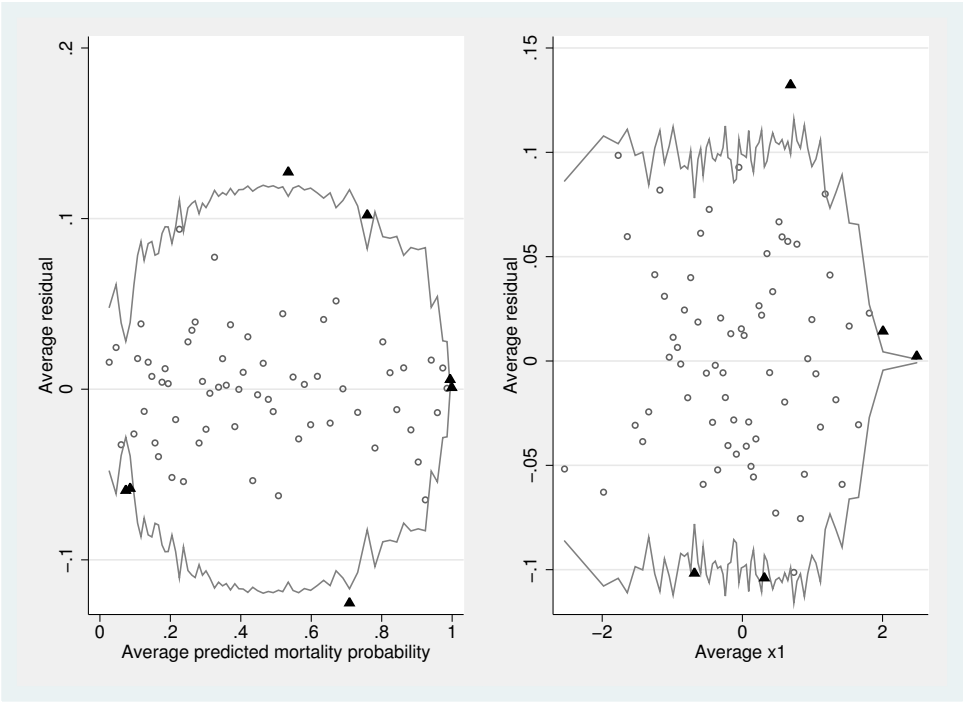

Figure 3. Binned residual plots for the model containing a quadratic term for `x1`

To illustrate the usefulness of binned residual plots, we consider the Medpar dataset from Hilbe (2009), available at http://www.crcpress.com/product/isbn/9781420075755. This dataset is a subset of the 1991 U.S. national Medicare inpatient hospital database for Arizona, and it consists of data from 1,495 randomly selected patients. The first model for in-hospital mortality included length of hospital stay, indicators for age over 80 years, and type of surgery (elective, urgent, or emergency, with elective as baseline):

```
. use medpar.dta, clear
. quietly logit died los age80 type2 type3
. predict pred_y, pr
. generate resid = died - pred_y
. sort pred_y, stable
```

Binned residual plots to assess the overall fit and the inclusion of length of stay were constructed and are displayed in the first row of figure 4. Because many patients have identical estimated mortality probabilities, the `stable` option of the `sort` command should be used.

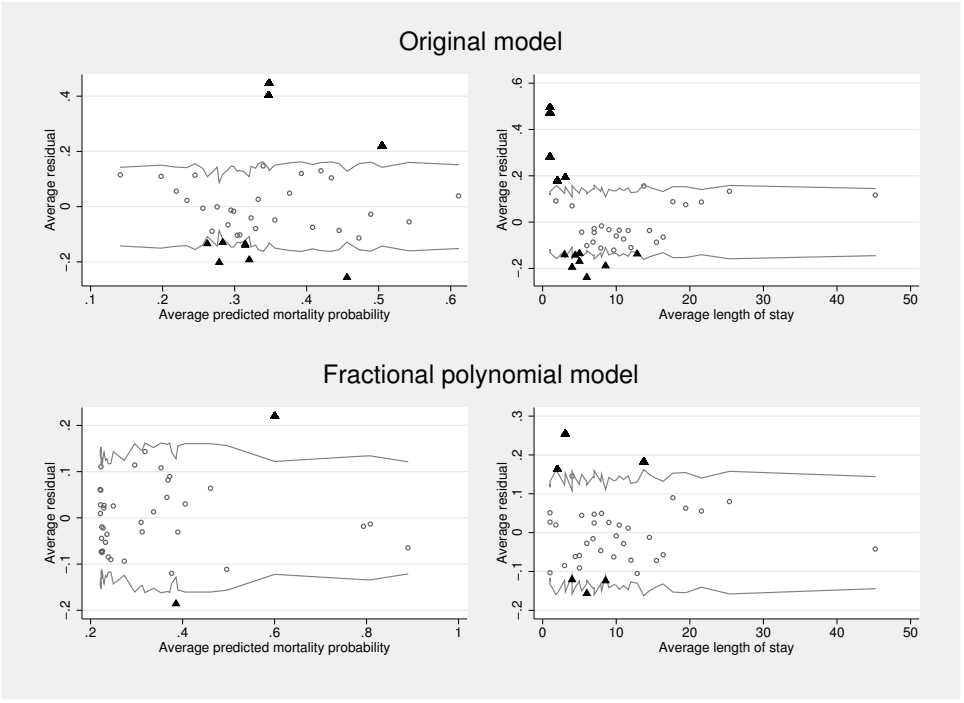

Figure 4. Binned residual plots for the Medpar example

The binned residual plot for length of stay indicates that a linear term for length of stay does not accurately capture the relationship between mortality and length of stay. A multivariable fractional polynomial logistic model is fit as follows:

```
. quietly mfp logit died los age80 type2 type3
```

The selected model contains indicators for age over 80 years and type of surgery, and a degree-1 fractional polynomial for length of stay with power -2. Binned residual plots to assess the overall fit and length of stay are displayed in the second row of figure 4. These plots indicate that this model is a much better fit to the data than that containing a linear term for length of stay.

References

Cox, N. J. 2004. Speaking Stata: Graphing model diagnostics. *Stata Journal* 4: 449–475.

———. 2010. Software update: gr0009_1: Speaking Stata: Graphing model diagnostics. *Stata Journal* 10: 164.

Gelman, A., and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Hilbe, J. M. 2009. *Logistic Regression Models*. Boca Raton, FL: Chapman & Hall/CRC.