

Speaking Stata: Graphing model diagnostics

Nicholas J. Cox
University of Durham, UK
n.j.cox@durham.ac.uk

Abstract. Plotting diagnostic information calculated from residuals and fitted values is a long-standard method for assessing models and seeking ways of improving them. This column focuses on the statistical mainstream defined by regression models for continuous responses, treated in a broad sense to include (for example) generalized linear models. After some comments on the history of such ideas (and even their anthropology and psychology), the commands available in official Stata are reviewed, and a `modeldiag` package is introduced. A detailed example on fuel-wood yield from fallow areas in Nigeria illustrates a variety of general points and specific tips.

Keywords: `gr0009`, `modeldiag`, `anovaplot`, `indexplot`, `ofrtpplot`, `ovfplot`, `qfrplot`, `racplot`, `rdplot`, `regplot`, `rhetplot`, `rvfplot2`, `rvlrplot`, `rvpplot2`, graphics, diagnostics, regression, generalized linear models, analysis of variance

1 Introduction

A common task in statistical graphics is looking at various flavors of residual and predicted (fitted) values after fitting a model. There are now many ideas on how these extra values may be used graphically to examine the fit between data and models and to seek possible means of improving models.

Diagnostic graphs have a key role in adding fine structure to judgments based, all too often, largely on single-valued summaries, such as R^2 (whether plain, adjusted, or pseudo-), AIC, or BIC. Naturally, these graphs should also complement, in a heuristic or exploratory manner, inferences based on specific tests of hypotheses.

Several different kinds of graph may be inspected in many modeling exercises, partly because each kind may be best for particular purposes and partly because in many projects a variety of models—in terms of functional form, choice of predictors, and so forth—may be entertained, at least briefly. It is therefore helpful to be able to produce such graphs very rapidly.

The focus here is on what may be fairly regarded as a central part of statistical modeling: regression treated in a suitably broad sense but emphasizing the modeling of continuous response variables. Thus we will not recapitulate problems or tools specific to particular areas, such as survival or time-series analysis, or material on categorical responses, covered so thoroughly with reference to Stata by Long and Freese (2003).

After a swift survey of the history of these ideas and some comments on variations in current practice, we will review official Stata commands and the `modeldiag` package. A detailed example closes the column.

2 History, anthropology, and psychology

As with many statistical ideas, the approaches discussed in this paper have both very long and much shorter roots. The idea of a residual, as a difference between observed and expected, is some centuries old and is exemplified in the experimental work of Galileo (who also had the idea that error distributions were likely to be symmetric and unimodal). For Galileo's statistical attitudes in particular and other related ideas in the 17th and 18th centuries, see Hald (1986) and Plackett (1988). The general method of inference from residual phenomena (meaning appearances) was strongly emphasized by John Herschel (1792–1871) in *A Preliminary Discourse on the Study of Natural Philosophy* (1830). His book has been described as the first work on philosophy of science in English written by a working scientist; it was widely influential in the 19th century, being studied carefully by Charles Darwin, among many others (Ruse 1979). Apposite quotations from Herschel's book appear as chapter epigraphs in the statistical monograph of Cook and Weisberg (1982).

Despite this splendid past, reports on analysis of residuals and the use of graphs to look at the results of regression-like models both appear to have been unusual in the literature before the early 1960s. (Scatterplots of raw data were naturally more common.) Before modern computers, and also afterwards, the usual algorithms for regression and analysis of variance led to sums of squares, mean squares, and the associated test statistics, rather than the set of individual residuals. In the face of calculation work that could be very time-consuming, just calculating a set of residuals may often have seemed a complication too far, even when data analysts thought about it. In addition, easy production of presentable graphs was available only to rather few researchers until very recently. Even many statistical packages produced ugly lineprinter graphs until the early 1990s. On the other hand, it is difficult to assess how often good data analysts looked at tables or even graphs of residuals informally before it became respectable, and even fashionable, to do so and to talk about it in print.

One striking exception to the general dearth before about 1960 of residual analysis appears in the work of the Danish scientist Thorvald Nicolai Thiele (1838–1910). Thiele worked in astronomy, mathematics, actuarial science, and statistics. He advocated graphical analysis of residuals checking for trends, symmetry of distributions, and changes of sign, and even warned against over-interpreting such graphs (Thiele 1889; Lauritzen 2002, 180–182).

Whatever the detailed prehistory, the modern history of such diagnostic graphics can be said to begin in the early 1960s with the work of Frank Anscombe, John W. Tukey, and others. Major references include Anscombe (1961), Tukey (1962), and Anscombe and Tukey (1963). (Incidentally, Anscombe [1918–2001] and Tukey [1915–2000] married sisters, which led Tukey to refer to Anscombe as his brother-in-squared-law.) Ideas percolated into textbooks, for example, Draper and Smith (1966, 1981, 1998)—their third edition remains a friendly and quite comprehensive survey of regression. The approach was soon defined by its own monographs (Belsley, Kuh, and Welsch 1980; Cook and Weisberg 1982; Atkinson 1985). The field is still active, with many new ideas that cannot be explored here (Cook 1998; Atkinson and Riani 2000).

Nevertheless this forty-year period has evidently been too short to establish any strong uniformity of methodology. Forays into intellectual anthropology or psychology appear necessary to explain some marked variations in practices from field to field. The logic of fitting and assessing regression-like models should transcend disciplinary boundaries, but fields do vary in what is preferred, or even compulsory, showing contrasts in tribal habits. There are fields like my own (geography, environmental sciences) in which a strongly graphical approach is not only welcome but positively expected. There are also fields in which regression-like modeling is central but use of diagnostic graphics appears rare and almost all the emphasis in model assessment is on figures of merit and formal test statistics.

Informal conversations bring up two points repeatedly to explain disinclinations to adopt a strongly graphical approach. First, researchers who may be working with a large number of variables often feel that there would just be too many graphs to work with, especially if many of those graphs could be equivocal or contradictory in their indications. The number of predictors can indeed be limiting, but there are also useful general graphs that are possible regardless of that number. Second, and seemingly more crucial, is that analysis practices tend to be dominated by whatever formats journals prefer or require for publication of results. Frequently, ritual displays of coefficients, standard errors, confidence intervals, and the like are considered essential but requests for graphical displays would be regarded as idiosyncratic or as posing unreasonable requests for journal space.

More detailed discussion of tribal habits in the use of regression would take us too far afield. For an incisive and much broader critique of many issues in contemporary regression methodology, see the polemical monograph of Berk (2004).

3 Existing commands in official Stata

Those with experience in using Stata for both modeling and graphics will often find it easy to get diagnostic graphs with just a few command lines. Thus suppose that you are checking an assumption that error terms follow a normal or Gaussian distribution. The best way to do this graphically is usually with a probability (meaning quantile–quantile) plot with ordered residuals on one axis and the corresponding expected quantiles on the other axis. Arguably that is much more informative than either a general purpose test (chi-square, Kolmogorov–Smirnov) or even a specific purpose test (Shapiro–Wilk, Shapiro–Francia). (The latter two are, in effect, producing numerical summaries of the information in the probability plots.)

In Stata, once a model has been run, this process is at most two commands: a **predict** command to get the residuals and a **qnorm** command to get the graph. Often the residuals will have been calculated already for another purpose, so only one command is needed. In this case, no special command appears needed, but if you were doing this repeatedly, the saving from two lines to one could make it worth your while to put the commands into a short wrapper program. (A related graph showing two quantile plots side by side will be discussed in more detail later.)

Conversely, I often produce observed versus fitted plots, on which more will also be said later. This also requires a `predict` to get the fitted values and a `scatter` to get the graph. At this point, however, I usually want to add a reference line of equality, and I realize that I want better axis titles. The last requires some labeling of the fitted variable or an axis title specification. I found myself doing this so frequently that an `ovfplot` with sensible defaults became a practical proposition.

Official Stata supplies a built-in bundle of commands originally written for use after `regress` and thus *post hoc* in character:

```
avplot and avplots  
  
cprplot and acprplot  
  
lvr2plot  
  
rvfplot and rvppplot
```

These were introduced in Stata 3.0 in 1992 and are documented at [R] **regression diagnostics**. More recently, in an update to Stata 7.0 in 2001, all but the first two were modified so that they may be used after `anova`.

Despite their many uses, this suite omits some very useful kinds of plots, while none of the commands may be used after other modeling commands. To make that point concrete, I find the logic behind generalized linear models very compelling and often want to use `glm`. I also find that physically inspired models typically lead to the brute force approach of `nl`. Evidently, none of the standard commands just mentioned will work after either of these.

Different in spirit, but worth a strong recommendation, are a bundle of commands introduced as part of the new graphics of Stata 8. `twoway lfit`, `twoway qfit`, `twoway ffit`, and their kin implement models on the fly for various functional forms, namely linear, quadratic, and fractional polynomials. They give graphs of data and fitted curves and (if desired) confidence intervals. In this territory, note also commands such as `lowess` and `locpoly` (see Gutierrez, Linhart, and Pitblado 2003).

Clearly, users have a choice. They can explore data using these latter commands and follow up graphs that appear successful with the formalities of, e.g., `regress` or `fracpoly`. I find it particularly helpful whenever an informal (or perhaps semiformal) exploration with `lowess` or `locpoly` either supports the notion of a linear approximation or convinces me that I need something quite different. These explorations rarely survive to the printed page, but they can nevertheless be invaluable aids in model development.

Alternatively, users may find that a simple model developed using such modeling commands can be plotted using one of these `twoway` types. So `twoway lfit` may also be used *post hoc* whenever a `regress` with one predictor appears adequate or at least interesting. It matters not if the next idea is then that curvature or some other nonlinearity or some obvious outliers need more care and attention.

Equally clearly, these **twoway** commands are limited to a few simple and standard forms and in no sense exhaust the repertoire of models that might be useful.

In this column, the main story concerns the use of a new set of commands, which as implied are biased to graphics useful for models predicting continuous response variables. The ideal followed in producing this set is to make minimal assumptions about which modeling command has been issued previously. The downside for users is that if the data and the previous model results do not match the assumptions, it is possible to get either bizarre results or an error message, but these are constitutional hazards in any case. More positively, it is the prerogative, and also the responsibility, of the user to decide what is justifiable. The programs discussed here are available with the Stata Journal software. In addition, they may be downloaded from SSC; see [R] **ssc** for details.

4 The **modeldiag** package

The principles followed in programming such commands include

- as far as possible, the command name by itself should produce a useful plot
- **predict** is used to produce temporary variables for residuals, fitted values, etc.
- each graph refers to the last model fitted
- each graph has reasonably smart default axis titles, etc.
- graphs implement Stata 8 graphics
- options are provided for key needs

The commands which have been written are as follows. First comes a group of general-purpose commands.

4.1 **ovfplot**

ovfplot plots observed versus fitted values for the response from an immediately previous **regress** or similar command, with a line of equality superimposed by default. Some merits of this kind of plot deserve mention. It is easy to understand, especially for presentation to users who do not specialize in statistical applications, and indeed it is often among many scientists' lists of favorites. It is generally applicable, as many kinds of models lead to predictions directly comparable with the response variable, whatever the number of predictors or the functional form. Residuals, measured on the same scale as the response, can be read off the plot as vertical differences. Another merit, which is also a profound defect, is that the observed versus fitted plot can be an optimistic or propaganda plot (Tukey 1972; Cox 2004c), making even a lousy model look fairly good.

4.2 `regplot`

`regplot` plots fitted or predicted values from an immediately previous `regress` or similar command. By default, the data for the response are also plotted.

With one syntax, no variable name is specified: `regplot` then shows the response and predicted values on the y -axis and the predictor named first in the `regress` or similar command on the x -axis. Thus with this syntax the plot shown is sensitive to the order in which predictors are specified in the estimation command.

With another syntax, a variable name is supplied, which may name any numeric variable: this is then used as the variable on the x -axis.

Thus in practice, `regplot` is most useful when the fitted values are a smooth function of the variable shown on the x -axis, or a set of such functions given also one or more dummy variables as predictors. However, other applications also arise, such as plotting observed and predicted values from a time-series model versus time.

By default, `regplot` shows the fitted values using `twoway mspline`. The `plotttype()` option may be used to specify another `twoway` `plotttype`.

A `separate()` option specifies that values of fitted and observed responses be plotted as separate groups corresponding to the distinct values of the variable specified. This is especially useful when a categorical predictor has been included in the model as one or more dummy variables. The `by()` option remains available as usual.

Note that `regplot` does not work after `anova`; see the comments on `anovaplot`, discussed later.

4.3 `rvfplot2`

`rvfplot2` plots residuals versus fitted values from an immediately previous `regress` or similar command. This is one of the main workhorses in this area. The underlying idea is that no news is good news: ideally, the scatter should be fairly even and patternless, with no hints of (for example) curvature, uneven scatter, or disturbance by outliers.

The residuals are, by default, those calculated by `predict, residuals` or (if the previous estimation command was `glm`) by `predict, response`. The fitted values are those produced by `predict` by default after each estimation command. `rvfplot2` is offered as a generalization of `rvfplot` in official Stata.

There is support for specifying several types of residual other than the default. An `rscale()` option specifies a transformed scale on which to show the residuals using Stata syntax and `X` as a placeholder for the residual variable name. Thus `rscale(X^2)` specifies squaring, to show relative contribution to residual variance; `rscale(abs(X))` specifies absolute value, to set aside sign; `rscale(sqrt(abs(X)))` specifies root of absolute value, a useful scale on which to check for heteroskedasticity.

Similarly, an `fscale()` specifies a transformed scale on which to show the fitted values using Stata syntax and `X` as a placeholder for the fitted variable name. Thus, for example, `fscale(2 * ln(X))` specifies twice the natural logarithm, which is the constant information scale for a generalized linear model with gamma error. Similarly, arguments of `2 * sqrt(X)`, `2 * asin(sqrt(X))`, and `-2 / sqrt(X)` specify the constant information scale for Poisson, binomial, and inverse Gaussian errors, respectively. See McCullagh and Nelder (1989, 398) for background.

A `lowess` option specifies that the residuals will be smoothed as a function of the fitted using `lowess` (options of which may be specified in turn).

4.4 `rvpplot2`

`rvpplot2` plots residuals versus values of a specified predictor (a.k.a., independent variable or carrier) from an immediately previous `regress` or similar command. The residuals are, by default, those calculated by `predict`, `residuals` or (if the previous estimation command was `glm`) by `predict`, `response`.

`rvpplot2` is offered as a generalization of `rvpplot` in official Stata.

There is support for specifying several types of residuals other than the default. A `force` option allows you to specify a predictor variable not included in the previous model. An `rscale()` option specifies a transformed scale on which to show the residuals using Stata syntax and `X` as a placeholder for the residual variable name. Thus `rscale(X^2)` specifies squaring, to show relative contribution to residual variance; `rscale(abs(X))` specifies absolute value, to set aside sign; `rscale(sqrt(abs(X)))` specifies root of absolute value, a useful scale on which to check for heteroskedasticity. A `lowess` option specifies that the residuals will be smoothed as a function of the predictor using `lowess` (options of which may be specified in turn).

Thus `rvpplot2` offers scope for easy plotting against either a predictor already in the model or a possible predictor not at present included in the model. The latter could be time or spatial position if there was concern about serial autocorrelation. A clear pattern in this plot will point up the possibility of modifying the model. With a predictor in the model, some evidence of curvature or other nonlinearity might point to a change in how it was included, for example, the addition of a quadratic term or a prior transformation. With a predictor not in the model, evidence of correlation might suggest adding that predictor to the model.

4.5 `indexplot`

`indexplot` plots estimation results (by default whatever `predict` produces by default) from an immediately previous `regress` or similar command versus observation number (i.e., `_n`).

Values are shown, by default, as vertical spikes starting at 0 using `twoway dropline`, but the graph may be recast to another `twoway` plot type.

4.6 qfrplot

qfrplot plots quantiles of fitted values, minus their mean, and quantiles of residuals from the previous **regress** or similar command.

Fitted values are whatever **predict** produces by default, and residuals are whatever **predict, res** produces. Comparing the distributions gives an overview of their variability and some idea of their fine structure, as plots appear side by side with aligned vertical scales. Note that the rationale of this graph is comparing distributions of (fitted – mean) and residuals; hence, it is vital that residuals be measured on the same scale as the response.

Options include observed versus normal (Gaussian) quantile–quantile plots. Note that **qfrplot** is essentially a wrapper for calls to **qplot** (Cox 2004a,b), itself a generalization, apart from one small detail, of official Stata’s **quantile** command.

Cleveland (1993) gives many side-by-side quantile plots of fit and residuals, which he calls “residual-fit spread plots”. See, for example, the graph on his page 41. However, he also uses this term for side-by-side time-series plots of fit and residuals (page 157). The command name and description here emphasize the use of a quantile plot.

4.7 rdplot

rdplot plots residual distributions from the previous **regress** or similar command. The residuals are, by default, those calculated by **predict, residuals** or (if the previous estimation command was **glm**) by **predict, response**.

The graph by default is a single or multiple dotplot, as produced by **dotplot**. Histograms as produced by **histogram** or box plots as produced by **graph box** or **graph hbox** may be selected. Oneway plots as implemented in **onewayplot**, skewness plots as implemented in **skewplot**, or quantile plots as implemented in **qplot** may also be selected. On the last three, see Cox (2004a,b).

Various options offer scope for grouping residuals in various ways, according to values of some other variable (e.g., a predictor).

4.8 rhetplot

rhetplot checks for residual heteroskedasticity after the previous **regress** or similar command.

rhetplot graphs standard deviations (optionally variances) of residuals for distinct groups formed by combinations of specified variables; standard deviations (optionally variances) of residuals against means of groups of a specified variable; or standard deviations (optionally variances) of residuals against means of groups of fitted values.

The residuals are, by default, those calculated by `predict`, `residuals` or (if the previous estimation command was `glm`) by `predict`, `response`. There is support for specifying several types of residual other than the default.

The graph is produced by `lowess`. The “smooth” curve shown (unless the number of groups specified is very small) is best regarded as an informal indication of the general pattern of variability of residuals.

Next come a group of commands designed for models based on time, although they may easily be extended to other situations when appropriate. Even commands that assume a previous `tsset` can be applied after `sorting` to a sensible order, and

```
. gen t = _n  
. tsset t
```

Conversely, the safeguard of requiring `tsset` gives users some protection against getting incorrect results, in particular with panel data.

4.9 ofrtplot

`ofrtplot` plots observed, fitted, and residuals versus “time” variables after the previous `regress` or similar command. It is primarily designed for time-series models, and by default the predictor is whatever has been `tsset` as the time variable. However, other variables may be specified, whether or not data have been `tsset`.

Observed values are for the response or dependent variable from the last model, fitted values are whatever `predict` produces by default, and residuals are whatever `predict`, `res` produces.

By default, the plot has two panels. In the top panel, observed and fitted are plotted against the predictor. In the bottom panel, residuals are plotted against the predictor, by default as spikes from zero. Optionally, plots may be superimposed, not separate.

4.10 rvlrplot

`rvlrplot` plots residuals versus lagged (i.e., lag 1) residuals for time-series data after the previous `regress` or similar command. Data must have been `tsset` previously.

By default, residuals are whatever `predict`, `res` produces after a model. There is support for specifying several types of residuals other than the default.

4.11 racplot

`racplot` plots the residual autocorrelation function after the previous `regress` or similar command. `racplot` calculates the residuals and then fires up `ac`. Data must have been `tsset` previously. There is support for specifying several types of residuals other than the default. An `rscale()` option specifies a transformed scale for the residuals using Stata syntax and `X` as a placeholder for the residual variable name.

Finally, in this listing, is a command especially dedicated to the results of `anova`:

4.12 `anovaplot`

`anovaplot` plots fitted or predicted values from an immediately previous one-, two-, or three-way `anova`. By default, the data for the response are also plotted. In particular, `anovaplot` can show interaction plots. The format of the graph may be varied by permuting the names of predictors used in `anova`.

Note especially that the graph format produced by `anovaplot` is appropriate for models with at most one continuous predictor, which should always be the predictor named first. With that caveat, `anovaplot` offers a way of showing parallel and diverging regression lines for models with one continuous predictor.

It is curious that analysis-of-variance people typically draw interaction plots but suppress the data, whereas regression people prefer to draw scatterplots showing both observed and fitted values. Admittedly, a complicated set of crossing lines showing interactions may seem to leave little scope for showing data effectively, while a relatively simple regression leaves plenty of scope, but the difference is nevertheless intriguing.

5 Example: wood volumes and fallow length in Nigeria

In many areas of the humid tropics, fallow areas are used for fuelwood and are indeed vitally important for local energy supply. With increasing population pressure and intensification of cropping, such fallows are under threat. The growth of trees on fallows is thus of great interest. The data for this example come from a paper by Adesina (1990), who looked at 80 fallows near Gbongan township in western Nigeria, asking: do fallows planted with the fast-growing species *Gliricidia sepium* yield more wood than “natural” or self-propagated fallows?

Data were collected for quadrats of 20 m × 20 m on slopes no greater than 2°, 40 on each of two fallow types. Of several variables measured, we will look at wood volume, in cubic meters, as a response, and length of fallow in years and fallow type, coded by 0 for natural and 1 for *Gliricidia*, as predictors.

The main purpose of this section is to provide some simple illustrations of `modeldiag` in action, without purporting to give an analysis fully sensitive to all the scientific or practical nuances of the problem. Adesina’s paper gives a most interesting description of the context but an analysis that includes no graphs and is based on bivariate correlations and *t* tests comparing means. Here we seize the opportunity to use length of fallow fully as a quantitative predictor within various models.

```
. scatter volume years, by(type) ms(oh)
```

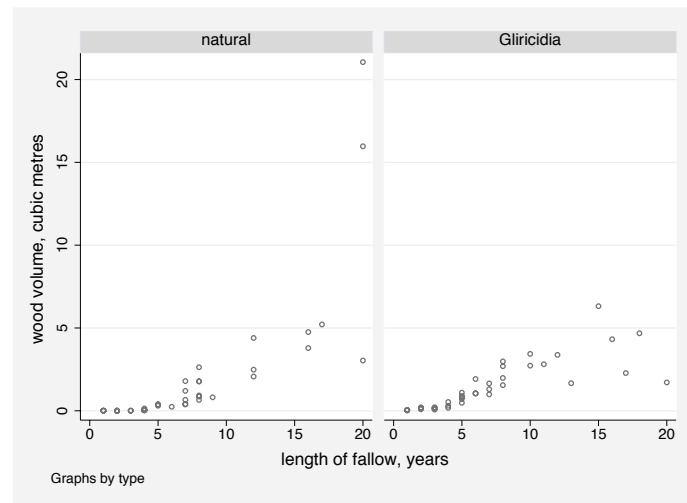


Figure 1: Scatterplot. The relationship between wood volume, fallow length, and vegetation type is obscured by the skewness of both volume and length.

This exploratory scatterplot (figure 1) and some simple summary statistics immediately reveal various basic features. Both **volume** and **years** are positively skewed (the moment measure of skewness is 4.372 in the case of **volume**), and relatively few plots have been fallow for more than (say) 10 years (17/80). It is thus difficult to discern structure, especially on the right-hand side of the plot. At worst, some outliers may be present. It will be easier to see what is going on if we transform the response. Cube root, sometimes a rather arbitrary transform, seems very natural here, as the units then become meters, and we are dealing with what may be called an equivalent length. Imagine a bundle of wood with given volume and of cubical shape; its sides will have this length. (In passing, note an excellent article on dimensional analysis and statistics by Finney [1977].) The cube root of volume is still skewed (0.691), but we are moving in the right direction and not making anything worse.

(Continued on next page)

```
. gen curtvol = volume^(1/3)
. label var curtvol "equivalent length, m"
. regress curtvol years
```

Source	SS	df	MS			
Model	18.4151275	1	18.4151275	Number of obs = 80		
Residual	5.33350034	78	.068378209	F(1, 78) = 269.31		
				Prob > F = 0.0000		
				R-squared = 0.7754		
				Adj R-squared = 0.7725		
Total	23.7486278	79	.300615542	Root MSE = .26149		
curtvol	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
years	.0919356	.0056022	16.41	0.000	.0807826	.1030886
_cons	.2591998	.0494769	5.24	0.000	.1606989	.3577007

More importantly, a trial regression looks good numerically, with clear-cut F and t results, R^2 of 0.775, and root mean squared error of 0.261 m.

However, a basic `regplot` shows that the *Gliricidia* data points possess considerable curvature, so that the model is missing some structure (figure 2). This is also shown by the residual versus fitted plot (figure 3).

```
. regplot, by(type)
```

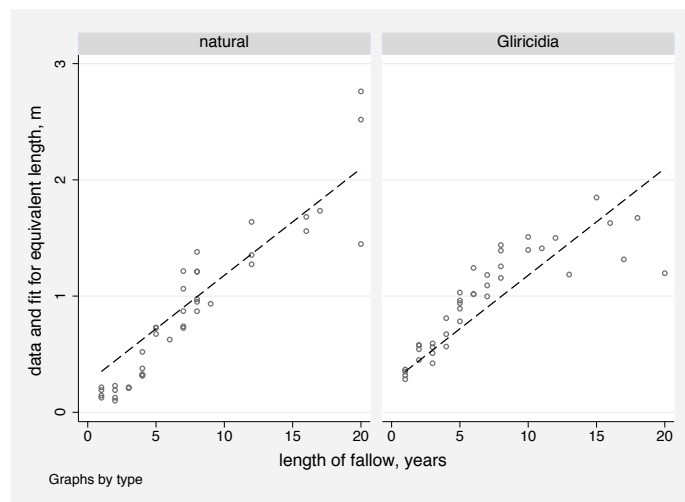


Figure 2: Regression plot. The regression of equivalent length on fallow length still leaves important curvature, especially for *Gliricidia*.

```
. rvfplot2, ms(oh)
```

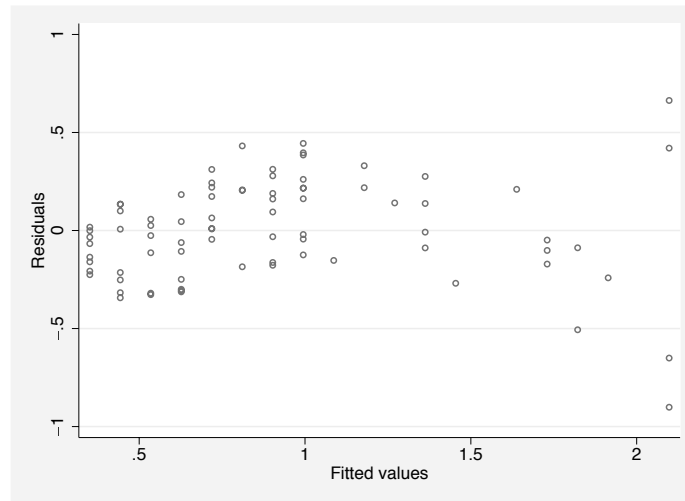


Figure 3: Residual versus fitted plot. Curvature is evident here too.

We add **type** as a dummy variable and the product of **type** and **years** as another predictor to permit differing slopes and intercepts. This boosts R^2 to 0.830 and reduces root mean squared error to 0.230 m. Unsurprisingly, much curvature remains (figures 4, 5, and 6). Note that in this case—with just one predictor whose coefficient is positive—the residual versus fitted plot and the residual versus predictor plot are the same graph, modulo the labeling of the x -axis. Nevertheless one may be more convenient to read than the other, depending on whether the scientist finds it easier to think on the predictor or the response scale.

```
. gen type_years = type * years
. regress curtvoll years type type_years
```

Source	SS	df	MS	Number of obs = 80		
Model	19.7122058	3	6.57073528	F(3, 76) = 123.72		
Residual	4.03642201	76	.053110816	Prob > F = 0.0000		
Total	23.7486278	79	.300615542	R-squared = 0.8300		
				Adj R-squared = 0.8233		
				Root MSE = .23046		

curtvoll	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
years	.1104499	.0067043	16.47	0.000	.0970971	.1238028
type	.4315831	.0873577	4.94	0.000	.257595	.6055711
type_years	-.0388651	.0099265	-3.92	0.000	-.0586355	-.0190947
_cons	.0446065	.0615559	0.72	0.471	-.0779928	.1672058

```
. regplot, by(type)
```

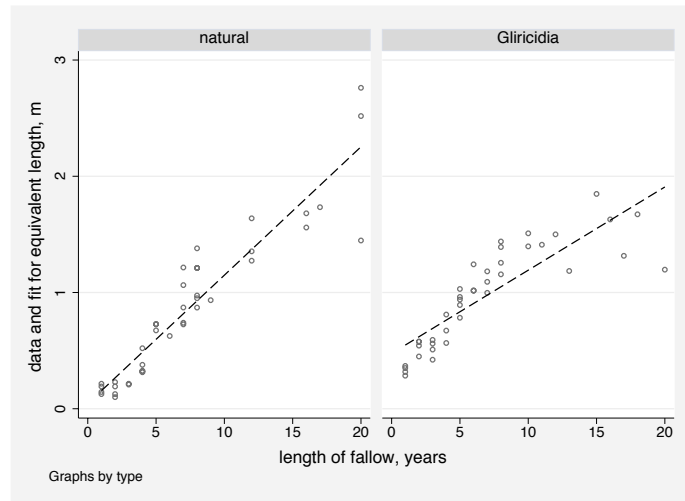


Figure 4: Regression plot. Allowing interaction is one thing, but the curvature for *Gliricidia* is another.

```
. rvfplot2, by(type) ms(oh)
```

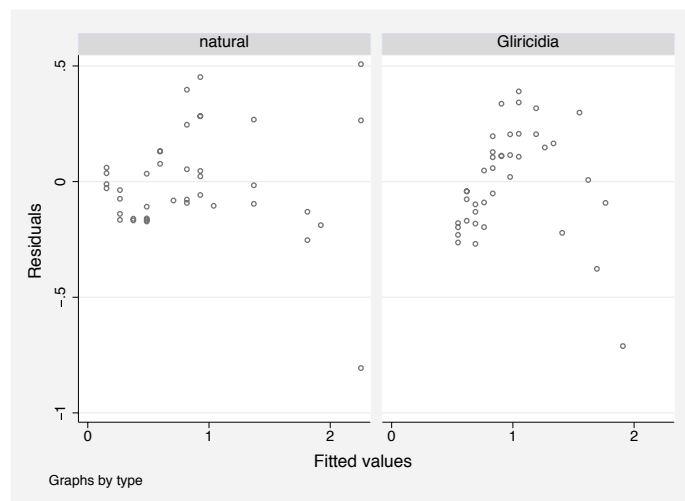


Figure 5: Residual versus fitted plot. Another way of seeing the curvature.

```
. rvpplot2 years, by(type) ms(oh)
```

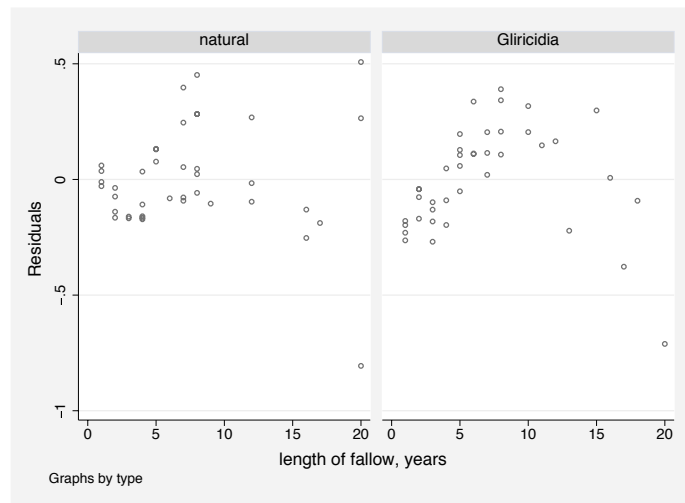


Figure 6: Residual versus predictor plot. Yet another way of seeing the curvature, in this case arguably clearer than the residual versus fitted plot.

A multiple dot plot shows that the residuals are heteroskedastic (figure 7). In using `rdplot`, there is a trade-off: enough groups are needed to get an idea of any fine structure, but not so many that there are too few data points in each group to summarize effectively. I often start with `group(3)` and go to more groups only if it seems sensible, but with a much larger dataset than that here, a larger number of groups would be justifiable.

(Continued on next page)

```
. rdplot, group(3) ms(oh)
```

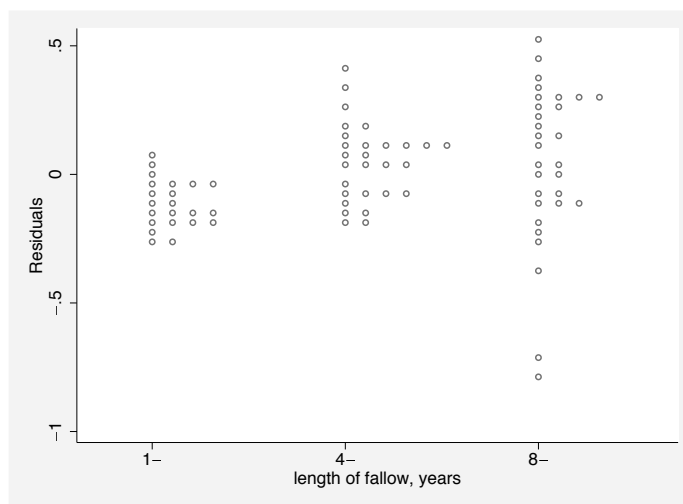


Figure 7: Residual distribution plot. Data points sliced into three groups according to values of the predictor with approximately the same size. Heteroskedasticity is evident.

So far then, although the **regress** output shows us a well-behaved dataset, the regression model is still failing to capture important structure. In addition, although it happens that the intercept is just above zero, there is no guarantee with a model of this kind that predictions are all positive, which is essential biologically. A positive prediction can be ensured by using a generalized linear model with logarithmic link. Using this, rather than a logarithmic transformation, has the signal advantage that results are returned on an intelligible scale, without any need for back-transformation. We have also some flexibility over choice of error distribution.

However, a logarithmic link with **years** as predictor would imply exponential growth over time, seemingly not appropriate for either type of vegetation. A power function appears more sensible than an exponential, which leads us to try log of years as a predictor.

(Continued on next page)


```

. gen logyears = log(years)
. gen type_logyears = logyears * type
. glm curtvoll logyears type type_logyears, link(log) nolog

```

Generalized linear models		No. of obs	=	80
Optimization	: ML: Newton-Raphson	Residual df	=	76
		Scale parameter	=	.0428134
Deviance	= 3.253821965	(1/df) Deviance	=	.0428134
Pearson	= 3.253821965	(1/df) Pearson	=	.0428134
Variance function:	V(u) = 1	[Gaussian]		
Link function	: g(u) = ln(u)	[Log]		
Standard errors	: OIM			
Log likelihood	= 14.57277095	AIC	=	-.2643193
BIC	= -329.7802023			

curtvoll	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
logyears	.9140198	.0595632	15.35	0.000	.7972781 1.030762
type	1.046955	.1861204	5.63	0.000	.6821658 1.411744
type_logyears	-.4247248	.0752181	-5.65	0.000	-.5721496 -.2773
_cons	-1.947677	.1541104	-12.64	0.000	-2.249727 -1.645626

With appropriate extra predictors to allow for an interaction, the generalized linear model looks good numerically. The output for `glm` does not supply an R^2 or a root mean squared error, but a simple program not documented here summons up both these measures from the correlation and differences between response and fitted. For detailed statistical arguments on why doing this is perfectly sensible, see Zheng and Agresti (2000). The values of 0.864 and 0.207 suggest some progress.

In looking at the model and data overall we can specify plotting against `years`, even though `logyears` was the predictor in the model (figure 8). Now the residuals look better (figures 9 and 10).

(Continued on next page)

```
. regplot years, by(type)
```

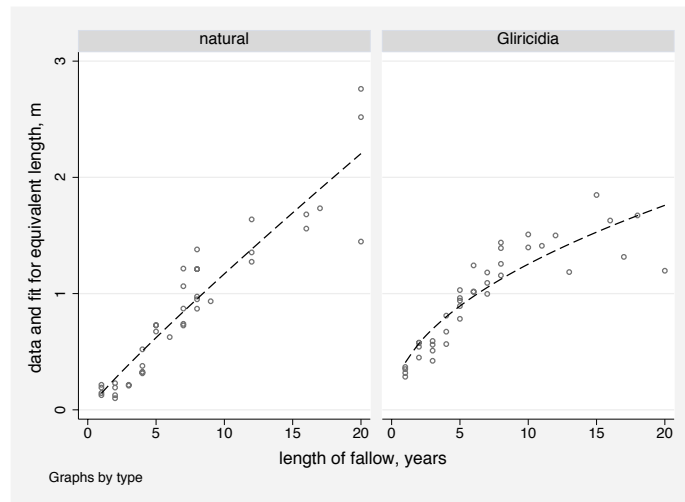


Figure 8: Regression plot. Generalized linear model fitted with logarithmic link and log years as one predictor, but plotted here versus years. This does a better job of capturing the different behavior.

```
. rvfplot2, by(type) ms(oh)
```

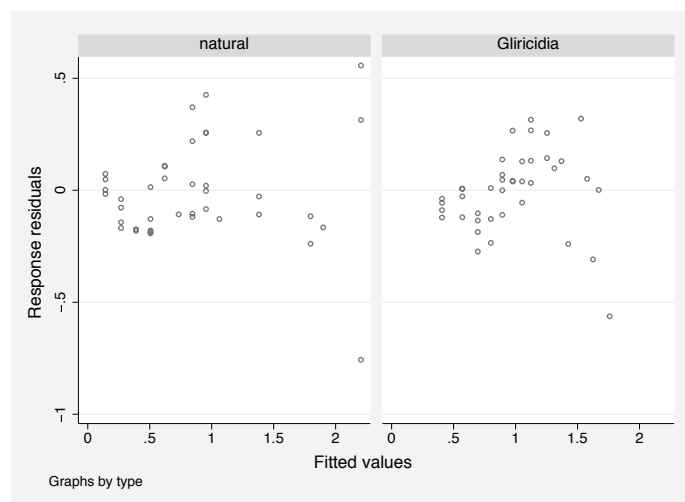


Figure 9: Residual versus fitted plot. The residuals are better behaved. Does important curvature persist?

```
. rvpplot2 years, by(type) ms(oh) force
```

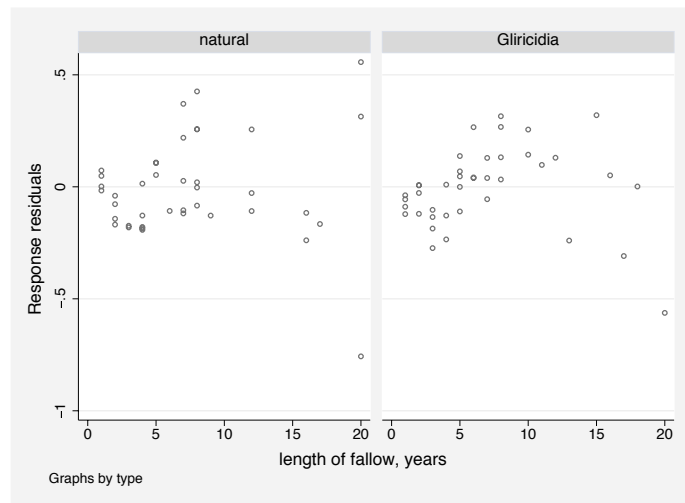


Figure 10: Residual versus predictor plot. Years is not in the model but can be used as an axis with the **force** option.

The quantile plot of fitted and residuals popularized by Cleveland (1993) is a nice summary of “how far we have come” compared with “how far we have yet to go” (figure 11). By default, we chose a normal (Gaussian) error family for the generalized linear model, so looking at residuals on a Gaussian scale is pertinent (figure 12). Nevertheless this plot avoids a key issue, whether error distributions are homoskedastic, which does not appear to be the case (figure 13). This leads to a switch to a gamma error family.

(Continued on next page)

```
. qfrplot
```

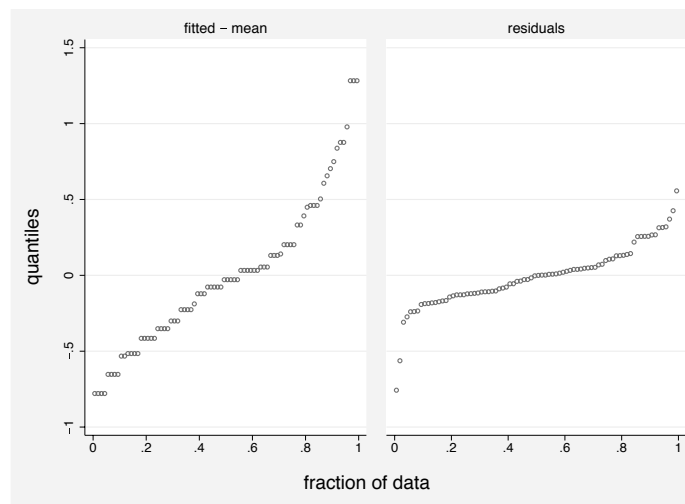


Figure 11: Quantile plot of fitted and residuals. Fitted – mean and residuals have the same scale, so they can be juxtaposed.

```
. qfrplot, gauss
```

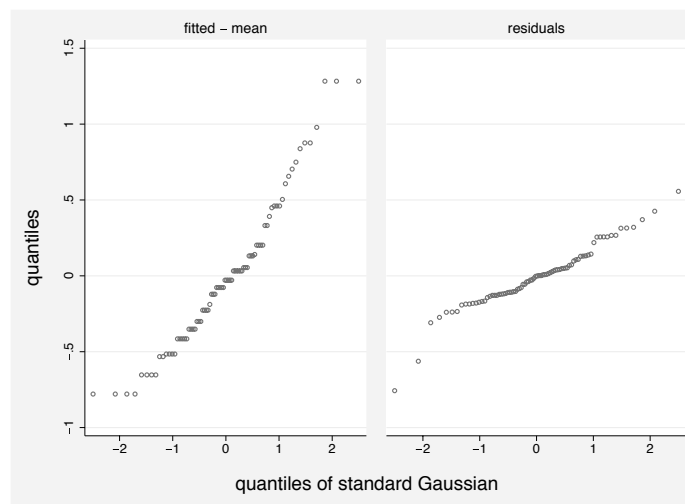


Figure 12: Quantile plot of fitted and residuals. A Gaussian scale is used, as that is the error family postulated. The assumption looks fair for the data as a whole.

```
. rdplot, group(3) ms(oh)
```

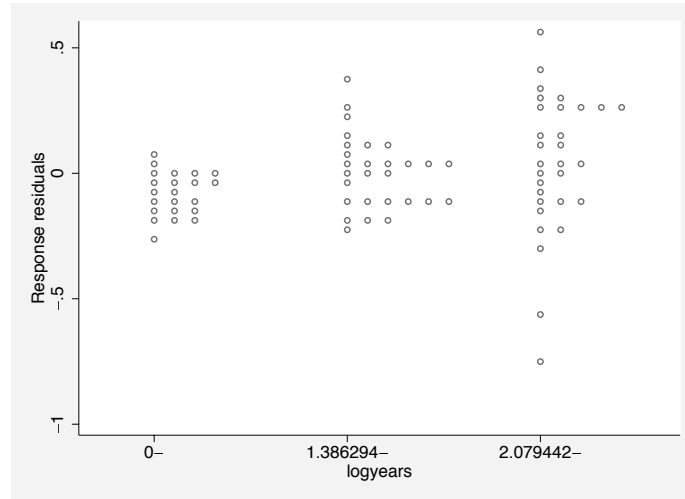


Figure 13: Residual distribution plot. The heteroskedasticity is still evident.

In one sense, the results for the fit look very similar, and every P -value in sight is excellent. R^2 drops a smidgen to 0.856 and root mean squared error rises correspondingly to 0.214 m, but experience teaches us not to be oversensitive to such small differences. The model gives a near-linear power function (power 0.964) for natural fallows and one much closer to the square root for *Gliricidia* ($0.964 - 0.383 = 0.581$) (figure 14). The residuals give no great cause for concern (figures 15 and 16); a sharp eye would wonder if the curvature of *Gliricidia* had been followed quite correctly, but there is an issue of how much weight to put on values for longer fallow lengths.

```
. glm curtvoll logyears type type_logyears, link(log) f(gamma) nolog
Generalized linear models          No. of obs      =       80
Optimization      : ML: Newton-Raphson      Residual df    =       76
                                                Scale parameter =   .0572479
Deviance          =   4.702906734      (1/df) Deviance =   .0618804
Pearson           =   4.350838306      (1/df) Pearson  =   .0572479
Variance function: V(u) = u^2          [Gamma]
Link function     : g(u) = ln(u)        [Log]
Standard errors   : OIM
Log likelihood    =  -56.88833941      AIC              =   1.522208
BIC               =  -328.3311175
```

curtvoll	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
logyears	.9642254	.0430776	22.38	0.000	.8797948	1.048656
type	.9961841	.1186355	8.40	0.000	.7636627	1.228705
type_logyears	-.3831086	.0638372	-6.00	0.000	-.5082273	-.2579899
_cons	-2.076803	.0821826	-25.27	0.000	-2.237878	-1.915728

```
. regplot years, by(type)
```

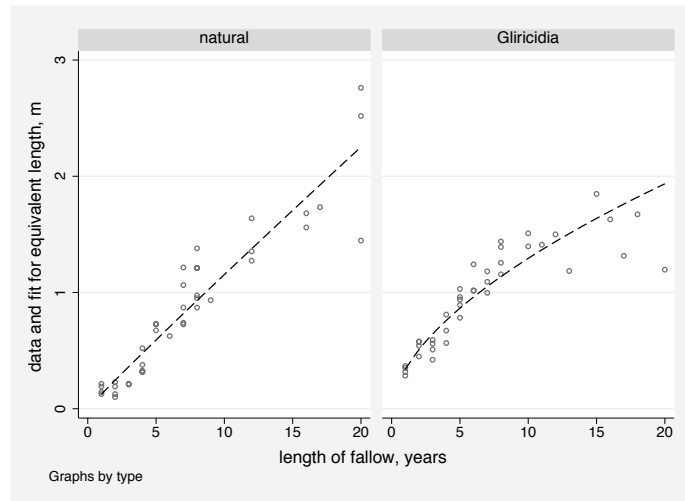


Figure 14: Regression plot. Fitted curves with gamma error assumption are similar to those with Gaussian error assumption.

```
. qfrplot
```

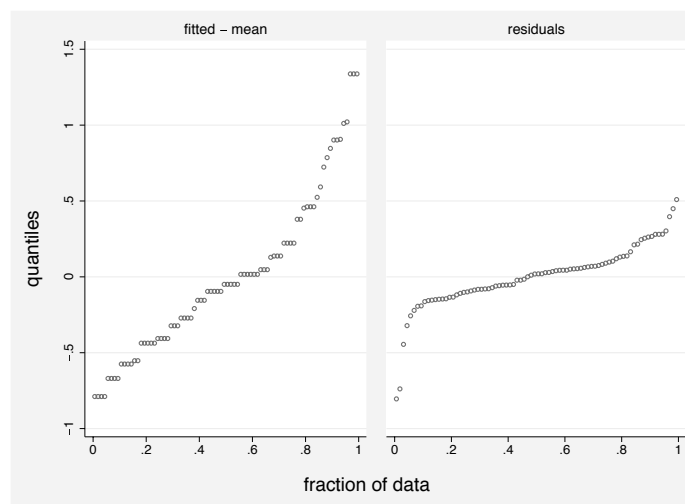


Figure 15: Quantile plot of fitted and residuals. A graphical alternative to an R^2 result.

```
. rdplot, group(3) ms(oh)
```

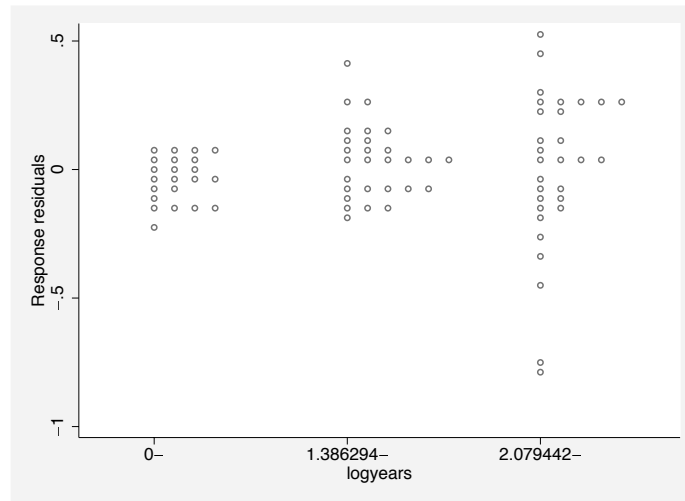


Figure 16: Residual plot. Heteroskedasticity is now expected given the gamma error assumption.

Offstage we saved the predictions from `glm` with Gaussian and gamma errors. After some prior surgery with `separate`, here is a line plot (figure 17). Both the similarity and the differences make sense. The gamma-based model is not constrained by an ideal of homoskedastic errors and is thus less sensitive to some rather low response values for longer fallow lengths, which seem rather suspicious. Until more ideas or more data arrive, the gamma-based model appears to have the edge.

(Continued on next page)

```
. line p?a*0 p?a*1 years, sort yti("equivalent length, m")
```

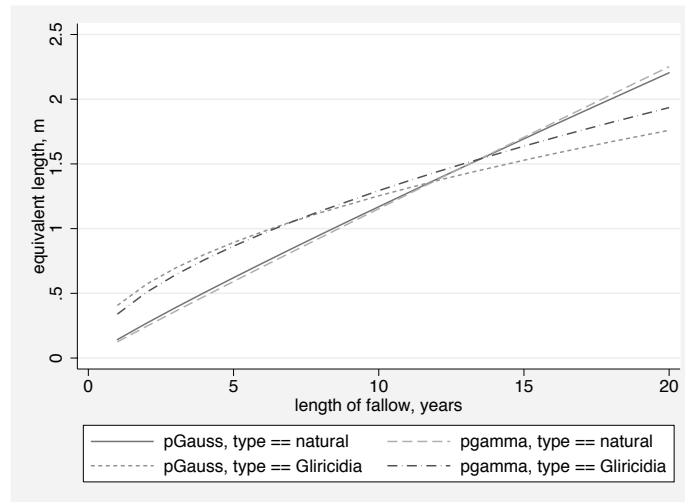


Figure 17: Line plot. The predictions for Gaussian and gamma error models are compared.

An observed versus fitted plot usually looks pretty good and provides an optimistic close (figure 18).

```
. ovfplot
```

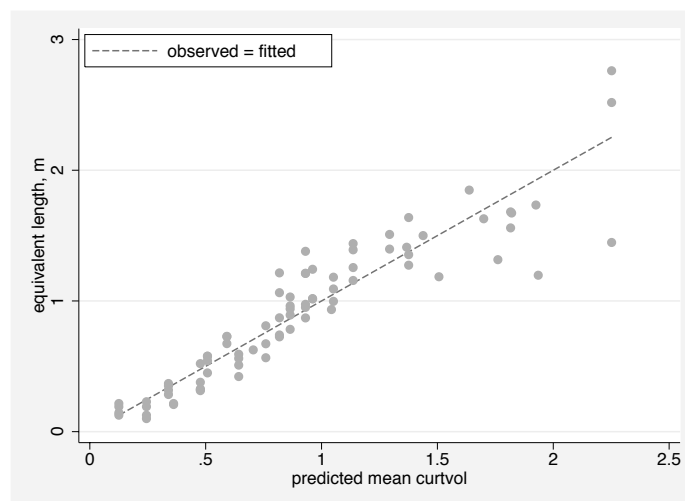


Figure 18: Observed versus fitted plot. A simple summary possible for many models.

As every modeler knows, fitting the data (or the apparent structure in the data) is only part of the battle. The ultracynical could comment that this example is, at root, showing that trees grow over time. Underlying the data, although at some removes from what we have, are presumably some monotonic growth curves. However, we have not data for individual trees monitored over time, but collective results for plots with varying number of trees from a spatial survey (cross-section, not panel data, in other words). No doubt we would benefit from extra predictors that might explain variations between plots, giving detail, say, on soils, microclimate, moisture, nutrients, and the precise history and pattern of land use at each site. However, these further predictors are not available and would have required an enormously bigger project. Adesina does give information on tree and herb diversity, which might serve as surrogates for competition from other species, but results not shown here indicate that they do not help substantially in improving the model, and in any case one should be wary of over-fitting. It is known that some wood is lost from fallows by casual harvesting before they are cleared for subsequent cultivation. Some of the data points for long fallows do look suspiciously low. One radical way to tackle this would be to repeat the analysis with only the shorter fallows, particularly as the comparison between natural and *Gliricidia* in early years is the heart of the matter.

6 Conclusions

The calculation of residuals—the bits left over, the parts of the data the model failed to reach—is the end of one process but also the beginning of a new one whenever we can see something that the model does not capture. Seeing that something is helped by having pictures to look at. We have shown that several Stata commands, official and user-written, exist to help.

What is exciting for Stata users interested in this approach is that many methods remain to be implemented in Stata, such as the ideas of Cook (1998) and Atkinson and Riani (2000). In addition, an open question is how far (and how) `avplot`, `avplots`, `cprplot`, `acprplot`, and `lvr2plot` may be generalized to be used with other commands. In this and other ways, the area of graphical diagnostics may be expected to develop further.

7 Acknowledgments

Kit Baum, Denis de Crombrughe, Phil Ender, Ken Higbee, and Andy Sloggett commented on earlier versions of programs discussed here.

8 References

- Adesina, F. A. 1990. Planted fallows for sustained fuelwood supply in the humid tropics. *Transactions, Institute of British Geographers* 15: 323–330.

- Anscombe, F. J. 1961. Examination of residuals. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1: 1–36.
- Anscombe, F. J. and J. W. Tukey. 1963. The examination and analysis of residuals. *Technometrics* 5: 141–160.
- Atkinson, A. C. 1985. *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*. Oxford: Oxford University Press.
- Atkinson, A. C. and M. Riani. 2000. *Robust Diagnostic Regression Analysis*. New York: Springer.
- Belsley, D. A., E. Kuh, and R. E. Welsch. 1980. *Regression Diagnostics*. New York: Wiley.
- Berk, R. A. 2004. *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage.
- Cleveland, W. S. 1993. *Visualizing Data*. Summit, NJ: Hobart Press.
- Cook, R. D. 1998. *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.
- Cook, R. D. and S. Weisberg. 1982. *Residuals and Influence in Regression*. New York: Chapman and Hall.
- Cox, N. J. 2004a. Software update: gr42.2: Quantile plots, generalized. *Stata Journal* 4(1): 97.
- . 2004b. Speaking Stata: Graphing distributions. *Stata Journal* 4(1): 66–88.
- . 2004c. Speaking Stata: Graphing agreement and disagreement. *Stata Journal* 4(3): 329–349.
- Draper, N. R. and H. Smith. 1966. *Applied Regression Analysis*. New York: Wiley.
- . 1981. *Applied Regression Analysis*. 2nd ed. New York: Wiley.
- . 1998. *Applied Regression Analysis*. 3rd ed. New York: Wiley.
- Finney, D. J. 1977. Dimensions of statistics. *Applied Statistics* 26: 285–289.
- Gutierrez, R. G., J. M. Linhart, and J. S. Pitblado. 2003. From the help desk: Local polynomial regression and Stata plugins. *Stata Journal* 3(4): 412–419.
- Hald, A. 1986. Galileo’s statistical analysis of astronomical observations. *International Statistical Review* 54: 211–220.
- Herschel, J. F. W. 1830. *A Preliminary Discourse on the Study of Natural Philosophy*. London: Longman, Rees, Orme, Brown, and Green; John Taylor. Facsimile reprint: Chicago: University of Chicago Press, 1987.

- Lauritzen, S. L. 2002. *Thiele: Pioneer in Statistics*. Oxford: Oxford University Press.
- Long, J. S. and J. Freese. 2003. *Regression Models for Categorical Dependent Variables Using Stata*. rev. ed. College Station, TX: Stata Press.
- McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.
- Plackett, R. L. 1988. Data analysis before 1750. *International Statistical Review* 56: 181–195.
- Ruse, M. 1979. *The Darwinian Revolution: Science Red in Tooth and Claw*. Chicago: University of Chicago Press.
- Thiele, T. N. 1889. *Almindelig Iagttagelseslære: Sandsynlighedsregning og mindste Kvadraters Methode*. Kjøbenhavn: C. A. Reitzel. English translation included in Lauritzen 2002.
- Tukey, J. W. 1962. The future of data analysis. *Annals of Mathematical Statistics* 33: 1–67 and 812.
- . 1972. Some graphic and semigraphic displays. In *Statistical Papers in Honor of George W. Snedecor*, ed. T. A. Bancroft and S. A. Brown, 293–316. Ames, IA: Iowa State University Press.
- Zheng, B. and A. Agresti. 2000. Summarizing the predictive power of a generalized linear model. *Statistics in Medicine* 19: 1771–1781.

About the Author

Nicholas Cox is a statistically minded geographer at the University of Durham. He contributes talks, postings, FAQs, and programs to the Stata user community. He has also co-authored fifteen commands in official Stata. He was an author of several inserts in the *Stata Technical Bulletin* and is Executive Editor of the *Stata Journal*.