The Stata Journal (2004)
4, Number 3, pp. 242–256

Graphing confidence ellipses: An update of ellip for Stata 8

Anders Alexandersson
Mississippi State University

Abstract. This paper describes an update of the `ellip` command for graphing confidence ellipses in Stata 8. Two of the most notable new features are the option to graph confidence ellipses around variable means and the ability to add inscribed lines. These features allow a geometric characterization of linear regression with unequal error variances, as in McCartin (2003).

Keywords: gr32_1, ellip, confidence ellipse, error-variance regression, elliptical distribution

1 Introduction

This paper describes an update of the `ellip` command (Alexandersson 1998) for graphing confidence ellipses in Stata 8. Most notably, the new features allow a geometric characterization of linear regression with unequal error variances, as in McCartin (2003).

The concept of an ellipse is over 2,000 years old.[1] Since Descartes (1596–1650), the "standard" or "normal" ellipse usually refers to an ellipse with a boundary constant of 1, centered at the origin and rotated along the Cartesian axes (nonrotated). This paper views the ellipse as a set of points, a point-conic, which may degenerate to a point.

The confidence ellipse has a much shorter history than the ellipse and is a more contested concept. The confidence ellipse is defined broadly as an ellipse-shaped joint $100(1-\alpha)\%$ confidence region for two parameters. The larger the confidence level is, the larger the confidence ellipse will be. The paper will discuss three related topics that should help to clarify the use and misuse of confidence ellipses and of the updated program: the geometry of linear regression, the boundary constant of confidence ellipses, and elliptical distributions.

2 The geometry of linear regression

It was Bravais who discovered the confidence ellipse mathematically. Bravais (1846, 273) graphed the regression line together with the "ellipse fondamentale" and its horizontal tangent. However, it was Galton (see *http://galton.org*) who discovered the confidence ellipse and regression empirically. Galton (1886, 245) had collected family height data

[1] Menaechmus (c. 375–325 BC) discovered the (nondegenerate) conics or, more revealing, the conic sections. Apollonius of Perga (c. 292–190 BC) based the theory of conics on *one* type of cone and used the names ellipse, parabola, and hyperbola. It is debatable whether Apollonius introduced or canonized the improved theory of conics and the new names; e.g., see Toomer (1976, 14–15).

and found that the isodensity contour lines from the bivariate normal distribution are "concentric and similar ellipses". Galton (1886, Plate X) graphed one such ellipse with its axes, the two regression lines, and its horizontal and vertical tangents.

Galton allowed only one of the variables to contain errors and used coordinate regression. Pearson (1901, 566) extended Galton's geometrical work by allowing both variables to contain errors. Pearson assumed equal error variances and, instead, used orthogonal regression and his χ^2 distribution. Pearson called the best and worst lines of fit the major and minor axes. Pearson distinguished between the "ellipse of residuals" and the "correlation ellipse".[2]

Jones (1937, 25) graphed a scatterplot and the Pearson correlation ellipse to show that the slopes of the major and minor axes vary depending on the relative values of the standard deviations. One of several ways to handle the problem of scale is to convert the variables to standardized variables. The result is called a reduced major axis, or geometric mean regression line. It minimizes the areas of the triangles between the points and the best fit lines and is easily calculated as the slope of the least-squares line divided by the correlation coefficient. Orthogonal regression is a difficult and controversial topic; *Stata Journal* 3(4) was devoted to the closely related topic of measurement error.

McCartin (2003) did not mention Jones' article and the related controversies but, in effect, provided an elegant graphical improvement. Let the ratio of the variance of y-error to the variance of x-error be the error variance ratio: $\lambda = \sigma_y^2/\sigma_x^2$. Further, m is the slope of the best line of λ-regression, s is the slope of the worst line of λ-regression, and n is the slope of the normal, $n = -1/s$. McCartin found that the best and worst line of λ-regression are conjugate diameters of the so-called Cramér concentration ellipse, the product of whose slopes equals $-\lambda$; that is, $m * s = -\lambda$. The ratio of the λ-regression slope to the normal slope of the ellipse equals the ratio of the error variances; that is, $m/n = \lambda$. The SAS software JMP has a good graphical implementation of λ-regression; e.g., it is interactive. But JMP does not inscribe the best line in the ellipse, and JMP does not easily graph s or n.

3 The boundary constant of confidence ellipses

Hotelling (1931) derived the generalized means test $T^2 = p(n-1)/(n-p) * F(p, n-p)$, where p is the number of parameters. Hotelling concluded that T^2 can be used for determining the ellipse with a certain confidence. The empirical probability, P, that a random point falls into the corresponding ellipse is determined by the χ^2 distribution with 2 degrees of freedom, that is, the boundary constant $c = \chi^2(1-P, 2)$. Shewhart (1931, 109–112) illustrated the empirical probability with "control ellipses" for variable means. Shewhart noted that, for the bivariate normal distribution of the standard ellipse, the probability $P = 1 - \exp^{(-c/2)}$. Like Shewhart, Jackson (1956) graphed control ellipses for variable means. But unlike Shewhart, Jackson used the Hotelling T^2 as the boundary constant.

[2] In comparison, Edgeworth (1887, 284) suggested the name "probabiloid" but it never became widely used.

Cramér (1946, 275–276 and 283–285) coined the terms inertia and concentration ellipses. The inertia ellipse is an ellipse with a boundary constant of 1/moment of inertia. The concentration ellipse is an ellipse with a boundary constant of 4 because it has the same first- and second-order moments as the given distribution. McCartin (2003) suggested that the Cramér concentration ellipse is the most representative of the data points without any *a priori* statistical assumptions concerning their origin.

When the sample size n is large, the distribution $2 * F(2, n-2)$ approaches the χ^2 distribution with 2 degrees of freedom. If we know the sampling distribution, which is often the case in linear regression, this statistic is preferred because the confidence ellipse is smaller. Most of the multivariate adjustments, for example Bonferroni, produce rectangular regions instead of ellipses. The Scheffé confidence ellipse is the projection of the ellipsoid in p dimensions on the confidence ellipse, that is, $c = p * F(p, n-p)$. Durand (1954) extended the Scheffé confidence ellipse for the analysis of variance to multiple regression and was the first to explicitly use the wording "confidence ellipse".

The confidence ellipse for predicting the next single observation in the population, or prediction ellipse, approximates a confidence ellipse containing a specified percentage of the population, or tolerance ellipse. Chew (1966) distinguished between formulas for confidence, prediction, and tolerance regions for the multivariate normal distribution according to the cases of known and unknown mean vector μ and covariance matrix Σ. For samples, Chew divided the χ^2 and T^2 boundary constants by the sample size n. For large samples, if Σ is known but μ is unknown, Chew recommended using χ^2/n for the confidence ellipse and $\chi^2/n*(n+1)$ for the prediction ellipse. For small samples, if Σ is unknown, Chew recommended using T^2/n for the confidence ellipse and $T^2/n*(n+1)$ for the prediction ellipse, which are what SAS has implemented. However, SAS allows only two parameters; that is, in SAS, $T^2 = 2(n-1)/(n-2) * F(2, n-2)$. Chew also referred to more complex boundary constants for small samples.

Batschelet (1981, 141) distinguished between the standard ellipse, which has a boundary constant of 1, and the Hotelling confidence ellipse. Batschelet argued that the standard ellipse is a descriptive tool, whereas the Hotelling confidence ellipse is a tool for statistical inference. A similar but more general distinction between data ellipses and confidence ellipses exists, and it corresponds to the distinction emphasized in this paper between confidence ellipses for means and confidence ellipses for regression coefficients. Except for a constant scale factor and their respective centers, the two types of confidence ellipses are inverses of each other. For instance, if the confidence ellipse for means is positively tilted, then the confidence ellipse for regression coefficients is negatively tilted.

The standard ellipse is sometimes referred to as a standard deviation ellipse because the standard deviations are the half-widths of its shadows on the axes. But rotating the ellipse can lead to misleading results. The locus of the standard deviations of the x coordinates forms a closed curve as the system is rotated around the origin. This standard deviation curve can be, but does not have to be, an ellipse; see Gong (2002). Distances from the ellipse center to the ellipse boundary do not describe the standard deviations along directions other than the principal or major axis.

4 Elliptical distributions

Much of the work in the last two decades on confidence ellipses no longer assumes the normal distribution. For example, Goldberg and Iglewicz (1992) suggested the robust elliptical plot ("relplot") for symmetric data and the robust quarter elliptical plot ("quelplot") for nonsymmetric data. Among the non-normal distributions, the elliptically contoured distributions, or elliptical distributions, have attracted much attention. The contours of constant density are ellipsoids

$$(x-\mu)'\Sigma^{-1}(x-\mu) = \text{constant}$$

which leads to the name "elliptically contoured distribution".

Elliptical distributions extend the multivariate normal distribution $Np(\mu, \Sigma)$. For a recent overview article on elliptical distributions, see Fang (2004). Elliptical distributions contain the normal, contaminated normal, Cauchy, and Student's t-distributions among others. Olive (2002) used the plot of classical versus robust Mahalanobis distance (DD plot) to diagnose elliptical symmetry, to detect outliers, and to assess numerical methods for transforming data toward an elliptical distribution. It is beyond the scope of this paper to implement DD plots. But there is a need for similar plots, especially since StataCorp has deprecated the only command in Stata, `hadimvo`, for detecting multivariate outliers. `hadimvo` has been deprecated, at least in part, because the confidence ellipses in the earlier Stata manuals could not be replicated from the data. The problem of replication is not unique to `hadimvo`. The following sections will present and discuss the `ellip` program. One of the program examples will illustrate that also the most famous confidence ellipse, Galton's "Plate X," cannot easily be replicated from the data.

5 The ellip command

5.1 Syntax

`ellip` *yvar xvar* [`if` *exp*] [`in` *range*] [`, means` *common_options*]

`ellip` *yvar* [*xvar*] [`if` *exp*] [`in` *range*], `coefs` [`pool(`#`) plopts(`*line_options*`)`
`pcopts(`*connect_options*`)` *common_options*]

where *common_options* are

`constant(`*statname* [#] | #`) level(`#`) formula(acosr|theta)`
`generate(`*ynewvar xnewvar*`) nograph replace evr(`#`) from(`#`) to(`#`)`
`npoints(`#`) overlay total tlabel(`*label*`) diameter(`#`) dlopts(`*line_options*`)`
`plot(`*plot*`)` *connect_options twoway_options*

5.2 Description

Stata does not offer an official command for graphing confidence ellipses. Alexandersson (1998) provided the program `ellip` for Stata 5, which graphs confidence ellipses for linear regression coefficients. The program `ellip` has been rewritten for Stata 8.2 as version 2.[3] Two of the most notable new features are the option to graph confidence ellipses around variable means and the ability to add inscribed lines, such as the major and minor axes and the normals of the major axis.

The Statistical Software Components (SSC) archive (see [R] **ssc**) contains the test script package for the `ellip` program. To learn more about and to install the test scripts, type

```
. ssc describe ellip
```

5.3 New or improved options

`means` | `coefs` specifies how to center the confidence ellipse. The default and the `means` option use two variable means, whereas `coefs` uses the first two regression coefficients from an immediately preceding `regress`. Default subtitles are "Means centered" and "Coefficient centered", respectively. If you restrict `regress` to a portion of the data using `if` or `in`, you generally want to use the same conditions with `coefs`. `coefs` is not allowed with `by()` because it would be misleading, since Stata remembers only the last set of estimates.

`constant(`*statname* $\left[\#\right]$ | $\#$`)` calculates or specifies the boundary constant c. The overall and `means` default is `constant(4)` and corresponds to a joint confidence interval of $1 - \exp^{(-c/2)} * 100$ or 86%. The `coefs` default is `constant(f 2)`. Available *statnames* are specified in table 1.

(*Continued on next page*)

[3] Updated versions of `ellip` for Stata 6 and Stata 7 were posted on the SSC archive. The older versions are now named `ellip5`, `ellip6`, and `ellip7`, respectively.

Table 1: Choices for `constant(`*statname* [#]`)`

statname	Definition; calculation of c
`sd`	Standard deviation; $c = \#$
	Cannot be used with `level(#)`
`t2`	Hotelling $T^2/n; c = \#(n-1)/n(n-\#) * F(\#, n-\#)$
`pt2`	Prediction $T^2; c =$ `t2` $* (n+1)$
`chi2`	(True) $\chi^2; c =$ `chi2`$(\#)$
`chi2_n`	Sample-adjusted $\chi^2; c =$ `chi2`$(\#)/n$
`pchi2_n`	Prediction $\chi^2; c =$ `chi2`$(\#)/n * (n+1)$
`f`	(Multiplier-fixed) F; $c = 2 * F(\#, n-\#)$
`f_scheffe`	Scheffé-adjusted F; $c = \# * F(\#, n-\#)$
`fadj`	Adjusted F (denominator); $c = 2 * F(2, n-\#)$

`plopts(`*line_options*`)`, to be used with `pool()`, affects the rendition of the pooled confidence ellipse; see [G] ***line_options***.

`pcopts(`*connect_options*`)`, to be used with `pool()`, specifies how points in the fractionally pooled curve are to be connected; see [G] ***connect_options***. Sort options are ignored.

`formula(acosr|theta)` is included only for curious users to show that two seemingly different formulas are equivalent. `formula(acosr)` is based on the angular distance, acos(r), and on the amplitudes. `formula(acosr)` is the default. `formula(theta)` is based on the rotation angle of the major axis against the x-axis, θ, and on the semimajor and semiminor axes, a and b.

`nograph` specifies that no graph be constructed. In contrast, the built-in `nodraw` option merely suppresses the display of the graph. `nograph` is often used together with `generate()`.

`replace` replaces any existing variables in `generate()`.

`evr(#)` specifies the error-variance ratio of *yvar* to *xvar* as a floating-point number between 0 and 99999. The default is 1 and corresponds to the major axis (classical orthogonal regression). A ratio of 0 corresponds to regression of x on y (reversed regression with horizontal tangents), and a very large number, say 99999, corresponds to regression of y on x (with vertical tangents); see McCartin (2003).

`from(#)` specifies the value from which the ellipse parameter runs. The default is 0. Specify this option only if you do not want to display the beginning of an ellipse.

`to(#)` specifies the value to which the ellipse parameter runs. The default is 2π. Specify this option only if you do not want to display the end of an ellipse. If `from()` is smaller than `to()`, as by default, the parameter runs clockwise. Otherwise, the parameter runs counterclockwise.

`overlay`, to be used with `by()`, creates an overlaid graph for all by-groups. Despite the use of `by()`, the graph has only one plot. The `overlay` option cannot be used with `by(, total)`.

`total`, to be used with `by()`, adds an extra graph with one confidence ellipse for the entire dataset. The extra graph is located immediately after the graphs for each by-group. The default label is "Total". The label is repeated for each variable in the varlist; a better solution will probably require changes to internal by-graph files, which is beyond the scope of this paper.

`tlabel(`*label*`)`, to be used with `total`, replaces the default label, "Total", with a user-specified label. The label is repeated for each variable in the *varlist*. In contrast, you cannot easily change the label or subtitle "Total" in `by()`; you would have to manually create the parts.

`diameter(#)` adds a diameter of the ellipse with slope $-999999 \leq \# \leq 999999$. A diameter of the ellipse is any chord that goes through the center of the ellipse. `diameter()` is incompatible with `pool()`.

`dlopts(`*line_options*`)`, to be used with `diameter()`, affects the rendition of the diameter; see [G] ***line_options***.

`plot(#)` adds other plots to the generated graph; see [G] ***plot_option***. By default, when this option is specified, the legend appears; see [G] ***legend_option***. By default, the descriptive text in the legend is obtained from the y-variables' variable labels. Examples of such plots are another confidence ellipse, a regression line, inscribing reference lines, and a scatterplot of *yvar* and *xvar* and of the midpoint.

connect_options are any of the options documented in [G] ***connect_options***. The options `sort[(`*varlist*`)]` and `cmissing(y | n ...)` are ignored. The default is `cmissing(n ...)`, meaning that ellipses are not connected to each other.

twoway_options are any of the options documented in [G] ***twoway_options***. These include options for titling the graph (see [G] ***title_options***), options for saving the graph to disk (see [G] ***saving_option***), and the `by()` option, which allows you to simultaneously graph confidence ellipses for different subsets of the data (see [G] ***by_option***). The `by()` option is not allowed with the option `coefs`. The `by()` suboptions `missing` and `total` are allowed. The `by()` suboption `total` may not be used with the option `overlay`.

5.4 Dialog

The `ellip` package includes a dialog-box program for `ellip`, `ellip.dlg`. To launch the dialog interactively, type `db ellip` from within Stata 8. Alternatively, GUI users can add `ellip` permanently to the **Graphics** submenu of the **User** menu by including the following in `profile.do`:

```
if _caller() > 7 {
        if "`c(console)'"=="" {
                window menu append submenu "stUserGraphics" ///
                        "Confidence Ellipses"
                window menu append item "Confidence Ellipses" ///
                        "for Means" "db ellip_means"
                window menu append item "Confidence Ellipses" ///
                        "for Regression Coefficients" "db ellip_coefs"
        }
}
```

For information on customizing your **User** menu, see [P] **window menu**.

5.5 Saved Results

`ellip` saves in `r()`:

Scalars

`r(mean_x)`	mean of x	`r(mean_y)`	mean of y
`r(min_x)`	minimum of x	`r(min_y)`	minimum of y
`r(max_x)`	maximum of x	`r(max_y)`	maximum of y
`r(ymin_x)`	x at `ymin`	`r(xmin_y)`	y at `xmin`
`r(ymax_x)`	x at `ymax`	`r(xmax_y)`	y at `xmax`
`r(amin_x)`	x at `amin`	`r(amin_y)`	y at `amin`, min_y for n1
`r(amax_x)`	x at `amax`, max_y for n2	`r(amax_y)`	y at `amax`
`r(bmin_x)`	x at `bmin`	`r(bmin_y)`	y at `bmin`
`r(bmax_x)`	x at `bmax`	`r(bmax_y)`	y at `bmax`
`r(s_x)`	se or sd of x	`r(s_y)`	se or sd of y
`r(lambda1)`	first eigenvalue of Σ	`r(lambda2)`	second eigenvalue of Σ
`r(m1)`	slope of best fit	`r(m2)`	slope of worst fit
`r(evr)`	error-variance ratio	`r(n)`	slope of normals n1, n2
`r(a)`	semimajor axis	`r(b)`	semiminor axis
`r(a_evr)`	`evr`-adjusted a	`r(b_evr)`	`evr`-adjusted b
`r(c)`	boundary constant	`r(theta)`	rotation angle (radians)
`r(e2)`	eccentricity squared	`r(r)`	correlation coefficient
`r(n1min_x)`	min_x for n1 if $r < 0$	`r(n1max_x)`	max_x for n1 if $r > 0$
`r(n2min_x)`	min_x for n2 if $r > 0$	`r(n2max_x)`	max_x for n2 if $r < 0$
`r(n1max_y)`	max_y for n1	`r(n2min_y)`	min_y for n2
`r(mmin_x)`	min_x for diameter slope m	`r(mmin_y)`	min_y for diameter slope m
`r(mmax_x)`	max_x for diameter slope m	`r(mmax_y)`	max_y for diameter slope m

6 Remarks

6.1 Example 1: the option by(, total)

The first example will illustrate the options `total` and `overlay`, both of which are used with the option `by()`. To illustrate `by()`, the Stata 8 graphics manual uses the familiar `auto` dataset and on page 81 mentions the scatterplot example:

```
. sysuse auto, clear
. scatter mpg weight, by(foreign, total)
```

The `by()` suboption `total` adds an extra graph for all by-groups combined, in addition to the graphs for each by-group. In `ellip`, the `by()` suboption `total` adds the extra graph as overlaid confidence ellipses:

```
. ellip mpg weight, by(foreign, total)
```

The option `total` is used with `by()` to add a graph with one confidence ellipse for the total data. To add the scatterplots to the confidence ellipses, type

```
. ellip mpg weight, by(foreign, legend(off)) total plot(scatter mpg weight)
```

Use the option `tlabel()` with the option `total` to specify another label than "Total". Figure 1 is the output of this command:

```
. ellip mpg weight, by(foreign, total legend(off))
>         total tlabel(Total as a by-group) plot(scatter mpg weight)
```

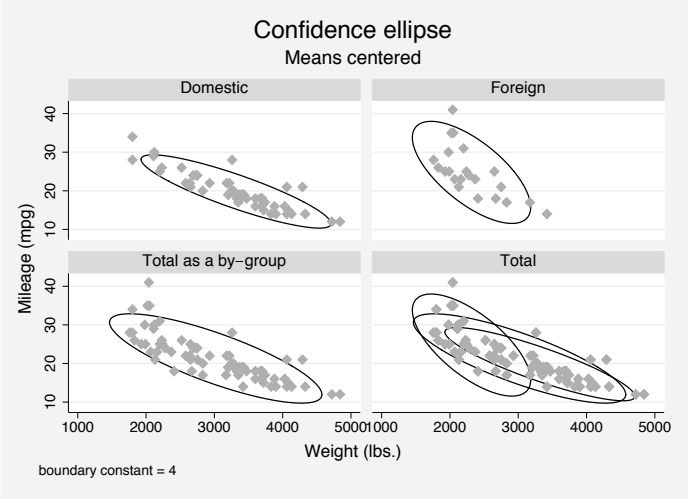

Figure 1: The option `by(, total)` on the `auto` dataset.

In figure 1, there are four graphs in a 2×2 array; the option `by(, total)` produced the (2,1) graph and the option `total` produced the (2,2) graph.

While the option `total` adds a by-graph, the option `overlay` is used with `by()` to reduce the number of by-graphs to just one overlaid graph, that is, to one graph with multiple plots. You may want to use a legend, but the default legend key from `plot()` may be misleading.

You may want to compare figure 1 with the output of this command:

```
. ellip mpg weight, by(foreign) total overlay plot(scatter mpg weight)
```

The result is one graph; the option `total` overlaid an extra confidence ellipse, and the option `overlay` produced the overlaying. If you want more flexibility, you may want to

use the `plot()` option or start over with the `generate()` option followed by overlaid twoway plots. For example, you might want to have different scatter symbols according to by-group.

6.2 Example 2: A geometric characterization of linear regression

`ellip` is a nongraph graphics command, that is, a graphics command implemented in terms of `graph twoway`. In [G] ***plot_option***, it is noted that the trade-off between graph and nongraph commands is one of greater flexibility versus easier use. To illustrate the trade-off, consider the perhaps most famous ellipse, that is, Galton (1886, Plate X). We can reproduce Galton's ellipse with the following four lines of code without using the `ellip` program:

```
. range t 0 `=2*_pi' 400
. generate x = 68.25 + 4 * cos(t)
. generate y = 68.25 + 3 * cos(t + acos(`=(4/3)/3'))
. twoway (line y x, yvarf(%4.2f) xvarf(%4.2f)), ylab(minmax) xlab(minmax)
```

In contrast, `ellip` has about 800 lines of code, and the certification script has more than 2,000 lines of code. The first line above constructs the variable `t`, taking on values from 0 to 2π, inclusive, with 400 observations. The second line generates the variable `x` with mean 68.25 and amplitude 4. The third line generates the variable `y` with mean 68.25, and amplitude 3, and correlation 4/9. Galton did not introduce the term "co-relation" or correlation until later, in 1888. But the correlation can be calculated as the ratio of the vertical tangent height to the shadow of the ellipse onto the y-axis, that is, as $(4/3)/3$. Marks (1982) showed that the slope of the line tangent to the ellipse when $x = 0$ is the correlation. When the correlation is 0, the ellipse is a circle, the tangent has slope 0, and the regression line is horizontal. When the correlation is 1 or -1, the ellipse is a straight line with slope 1 or -1, respectively, and is the major axis of the ellipse. The last line above graphs the ellipse in terms of `twoway line`.

Wachsmuth, Wilkinson, and Dallal (2003) suggested that Galton's height data are better fitted by a piecewise linear model than by a simple linear model. Using piecewise linear models is one way to find a more reasonable specification if we can reject the linearity assumption in simple linear regression. McCartin (2003, 9) suggested using e^2 as a measure of linearity instead of the correlation coefficient r because r can be zero, even when the ellipse is very elongated. The linearity of the ellipse can be defined as $e^2 = 1 - (b/a)^2$ where $0 < e < 1$ is the eccentricity of the ellipse. For `ellip parentht childht` on Galton's data, e^2 is approximately .7, whether you use the table 1 data [N=928], the Plate X data [N=314], or the fictional ellipse data. In comparison, for `ellip mpg weight` on the Stata auto data, e^2 equals .999. From this perspective, the authors correctly concluded that the Galton family-height dataset is not the preeminent linear regression dataset. But Galton never claimed that it was, and McCartin showed that Galton's concept of linear regression has not lost its power to surprise. To illustrate McCartin's (2003, 14) main results, consider the following Stata code, which creates the graph in figure 2:

```
. * reproduce Galton's confidence ellipse data
. range t 0 `=2*_pi' 4000
obs was 74, now 4000
. generate x = 68.25 + 4 * cos(t)
. generate y = 68.25 + 3 * cos(t + acos(`=(4/3)/3'))
. * extract major axis
. ellip y x, constant(2) nograph
. local m1 = r(m1)
. * create inscribing lines for best fit m1 and normal n2 for evr(.3)
. ellip y x, constant(2) evr(.3) nograph
. assert reldif( 1.067779930389807, r(m1)) < 1e-5
. assert reldif( -.2809567697067325, r(m2)) < 1e-5
. assert reldif( 3.559266434632692, r(n)) < 1e-5
. assert reldif( r(m1)/r(n), r(evr)) < 1e-5
. assert reldif( r(m1)*r(m2), -r(evr)) < 1e-5
. generate m_y = r(amin_y)
. quietly replace m_y = r(amax_y) in 2/2
. generate m_x = r(amin_x)
. quietly replace m_x = r(amax_x) in 2/2
. generate n2_y = r(n2min_y)
. quietly replace n2_y = r(amax_y) in 2/2
. generate n2_x = r(n2min_x)
. quietly replace n2_x = r(amax_x) in 2/2
. * reproduce Galton's confidence ellipse
. ellip y x, constant(2) legend(off) plot(
>         line m_y m_x, clpat(solid) text(70.5 70 "m") ||
>         line n2_y n2_x, clpat(solid) text(67 70 "n"))
>         ylab(65.25 68.25 71.25) xlab(64.25 68.25 72.25, grid)
>         ytick(65.25(1)71.25) xtick(64.25(1)72.25)
>         ytitle("Mid-parents") xtitle("Adult children")
>         diameter(`m1') dlopts(clpat(dot))
. assert reldif( .6759268231824285, r(e2)) < 1e-5
```

(*Continued on next page*)

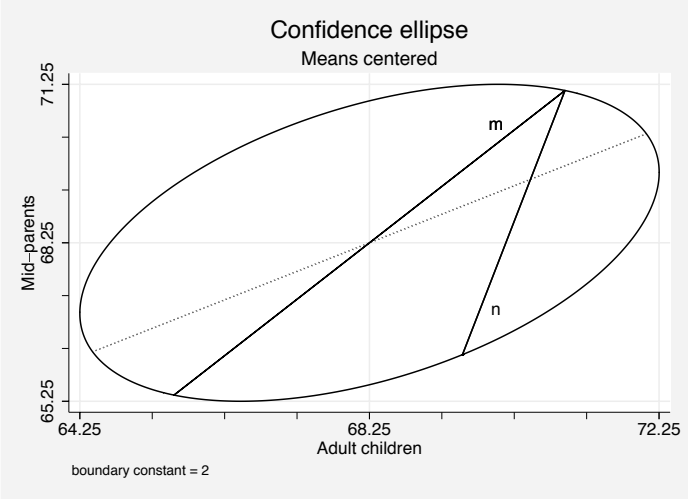

Figure 2: Error-variance regression for Galton's confidence ellipse ($m/n = \lambda$).

In figure 2, m refers to the best fit of the error-variance regression, n refers to its normal, and the dotted line is the major axis. Figure 2 does not have a legend in order to reproduce more closely McCartin's figure 8. In `ellip`, the equivalent saved results for McCartin's symbols m, s, n, and λ are `r(m1)`, `r(m2)`, `r(n)`, and `r(evr)`. McCartin found that $m/n = \lambda$ and $m * s = -\lambda$. Therefore, $1.0677/3.5593 = .3$, and $1.0677 * (-.2810) = -.3$. In addition to being useful for experimental data analysts, McCartin's geometric characterization of error-variance regression is likely to be a useful tool for teaching the concept of linear regression and the relationship between least squares and orthogonal regression.

6.3 Methods and formulas

Two versions of the parametric scalar equations of the confidence ellipse are implemented in the option `formula(acosr | theta)`. In the following equations, x and y are two arcs or variables that define the confidence ellipse in the (X,Y) coordinate system, $\overline{x}$ and $\overline{y}$ are the means for x and y, a is the major axis, b is the minor axis, θ is the rotation angle of the major axis around the x-axis, t is the variable angle that goes through the ellipse ($0 \leq t \leq 2\pi$ radians), and r is the linear correlation coefficient. Further, x_{amp} and y_{amp} are the amplitudes $s_x\sqrt{c}$ and $s_y\sqrt{c}$, where c is the boundary constant. In confidence ellipses for means, s_x and s_y are the standard deviations for x and y. In confidence ellipses for regression coefficients, s_x and s_y are the standard errors for $\overline{x}$ and $\overline{y}$, respectively.

Batschelet (1981) and others use the `formula(theta)` equations

$$x = \overline{x} + a * \cos(\theta)\cos(t) - b * \sin(z)\sin(t)$$
$$y = \overline{y} + a * \sin(\theta)\cos(t) + b * \cos(\theta)\sin(t)$$

but `ellip` defaults to the `formula(acosr)` equations

$$x = \overline{x} + x_{\text{amp}}\cos(t)$$
$$y = \overline{y} + y_{\text{amp}}\cos\{t + \text{acos}(r)\}$$

because r is easier to calculate and explain than θ. The shorter arc includes the term acos(r), which Batschelet (1981, 242) referred to as the "angular distance". The SSC package `ellip` provides formal derivations of the `formula(acosr)` equations.

7 Statistical applications and possibilities

`ellip` is useful to the practitioner, not only to illustrate ideas about linear regression and to plot parameter estimates, but also in statistical applications. A typical application is to overlay confidence ellipses on top of scatterplots to display the location, shape, and outliers of elliptical distributions. In addition to example 1, the certification script in the SSC archive provides several examples with ellipse-enhanced scatterplots. When two confidence ellipses served as a hinge and a fence of the scatterplot, Goldberg and Iglewicz (1992) viewed this as an extension of the boxplot.

To overlay graphs in Stata 8, the commands must be `graph twoway` or be implemented in terms of `graph twoway` with the `plot()` option. This requirement limits the plottypes on which you can overlay `ellip`. Two possibilities are the scatterplot matrix (`graph matrix`) and convex hull plots (`cvxhull` in the SSC archive). The scatterplot matrix is a clever arrangement of two-dimensional scatterplots but tends to get cluttered when there are many variables. Convex hull plots can replace scatterplots to reduce clutter from multigroup scatterplots of large datasets. Friendly (2002) suggested enhancing or replacing the plots in the scatterplot matrix with confidence ellipses and lowess smoothing. When the plots were replaced, Friendly called the result a correlation matrix graph.

Although newer applications tend to have less impact, the `pool()` option in `ellip` is useful for checking the robustness of pooling observations. McCartin's geometric interpretation of regression, which was illustrated in example 2, is another novel application of confidence ellipses and is useful to the experimenter who needs to allow unequal error variances.

Confidence ellipses functioning as control ellipses have long been used to identify outliers in one cluster. `ellip` works well with `hadmimvo` to identify multivariate outliers in one cluster. However, little is generally known about the distributions of the clusters representing the outliers. As the joke goes, "I'm not an outlier. I just haven't found my distribution yet!" Nevertheless, if the distributions are elliptical, consider displaying them with confidence ellipses.

8 Acknowledgments

J. B. Douglas helped to clarify the `formula(acosr)` equations, and Brian McCartin and Kai Tang Fang provided prepublication copies of their papers.

9 References

Alexandersson, A. 1998. gr32: Confidence ellipses. *Stata Technical Bulletin* 46: 10–13. In *Stata Technical Bulletin Reprints*, vol. 8, 54–57. College Station, TX: Stata Press.

Batschelet, E. 1981. *Circular Statistics in Biology*. London and New York: Academic Press.

Bravais, A. 1846. Analyse mathématique sur les probabilités des erreurs de situation d'un point. *Mémoires présentés par divers savants à l'Académie royale des sciences de l'Institut de France* 9: 255–332.

Chew, V. 1966. Confidence, prediction, and tolerance regions for the multivariate normal distribution. *Journal of the American Statistical Association* 61(315): 605–617.

Cramér, H. 1946. *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.

Durand, D. 1954. Joint confidence regions for multiple regression coefficients. *Journal of the American Statistical Association* 49(265): 130–146.

Edgeworth, F. Y. 1887. On observations relating to several quantities. *Hermathena* 6: 279–285.

Fang, K. T. 2004. Elliptically contoured distributions. In *Encyclopedia of Statistical Sciences, 2nd ed.*, ed. S. Kotz, N. L. Johnson, C. B. Read, N. Balakrishnan, and B. Vidakovic, forthcoming. New York: John Wiley & Sons.

Friendly, M. 2002. Corrgrams: Exploratory displays for correlation matrices. *American Statistician* 56(4): 316–324.

Galton, F. 1886. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland* 15: 246–263.

Goldberg, K. and B. Iglewicz. 1992. Bivariate extensions of the boxplot. *Technometrics* 34(3): 307–320.

Gong, J. 2002. Clarifying the standard deviational ellipse. *Geographical Analysis* 34(2): 155–167.

Hotelling, H. 1931. The generalization of Student's ratio. *Annals of Mathematical Statistics* 2(3): 360–378.

Jackson, J. 1956. Quality control methods for two related variables. *Industrial Quality Control* 12(7): 4–8.

Jones, H. 1937. Some geometrical considerations in the general theory of fitting lines and planes. *Metron* 13(1): 21–30.

Marks, E. 1982. A note on a geometric interpretation of the correlation coefficient. *Journal of Educational Statistics* 7(3): 233–237.

McCartin, B. 2003. A geometric characterization of linear regression. *Statistics: A Journal of Theoretical and Applied Statistics* 37(2): 101–117.

Olive, D. 2002. Applications of robust distances for regression. *Technometrics* 44(1): 64–71.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 6th ser., 2(11): 559–572.

Shewhart, W. A. 1931. *Economic Control of Quality of Manufactured Product*. New York: D. Van Nostrand.

Toomer, G. J. 1976. *Diocles: On Burning Mirrors*. New York: Springer.

Wachsmuth, A., L. Wilkinson, and G. E. Dallal. 2003. Galton's bend: A previously undiscovered nonlinearity in Galton's family stature regression data. *American Statistician* 57(3): 190–192.

About the Author

Anders Alexandersson has an MA in Social and Public Policy from Duquesne University and an MS in Business Informations Systems from Mississippi State University. He currently works as a computer consultant at Mississippi State University.