*Asian Journal of Statistics and Applications*
Vol. 1, Issue 1, 2024, pp. 17-29
https://doi.org/10.47509/AJSA.2024.v01i01.02

# GUESSING CORRELATION FROM A SCALED SCATTER PLOT AND A COVERAGE ELLIPSE

SARKAR, Jyotirmoy[a] and TULLY, Collin[b]

[a] *Department of Mathematical Sciences, Indiana University Indianapolis, Indianapolis*
[b] *Department of Computer Science, Purdue University, West Lafayette*

**ABSTRACT**
Based on a random sample, the relationship between two continuous variables is depicted by a scatter plot of the bivariate data $(x, y)$. If the relation is deemed linear, then the direction and strength of the linear association is measured by Pearson's product moment correlation coefficient $r$, which is computed using a calculator or software. Although $r$ is routinely calculated in many quantitative disciplines, most users are unskilled at guessing the magnitude of the correlation from the scatter plot alone.

To overcome this wide-spread deficiency, we propose two alterations to a scatter plot: (1) control its aspect ratio, and (2) superimpose a coverage ellipse. These two simple changes reveal to the viewer not only the correlation, but also the two regression lines, the coefficient of determination and any potential outlier.

**KEYWORDS**
linear regression model, bivariate normal distribution, least-squares method, circumscribe, coverage ellipse

Dear readers, to get the most out of this paper, please read it linearly from start to finish, and guess the correlation in each Figure as and when we ask you to.

## 1. Measuring Linear Relationship between Two Continuous Variables

When two continuous variables $x$ and $y$ are measured on $n$ items, we depict the two variables simultaneously using a scatter plot $\mathcal{S}$, which displays the $n$ points $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$ on the plane, with the horizontal coordinate showing the $x$-value, and the vertical coordinate the $y$-value. The scatter plot reveals the relationship, if any, between the two variables. If the relationship is approximately linear (that is, if the points hover around a line with some random departure), then the direction and strength of the linear relationship is measured by *Pearson's product moment correlation coefficient* $r$ (see [2]) defined as

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}. \tag{1}$$

CONTACT Author[a]. Email:jsarkar@iu.edu

**Article History**
Received : 14 March 2024; Revised : 13 April 2024, Accepted : 25 April 2024; Published : 30 June 2024

**To cite this paper**
Sarkar, Jyotirmoy & Tully, Collin (2024). Guessing Correlation from a Scaled Scatter Plot and a Coverage Ellipse. *Asian Journal of Statistics and Applications.* 1(1), 17-29.

Throughout the paper, $\Sigma$ represents summation as $i$ ranges over $\{1, 2, 3, \ldots, n\}$.

It turns out that $r$ is a real number in $[-1, 1]$. The interested reader may refer to [6] for a proof based on the Cauchy-Schwartz inequality. We will give a simpler proof in Section 2. If either $x$ or $y$ or both are linearly transformed, then both the numerator and the denominator of (1) are multiplied by the same factor, and so $r$ remains unchanged.

Based on the scatter plot most users can decide if $r$ is positive or negative (except perhaps when $|r|$ is close to 0). However, rarely one can guess the numerical value of $r$, correct to, say, within 0.025. We invite the reader to guess the correlation in the two scatter plots in Figure 1. Kindly jot down your guesses on a piece of paper. The answers will be revealed towards the end of this paper. No peeking, please.

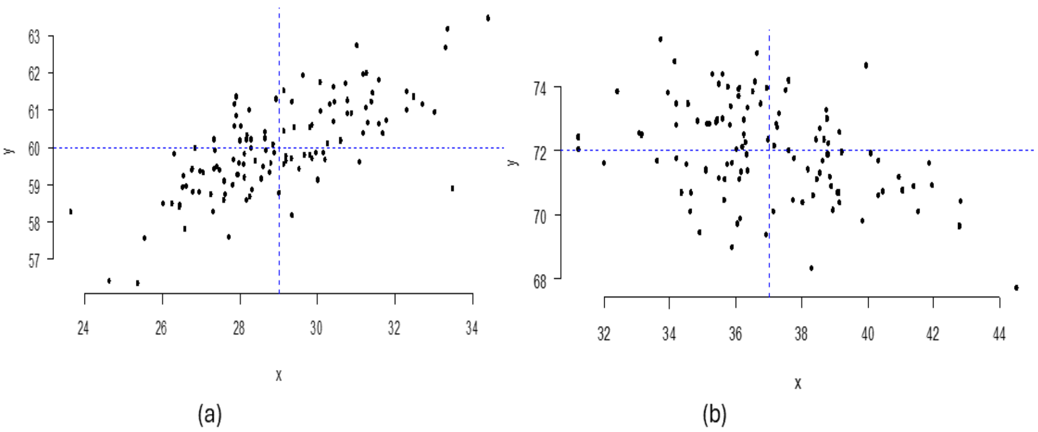


**Figure 1.** Guess the correlation in these two scatter plots, where we have superimposed vertical and horizontal lines through $(\bar{x}, \bar{y})$.

If the relationship between $x$ and $y$ is approximately linear, what is the underlying line around which the points hover? Assuming the linear regression model (LRM),

$$y_i = a + bx_i + \epsilon_i, \quad \text{where} \quad \epsilon_i \sim N(0, \sigma^2) \quad \text{independently}, \tag{2}$$

we obtain the "best" linear regression line (BLRL) $\hat{y} = a + bx$ by applying the least-squares principle: Minimize the total squared vertical distances of the observed $y$-values from the line, or $\sum(y_i - a - bx_i)^2$. Using calculus, we can show that the optimal choices of the slope and the intercept of the BLRL are given by

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \quad \text{and} \quad a = \bar{y} - b\bar{x}. \tag{3}$$

Hence, the BLRL of $\hat{y}$ on $x$ becomes

$$\hat{y}_i = a + b\, x_i = \bar{y} + b\,(x_i - \bar{x}). \tag{4}$$

## 2. Removing Location and Scale, Focus Only on Correlation

The variance of $x$ is defined as $\mathrm{var}(x) = s_x^2 = \sum(x_i - \bar{x})^2/(n-1)$, or the "average" squared deviation from the mean. [We do not divide by $n$, but by $(n-1)$, because among all $n$ deviations $(x_1 - \bar{x}), (x_2 - \bar{x}), (x_3 - \bar{x}), \ldots, (x_n - \bar{x})$, one is redundant as $\sum(x_i - \bar{x}) = 0$. This reduces the degrees of freedom to $(n-1)$.] The positive square-root of $\mathrm{var}(x)$, $s_x$, is called the standard deviation (SD) of $x$. Thus, $\bar{x}$ is a location measure, and $s_x$ is a scale measure. Next, the standardized value (or $z$-score) of $x_i$ is defined as $z(x_i) = (x_i - \bar{x})/s_x$. The $z$-score is a location- and scale-free measure which quantifies by how many SDs a datum $x_i$ is away from the mean $\bar{x}$.

Likewise, we define the variance of $y$, $\mathrm{var}(y) = s_y^2$, the SD $s_y$ of $y$, and the standardized value $z(y_i)$ of $y_i$. Note that the $z$-scores are pure numbers (with no units attached), with mean 0 and SD 1; that is, $\bar{z}(x) = 0 = \bar{z}(y)$ and $s_{z(x)} = 1 = s_{z(y)}$.

Finally, we define $\mathrm{cov}(x, y) = s_{xy} = \sum(x - x_i)(y - y_i)/(n-1)$ called the covariance between $x$ and $y$. Starting from (1), $r$ can be written as

$$r = \frac{s_{xy}}{s_x s_y} = \mathrm{cov}(z(x_i), z(y_i)) = \frac{1}{n-1}\sum z(x_i)z(y_i), \qquad (5)$$

or roughly the "average" product of $z$-score of $x_i$ and $z$-score of $y_i$.

A more elementary proof that $-1 \le r \le 1$ follows from

$$0 \le \mathrm{var}(z(x_i) \pm z(y_i)) = \mathrm{var}(z(x_i)) + \mathrm{var}(z(y_i)) \pm 2\mathrm{cov}(z(x_i), z(y_i)) = 2 \pm 2r.$$

Next, from (1) and (3), we have $b = rs_y/s_x$, and from (4), the BLRL becomes

$$\frac{\hat{y}_i - \bar{y}}{s_y} = r\,\frac{x_i - \bar{x}}{s_x}.$$

which is also known as the regression equation (see [1]); or equivalently, as

$$z(\hat{y}_i) = r \cdot z(x_i). \qquad (6)$$

Equation (6) motivates us to replace the usual scatter plot $\mathcal{S}$ of $(x_i, y_i)$ into a doubly linearly transformed scatter plot $\tilde{\mathcal{S}}$ of standardized values $(z(x_i), z(y_i))$ drawn with aspect ratio 1:1; that is, we ensure that the unit measures on the horizontal and the vertical axes are *exactly the same*. To aid the viewer, we also show a background grid $\{x = k, y = k : k = -4, -3, \ldots, 3, 4\}$. Now, to guess the correlation $r$, the user must estimate only the slope of the BLRL $z(\hat{y})$ on $\tilde{\mathcal{S}}$. We invite the readers to guess and write down the correlation in the two standardized scatter plots in Figure 2.

Likewise, we can construct the BLRL of $\hat{x}$ on $y$. We refer the reader to papers [3]–[5] to gain a perspective on visualizing the correlation via the two regression lines or two coverage ellipses.

## 3. On $\tilde{\mathcal{S}}$ superimpose the BLRL $z(\hat{y})$ or a Coverage Ellipse or Both

Since the users must estimate only the slope of the BLRL $z(\hat{y})$ on $\tilde{\mathcal{S}}$ to guess the correlation $r$, their tasks will be simpler if we superimpose the BLRL $z(\hat{y})$ on $\tilde{\mathcal{S}}$.

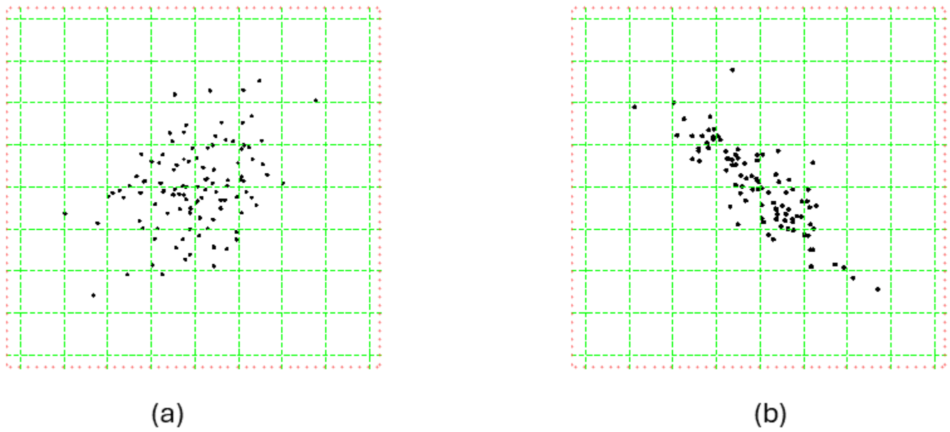


**Figure 2.** Guess the correlation in these two standardized scatter plots, with a background grid $\{x = k, y = k : k = -4, -3, \ldots, 3, 4\}$.

Furthermore, let us superimpose on $\tilde{\mathcal{S}}$ a $100(1-\alpha)\%$ coverage ellipse $\mathcal{E}_{1-\alpha}$. The idea of the coverage ellipse comes from a bivariate normal distribution. Since the contour plot of a bivariate normal distribution is a family of concentric ellipses that are dilation of one another about the center at the mean vector (see [7]), the probability that $(X, Y)$ falls within any ellipse in this family is the highest among all regions having the same Euclidean area. Consequently, $100\alpha\%$ of the data are expected to fall outside $\mathcal{E}_{1-\alpha}$. (The actual number of points that fall outside $\mathcal{E}_{1-\alpha}$ is distributed as a binomial $(n, \alpha)$ distribution, or approximately a Poisson$(n\alpha)$ distribution.) We recommend choosing $\alpha = .02$, but the user can choose any other value. Note that while the bivariate normal distribution satisfies almost all assumptions of the LRM (2) except homoscedasticity (equal variance of random variables $\epsilon_i$), the converse is not true. However, we take liberty to use the coverage ellipse $\mathcal{E}_{1-\alpha}$ to detect potential outliers in an LRM.

Using a coverage ellipse $\mathcal{E}_{1-\alpha}$, a trained viewer can reconstruct the BLRL $z(\hat{y})$ on $z(x)$ as follows: On the coverage ellipse, join the leftmost point to the rightmost point. The justification is given in Section 5. By symmetry, this line must pass through the center $(0, 0)$ of the coverage ellipse. Thereafter, using the background grid, the viewer can read off the slope of the BLRL $z(\hat{y})$ on $z(x)$ to guess the correlation $r$. Indeed, the horizontal coordinate of the rightmost point of coverage ellipse $\mathcal{E}_{.98}$ being 2.8, the viewer must only read off the vertical coordinate using the background grid, and divide it by 2.8. [You may draw coverage ellipse $\mathcal{E}_{.9889}$ to make the denominator exactly 3.]

Of course, we can help the user further by drawing the BLRL $z(\hat{y})$ on $z(x)$ together with the coverage ellipse. In Figure 3, we have superimposed the BLRL $z(\hat{y})$ and/or $\mathcal{E}_{.98}$ on the standardized scatter plots. We invite readers to guess the correlation in all four standardized scatter plots and write them down.

The coverage ellipse has two additional advantages: (1) one can draw not only the BLRL of $z(\hat{y})$ on $z(x)$ (by joining the leftmost and the rightmost points on the ellipse), but also the BLRL of $z(\hat{x})$ on $z(y)$ (by joining the bottommost and the topmost points), and (2) it flags any potential regression-outlier or bivariate-outlier, in addition to $x$-outliers and $y$-outliers, which may have been already detected using univariate techniques such as dot plot and box plot.

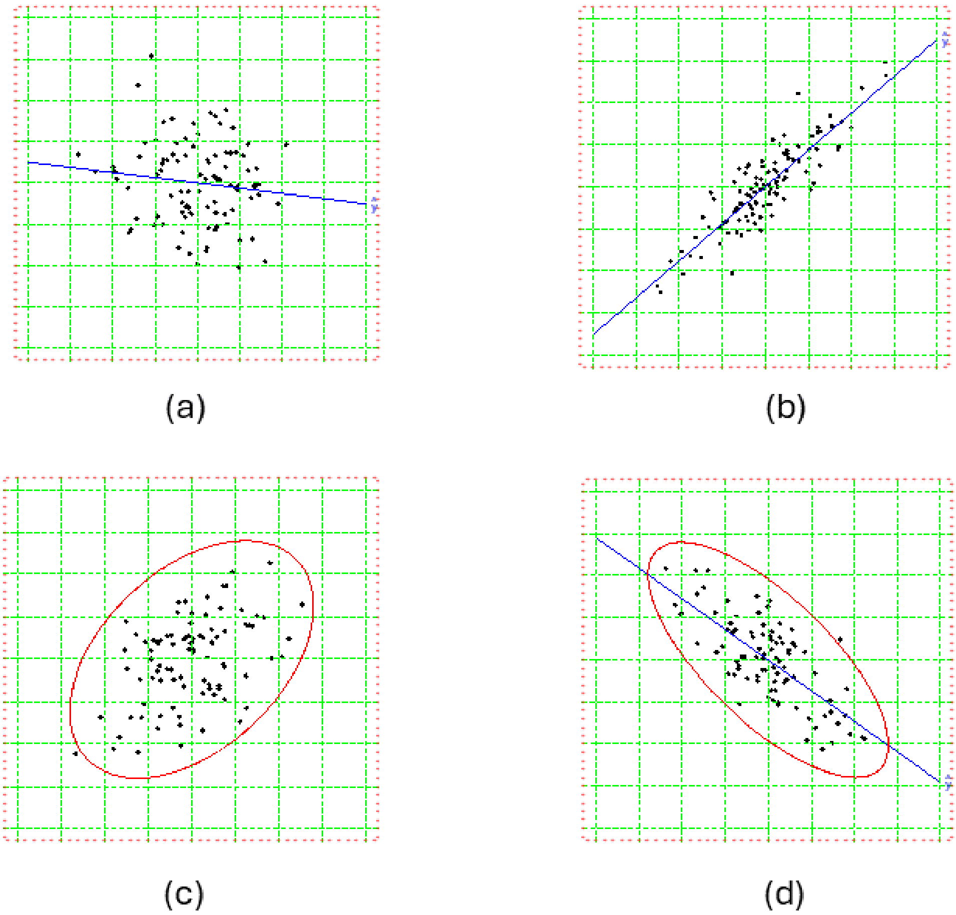


**Figure 3.** Guess the correlation in these four **standardized** scatter plots, with a background grid $\{x = k, y = k : k = -4, -3, \ldots, 3, 4\}$, and the BLRL $\hat{y}$, and/or the coverage ellipse $\mathcal{E}_{.98}$.

## 4. Guessing Correlation Without Standardizing

Having explained how to guess the correlation in a scatter plot of standardized variables drawn with aspect ratio 1 : 1, we can return to the original scatter plot $\mathcal{S}$ before standardization. We recommend drawing the scatter plot $\mathcal{S}$ by controlling the aspect ratio so that the physical horizontal distance (on paper or screen) representing $s_x$ equals the physical vertical distance representing $s_y$. Let us call this the *scaled scatter plot*. Then correlation $r$ is simply the physical slope of the BLRL of $\hat{y}$ on $x$, disregarding the scales of the coordinate axes. Also, then the two axes of a coverage ellipse have physical slopes 1 and $-1$.

Of course, estimating the slope is easier if we also print one or both BLRLs and/or the coverage ellipse. Thus, we preserve the meaningful units of measurements and yet let users guess the correlation $r$ and detect potential outliers. We strongly advocate for such alterations to any scatter plot. We invite readers to guess and write down the correlation in all four unstandardized but scaled scatter plots in Figure 4, where we have superimposed $\mathcal{E}_{.98}$ and/or BLRL $\hat{y}$ and BLRL $\hat{x}$.

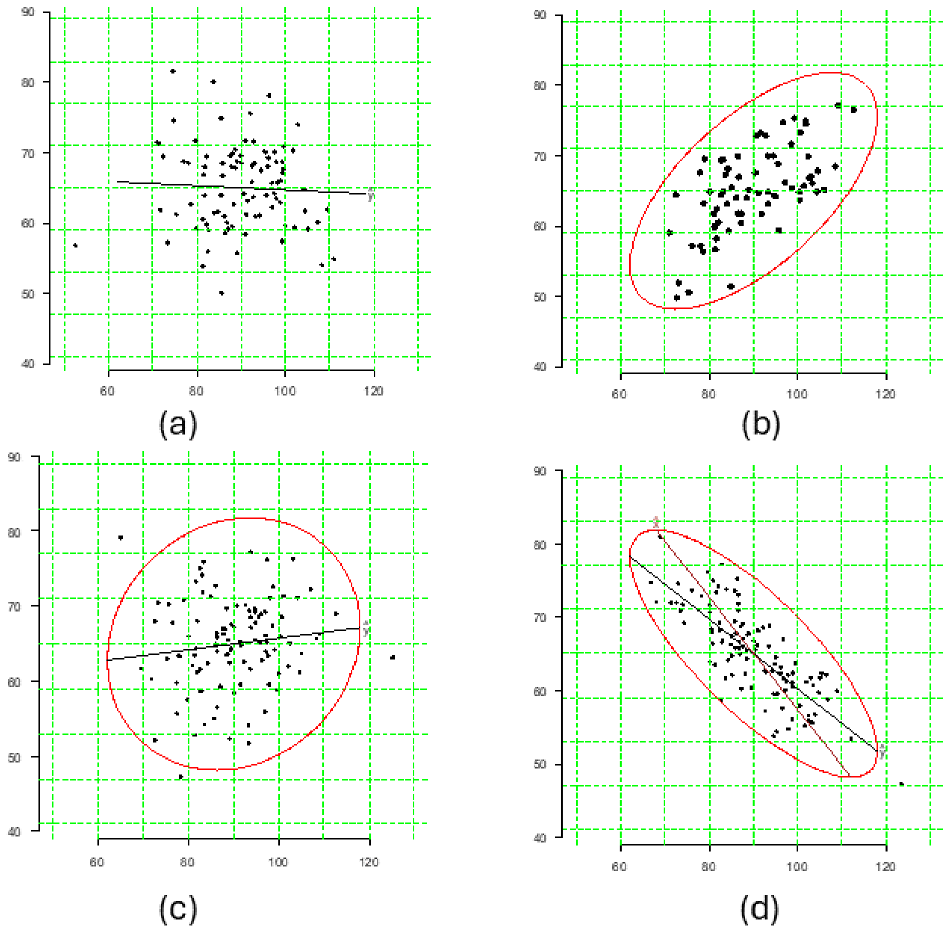


**Figure 4.** Guess the correlation in these four **scaled** scatter plots, with a background grid $\{x = \bar{x} + ks_x, y = \bar{y} + ks_y : k = -4, -3, \dots, 3, 4\}$, and the BLRL $\hat{y}$ and/or the coverage ellipse $\mathcal{E}_{.98}$.

## 5. Correlation and Coverage Ellipse on a Scaled Scatter Plot

Section 3 stated how to obtain the BLRL $z(\hat{y})$, given a coverage ellipse on $\tilde{\mathcal{S}}$: Join the leftmost point to the rightmost point. Likewise. to obtain the BLRL $z(\hat{x})$, join the bottommost point to the topmost point. Here we justify why the method works. We prove mathematical results establishing that the correlation uniquely determines the coverage ellipse on a scaled scatter plot and vice versa.

**Proposition 5.1** (Boundary Point to Ellipse)**.** *Given a square with sides horizontal and vertical, let $R(r, 1)$ be a point on the top side $ZW$ representing $[-1, 1]$. Then one and only one ellipse passes through $R$, has axes on the diagonals of the given square, and is tangent to all four sides of the square.*

**Proof of Proposition 5.1**:
The given square has two pairs of orthogonal lines of symmetry — horizontal and vertical mid-lines, and the diagonals. Only the latter pair can serve as the orthogonal lines of symmetry for the required ellipse passing through the given point $R$. Clearly, the diagonals of the given square are on lines $y = x$ and $y = -x$, respectively.

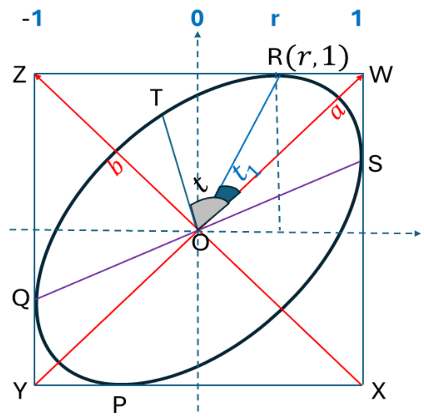


**Figure 5.** Given a square $XYZW$ and a point $R$ on the top edge $ZW$, there is a unique coverage ellipse passing through $R$ and tangent to all four sides of the square.

Consider any ellipse $\mathcal{E}$ centered at $(0,0)$ with half-axes $a$ and $b$ in the directions $y = x$ and $y = -x$, respectively. As $T$ varies over the ellipse, let $t$ denote the angle of rotation from $OW$ to $OT$. Then the parametric equation of the ellipse is

$$\mathcal{E} = \{(x(t) = (a\cos t - b\sin t)/\sqrt{2}, y(t) = (a\cos t + b\sin t)/\sqrt{2}) : 0 \le t \le 2\pi\}$$

Let $t_1$ be the argument corresponding to the boundary point $R = (r,1)$ where $y(t)$ is maximized; that is, $x(t_1) = r$, $y(t_1) = 1$ and $y'(t_1) = 0$. Equivalently,

$$\begin{aligned} a\cos(t_1) - b\sin(t_1) &= \sqrt{2}\,r \\ a\cos(t_1) + b\sin(t_1) &= \sqrt{2} \\ -a\sin(t_1) + b\cos(t_1) &= 0 \end{aligned} \tag{7}$$

The half-difference and the half-sum of the first two equations of (7) yield

$$b\sin(t_1) = \frac{1-r}{\sqrt{2}} \tag{8}$$

$$a\cos(t_1) = \frac{1+r}{\sqrt{2}} \tag{9}$$

which when put into $ab$ times the third equation of (7), yields

$$a^2(1-r) = b^2(1+r) = \kappa, \text{ say} \tag{10}$$

To evaluate $\kappa$, multiply the Pythagorean identity $1 = \sin^2 t_1 + \cos^2 t_1$, by $2a^2b^2$ (Why?), and into that resulting identity, substitute (8) and (9), to get

$$\begin{aligned} 2a^2b^2 &= 2a^2b^2(\sin^2(t_1) + \cos^2(t_1)) \\ &= a^2(1-r)^2 + b^2(1+r)^2 = \kappa\,[(1-r) + (1+r)] = 2\kappa \end{aligned}$$

Hence, $\kappa = a^2b^2$, substituting which in (10), we get $a^2 = 1 + r$ and $b^2 = 1 - r$. Since $a$ and $b$ are unique in $r$, the ellipse $\mathcal{E}$ is unique.

Lastly, by the reflection symmetry of $\mathcal{E}$ about its axes, it must be tangent to the other three sides $WX, XY, YZ$ of the given square at $S, P, Q$, respectively. □

**Proposition 5.2** (Ellipse to Boundary Point)**.** *Given an ellipse with center* $(0, 0)$ *and axes on the lines* $y = x$ *and* $y = -x$*, respectively, its circumscribing rectangle with horizontal and vertical sides is a square. If the top side* $ZW$ *of that square represents the line segment* $[-1, 1]$*, then the point* $R$ *where the ellipse is tangent to the top side of the square splits it in the ratio* $1 + r : 1 - r$*; hence,* $R$ *represents* $r$*.*

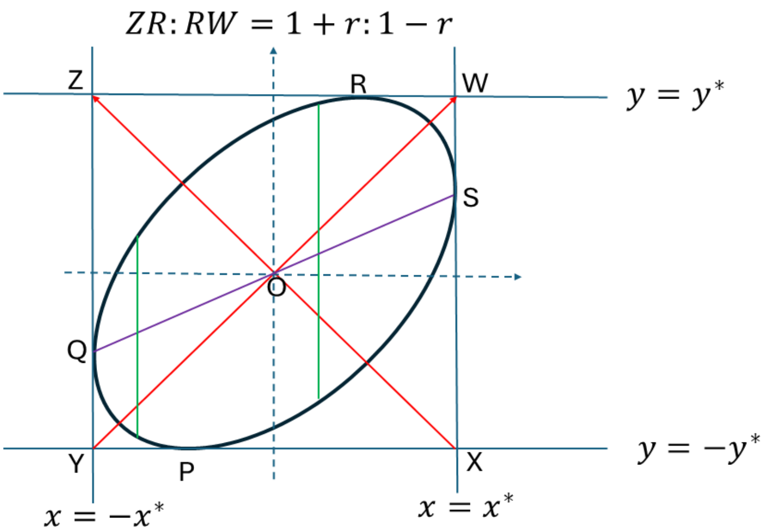


**Figure 6.** Given an ellipse with axes on $y = x$ and $y = -x$, its circumscribing rectangle with horizontal and vertical sides is a square. The top edge is split in the ratio $1 + r : 1 - r$ at the tangent point $R$.

**Proof of Proposition 5.2**:

Since the given ellipse $\mathcal{E}$ has center $(0, 0)$, we can sandwich it between horizontal lines $y = -y^*$ and $y = y^*$ touching it at $P, R$, respectively. Also, we can sandwich it between vertical lines $x = -x^*$ and $x = x^*$ touching it at $Q, S$, respectively. Let these four tangent lines intersect to form rectangle $XYZW$. Because $\mathcal{E}$ has reflection symmetry about each of its axes $y = x, y = -x$, by folding the figure along $YW$, we see that $WZ$ falls on $WX$ and $R$ falls on $S$, proving that $x^* = y^*$. Thus, the unique circumscribing rectangle is a square.

Re-scaling the circumscribing square (that is, dividing all measures by $x^*$, we let $ZW$ represent $[-1, 1] \times \{1\}$. Let us define $r$ as the common ratio

$$\frac{ZR - RW}{ZW} = \frac{XS - SW}{XW} = \frac{XP - PY}{XY} = \frac{ZQ - QY}{ZY}$$

Then $R = (r, 1)$ splits $ZW$ in the ratio $(1 + r) : (1 - r)$. Given the square $XYZW$ and the point $R = (r, 1)$ on $ZW$, in view of Proposition 5.1, $\mathcal{E}$ is the unique ellipse tangent to all four sides of the square $ZWXY$ and passing through $R, S, P, Q$. □

### *5.1.* Connecting the boundary point $R(r, 1)$ to the correlation $r$

In Proposition 5.1, we can begin by choosing $R(r, 1)$, where $r$ is the correlation between variables $x$ and $y$. In Proposition 5.2, the slope of $QS$, where $Q$ is the leftmost point and $S$ is the rightmost point of the coverage ellipse, is $r$. It remains to show that this ratio $r$ is indeed the correlation between variables $x$ and $y$ over whose scaled scatter plot we are given a coverage ellipse $\mathcal{E}$. In the next paragraph, we show more: We justify that $QS$ is the BLRL $\hat{y}$ on $x$. Likewise, $PR$ is the BLRL $\hat{x}$ on $y$.

In a LRM (and also for a bivariate normal distribution), the conditional distribution of $Y|X = x$ is normal with mean $a + bx$ on the BLRL $\hat{y}$ on $x$. Hence, the midpoint of any vertical line segment terminated by the ellipse (see the two green lines in Figure 6) is on the BLRL $\hat{y}$. But a line is uniquely determined by any two points on it. So, the line joining the leftmost point $Q$ and the rightmost point $S$ of the coverage ellipse is the BLRL $\hat{y}$ on $x$, and its slope is the correlation $r$, provided the physical horizontal space representing $s_x$ equals the physical vertical space representing $s_y$ (and hence the axes of the coverage ellipse have slopes 1 and $-1$, respectively). Likewise, the midpoint of any horizontal line segment terminated by the ellipse is on the BLRL $\hat{x}$ on $y$, which can be found by joining the bottommost point $P$ and the topmost point $R$ of the coverage ellipse.

## 6. Guessing the correlation $r$ from a coverage ellipse

Given a coverage ellipse on a scaled scatter plot, here are several equivalent ways to ascertain the value of $r$. Refer to Figure 7 for all points and lines mentioned below.

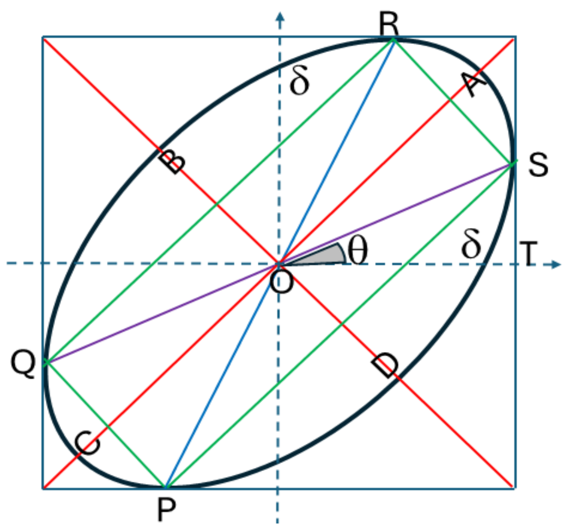


**Figure 7.** The coverage ellipse, its circumscribing square, and several reference lines offer different ways to guess the correlation.

(1) The line $QS$ joining the leftmost point $Q$ and the rightmost point $S$ of the coverage ellipse is the BLRL $\hat{y}$ on $x$, and its slope is $r$.
(2) The line $PR$ joining the bottommost point $P$ and the topmost point $R$ of the coverage ellipse is the BLRL $\hat{x}$ on $y$, and its slope is $1/r$.

(3) Let $\beta = OA : OB$ be the ratio of the half-axes of the ellipse with slopes 1 and $-1$, respectively. Then $\beta = \sqrt{(1+r)/(1-r)}$. Hence, $r = (\beta^2 - 1)/(\beta^2 + 1)$.
(4) Let $\gamma = QR : RS$ be the ratio of the sides of rectangle $PQRS$. Then $\gamma = (1+r)/(1-r)$. Hence, $r = (\gamma - 1)/(\gamma + 1)$.
(5) The horizontal and the vertical lines through $O$ intersect the ellipse at a distance $\delta = \sqrt{1-r^2}$ from $O$. Hence, $r = \sqrt{1-\delta^2}$.
(6) Let $\theta = \angle TOS$. Then $r = \tan\theta = ST/OT$.

We encourage readers to apply several of the above-mentioned methods to guess the value of the correlation based on a coverage ellipse on a scaled scatter plot, and commit to the value only when all guesses agree. We invite them to guess and write down the correlation in all four scaled scatter plots in Figure 8, where we have superimposed $\mathcal{E}_{.98}$ and some reference lines.

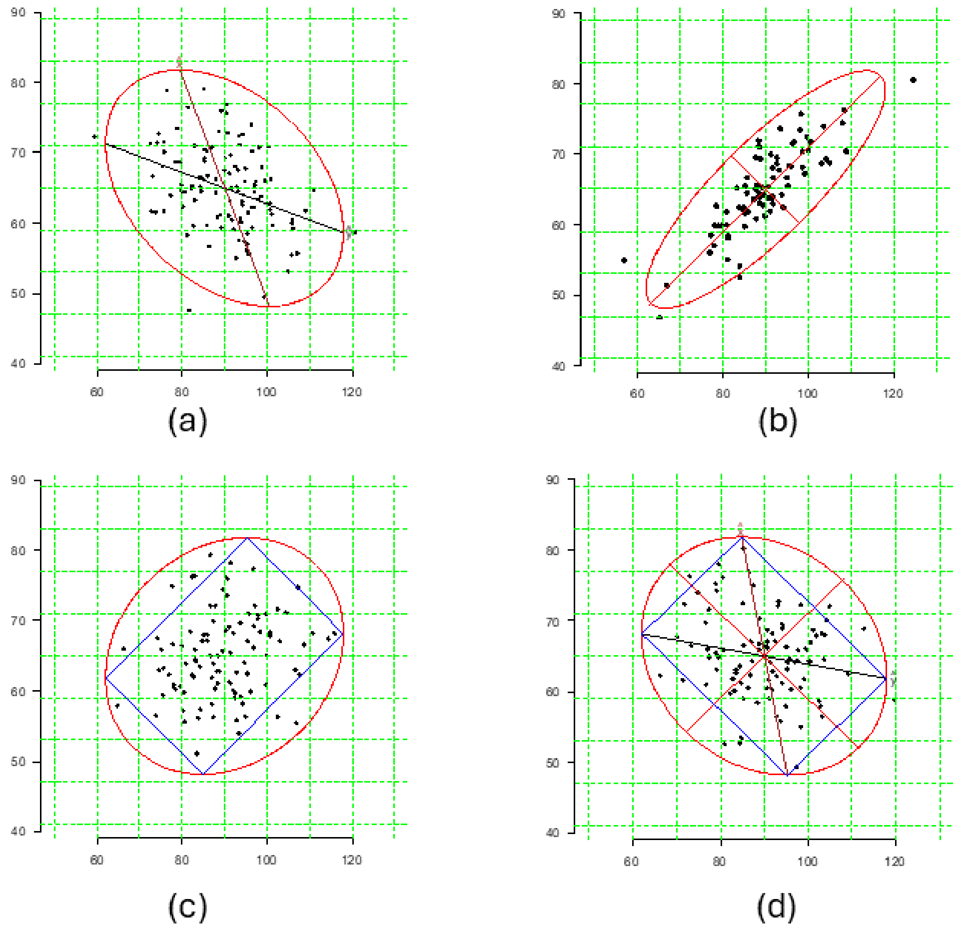


**Figure 8.** Guess the correlation in these four **scaled** scatter plots, with a background grid $\{x = \bar{x} + ks_x, y = \bar{y} + ks_y : k = -4, -3, \ldots, 3, 4\}$, the coverage ellipse $\mathcal{E}_{.98}$ and/or some reference lines.

Given the coverage ellipse, the reader can impose one or more of these reference lines as desired. This is why we recommend printing a coverage ellipse on a scaled scatter plot drawn with an aspect ratio that ensures $s_x$ and $s_y$ are given equal Euclidean length. In Figure 9, we do not provided the background grid any more, instead we draw the ellipses corresponding to $r = .1(.1).6$ (in red) and $r = .70(.05).95$ (in blue), together with the ratio $\beta = \sqrt{(1+r)/(1-r)}$ of the two axes. In each panel, the black ellipse, a circle corresponding to $r = 0$, serves as a reference. Readers may want to develop a photographic memory of these ellipses. For $r < 0$, the corresponding ellipse is a vertical reflection of the ellipse corresponding to $|r|$.

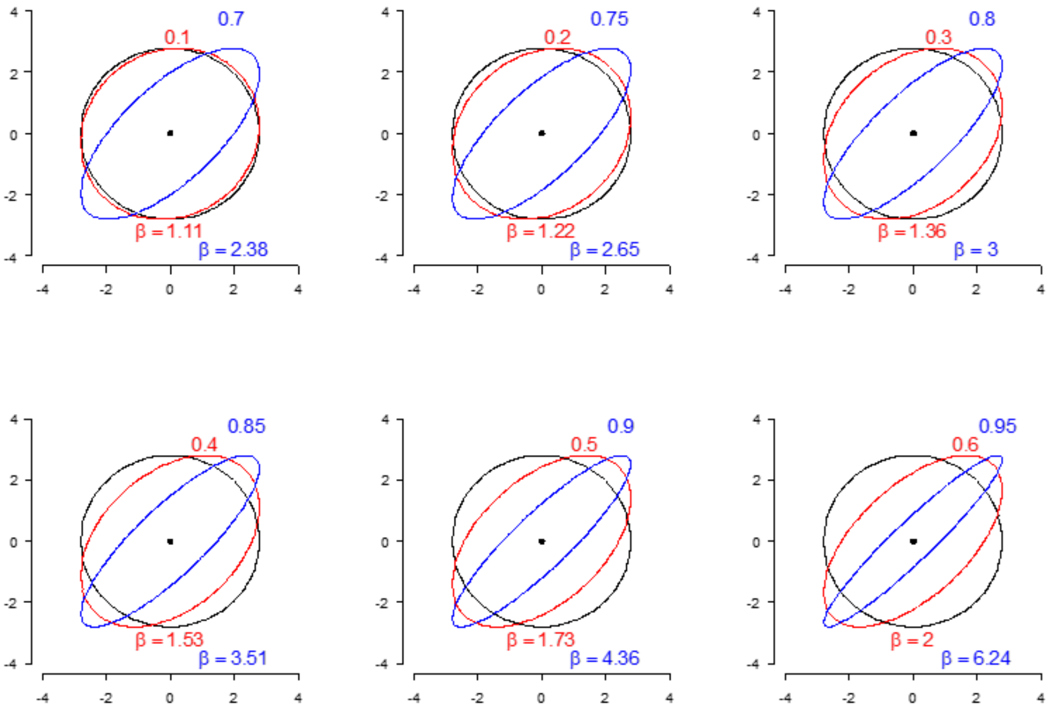


**Figure 9.** Develop a visual memory of the correlation ellipses for $r = .10(.10).60$ (in red) and for $r = .70(.05).95$ (in blue), with associated ratio $\beta = \sqrt{(1+r)/(1-r)}$ of axes.

We invite readers to guess the correlations in Figure 10 by imposing any reference line(s) they wish, or by relying on their visual memory of Figure 9. They may also guess the values of $\bar{x}, \bar{y}, s_x, s_y$ by using the coordinate scales. Please do so before reading the next section.

## 7. Summary

By virtue of Propositions 5.1 and 5.2, there is a one-to-one correspondence between correlation $r$ and $45^o$-slanted ellipse with a predetermined coverage. While the former is a numeric quantity, the later is a visual entity. If on a scaled scatter plot a 98% coverage ellipse is superimposed, then a trained viewer will decipher the correlation $r$, the BLRL $\hat{y}$ on $x$ and the BLRL $\hat{x}$ on $y$, and the coefficient of determination $r^2$, which measures the goodness-of-fit of the LRM by reporting the proportion of total variation in $y$ explained by the model. They will also detect any potential outlier.

We invite the readers to check how well they have guessed the correlation in Figures 1–4, and 8. We hope that guessing got progressively easier from start to finish, and in

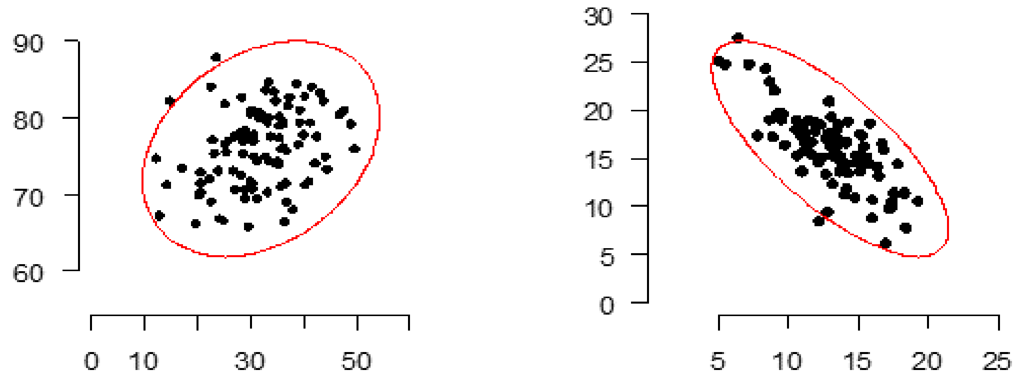


**Figure 10.** Guess the correlation in this scaled (that is, $s_x$ and $s_y$ are given equal Euclidean length) scatter plot showing the coverage ellipse $\mathcal{E}_{.98}$. Guess also $\bar{x}, \bar{y}, s_x, s_y$.

Figure 10, they have correctly guessed (a) $r = 0.300, \bar{x} = 32, \bar{y} = 76, s_x = 8, s_y = 5$ and (b) $r = -0.746, \bar{x} = 13, \bar{y} = 16, s_x = 3, s_y = 4$.

```
Correlation values in Figures 1-4, 8.
Figure 1: (a)  .733, (b) -.393;
Figure 2: (a)  .354, (b) -.838;
Figure 3: (a) -.127, (b)  .878, (c)  .444, (d) -.726;
Figure 4: (a) -.050, (b)  .630, (c)  .130, (d) -.786;
Figure 8: (a) -.375, (b)  .839, (c)  .182, (d) -.186.
```

Any reader, who has diligently written down the guessed correlation values as and when they were asked to do so, will appreciate how comprehension of correlation is aided by the two simple alterations we have proposed: (1) control the aspect ratio to draw a scaled scatter plot, and (2) superimpose a 98% coverage ellipse.

## Acknowledgement

We are grateful to Jack Dwyer for collecting data on students guessing the correlation from standardized scaled scatter plots with or without BLRL and/or coverage ellipse.

## References

[1] Galton, Francis (1886). Regression Towards Mediocrity in Hereditary Stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*. 15: 246–263. doi:10.2307/2841583. ISSN 0959-5295. JSTOR 2841583.

[2] Kendall, M. G., Stuart, A. (1973) *The Advanced Theory of Statistics, Vol 2: Inference and Relationship*, Griffin. ISBN 0-85264-215-6 (Section 31.19)

[3] Jyotirmoy Sarkar and Mamunur Rashid (2021). Two regression lines suffice to determine $r^2$ and $r$, *Educational Research Quarterly*, 45(2), 21–30.

[4] Jyotirmoy Sarkar and Mamunur Rashid (2021) Visualizing bivariate data: What's your point of view? *Journal of Probability and Statistical Science*, 19(1), 83–95.

[5] Jyotirmoy Sarkar, Mamunur Rashid and Sidhanta Phuyal, (2022). Visualizing bivariate

statistics using ellipses over a scatter plot. *Journal of Probability and Statistical Science*, 20(1), 150-165.

[6] Solomentsev, E. D. (2001) [1994], Cauchy inequality, *Encyclopedia of Mathematics*, EMS Press.

[7] Eric W. Weisstein, Bivariate Normal Distribution, MathWorld—A Wolfram Web Resource. https://demonstrations.wolfram.com/TheBivariateNormalDistribution/