Journal of Probability and Statistical Science

20*(1), 150-165, August 2022*

Visualizing Bivariate Statistics Using Ellipses Over a Scatter Plot

Jyotirmoy Sarkar
Indiana University-Purdue University

Mamunur Rashid
DePauw University

Siddhanta Phuyal
DePauw University

ABSTRACT

A scatter plot shows the relationship between two quantitative variables $\boldsymbol{x}$ and $\boldsymbol{y}$. Sometimes, we can predict one variable as a linear function of the other using the least squares regression lines of $\boldsymbol{y}$ on $\boldsymbol{x}$ or $\boldsymbol{x}$ on $\boldsymbol{y}$. These two regression lines together suffice to identify the mean vector, the coefficient of determination, the correlation coefficient, and the ratio of the standard deviations (SD). So, do our proposed summary ellipses. Additionally, the inner ellipse reveals the SDs and the outer ellipse flags potential outliers.

Keywords: dot plot, boxplot; Gaussian interval plot; summary ellipse

---

□ Received September 2021, revised December 2021, in final form January 2022.

□ Jyotirmoy Sarkar is affiliated with the Department of Mathematical Sciences, Indiana University-Purdue University, Mamunur Rashid and Siddhanta Phuyal are affiliated with Department of Mathematical Sciences, DePauw University, Greencastle, IN 46135

1. INTRODUCTION

For a quantitative variable, a dot plot (Wilkinson 1999) depicts the data unabridged, whereas a boxplot (Spear 1952, 1969) depicts only a few percentiles and flags potential outliers. If the variable is approximately normally distributed, as an alternative to a boxplot, Sarkar and Rashid (2021) proposed a Gaussian interval (GI) plot to depict two intervals: $(\bar{x} - cs, \bar{x} + cs)$ for $c = 1$ and $c = 2.33$, using a shorter, solid two-headed arrow and a longer, dotted two-headed arrow, respectively. The longer arrow may be shortened from its intended expanse to reach only up to the most extreme value(s) within the expanse. From the GI plot the mean (the center of the shorter arrow), and the standard deviation (the half-length of the shorter arrow) can be read off easily, and all points outside the larger arrow (roughly 2%) can be identified as potential outliers, together with their frequencies. An R package of GI plot is available on CRAN (R Core Team 2021.)

Two quantitative variables measured on the same set of items are depicted simultaneous with a possible small distortion (due to overlapping points whose frequencies are lost) by using a scatter plot (Cleveland 1993), where the horizontal and the vertical axes represent the two variables. How should the scatter plot be summarized so that most of the desirable summary statistics can be recovered?

If no two scatter points coincide (or if the frequencies of the points are recoverable), then the scatter points can be projected onto the horizontal or the vertical axis to reconstruct the dot plot of the corresponding variable. Thereafter, these dot plots can be summarized into box plots and GI plots. Naturally, one can superimpose on the scatter plot the GI plot

of x drawn parallel to the x-axis, and the GI plot of y drawn parallel to the y-axis, and the two GI plots intersecting at $(\bar{x}, \bar{y})$. See Figure 1(a). However, these two GI plots do not exhibit the relationship between x and y. Something else in needed.

To capture the correlation coefficient and the coefficient of regression, one might wish to draw the two regression lines. We will review how to extract some summary statistics from the two regression lines together. However, in this paper, we propose to superimpose on the scatter plot two Gaussian summary ellipses given by

$$(x, y): (x - \bar{x})' S^{-1} (y - \bar{y}) = c^2, \text{ with } c^2 = 1 \text{ and } qf(.98, 2, n-2) \qquad (1)$$

where $qf(.98, 2, n-2)$ is the 98th percentile of an F-distribution with degrees of freedom 2 and $(n-2)$. See Figure 1(b). These ellipses are concentric at $(\bar{x}, \bar{y})$; they share the same directions of axes but differ in magnification factors. Under the assumption of bivariate normal distribution, the smaller ellipse contains roughly a proportion *pf*(1, 2, *n*-2) of the data, obtained by evaluating at 1 the cumulative distribution function of an F-distribution with degrees of freedom 2 and $(n-2)$, and the larger ellipse contains about 98% of the data. See Johnson and Wichern (2007). Thus, points outside the larger ellipse (roughly 2%) are flagged as potential outliers. What the two arrows of the GI plot do for the univariate case, the two ellipses over a scatter plot do the same for the bivariate case, and then some.

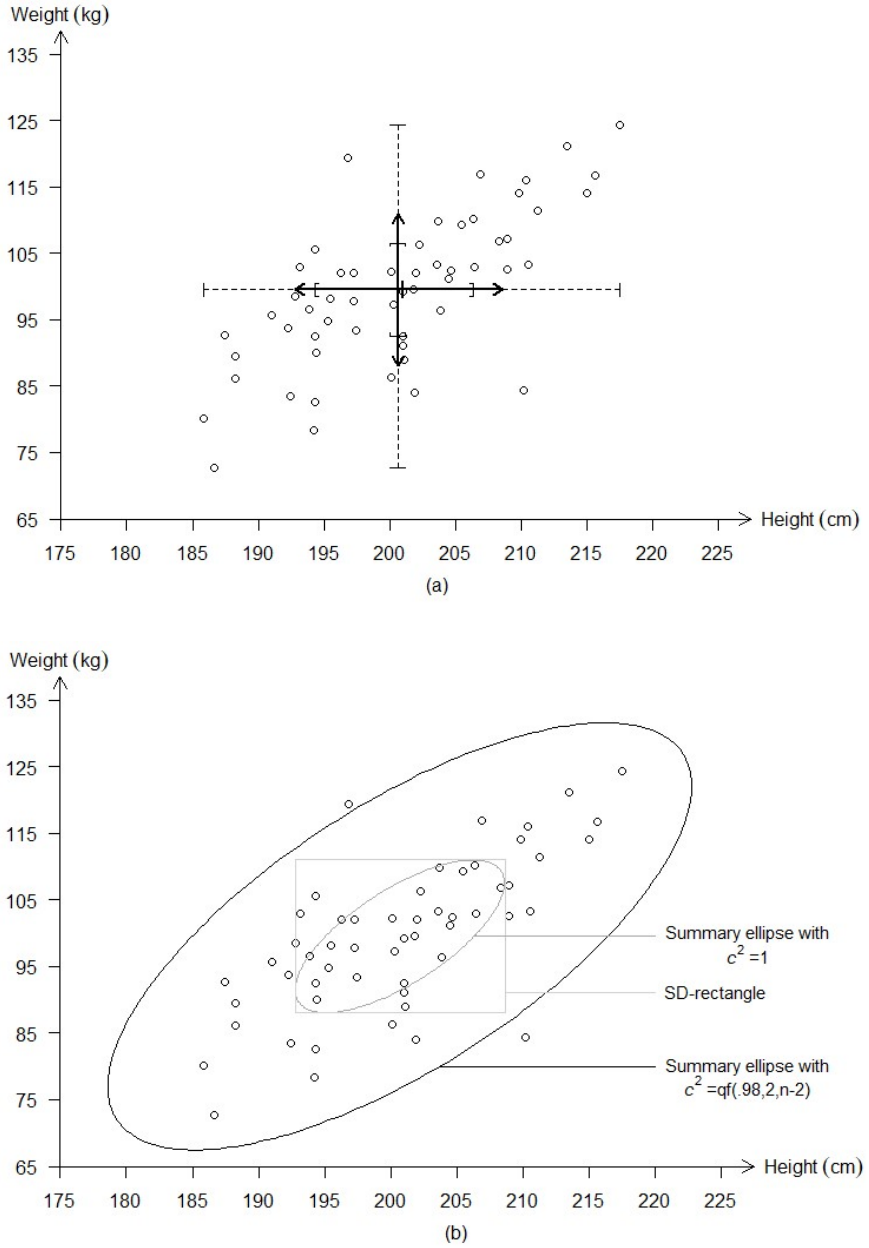

Figure 1. Superimpose on the scatter plot (a) the GI plots of x and y, respectively parallel to the horizontal and the vertical axes, intersecting at $(\bar{x}, \bar{y})$, and (b) two concentric summary ellipses corresponding to $c^2 = 1$ and c^2=qf(.98, 2, n-2) with center $(\bar{x}, \bar{y})$ and orientations that reveal all bivariate summary statistics.

We shall explain how to extract from Figure 1(b) all summary statistics associated with a linear regression model or a bivariate normal model for two quantitative variables — the mean vector, the standard deviations (SD), the correlation coefficient, the two regression lines, and the coefficient of determination.

Section 2 reviews the relationship between the two quantitative variables depicted by a scatter plot. Section 3 explains how the two regression lines suffice to recover most bivariate statistics, except the two SDs. Section 4 explains how to recover all bivariate statistics starting from the two summary ellipses shown in Figure 1(b). Section 5 discusses some implications.

2. RELATIONSHIP IN BIVARIATE DATA

The primary purpose of a scatter plot is to depict the relationship between two quantitative variables with an aim to predict one variable as a function of the other. For example, a person's weight and height are "statistically linear," which according to the regression model means: "Variable y is a linear function of x, plus a random error or noise variable." Oftentimes, the error is assumed to have a normal distribution. Sometimes one may assume that the two variables are distributed as bivariate normal. For example, a student's score in the midterm and the final exams are linearly related. (More knowledgeable students tend to score higher in both exams; less knowledgeable students tend to score lower in both exams.) Using the score in one exam, we can predict (with some error) the score in the other exam. Moreover, the two scores may be jointly bivariate normally distributed, in which case each score is univariate normal, and any linear combination of the two scores is also univariate normal.

The strength and direction of the linear relation between the two quantitative variables is measured by Pearson's product moment correlation coefficient given by

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\,\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{s_{xy}}{\sqrt{s_{xx}}\,\sqrt{s_{yy}}} = \frac{s_{xy}}{s_x\,s_y} \quad (2)$$

Where $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y})$ is called the covariance between x and y; s_{xx} is called the variance of x, its positive square root s_x is called the SD of x, and similarly s_{yy} and s_y are the variance and SD of y. When $0 < r < 1$ (or when $-1 < r < 0$), the linear regression of y on x, according to the least square method, is the line $\hat{y}$ given by

$$\hat{y} - \bar{y} = r\frac{s_y}{s_x}(x - \bar{x}) \quad (3)$$

with a positive (or negative) slope rs_y/s_x; and the linear regression of *x* on *y* is the line

$$\hat{x} - \bar{x} = r\frac{s_x}{s_y}(y - \bar{y}) \quad (4)$$

with a positive (or negative) slope rs_x/s_y. Finally, the coefficient of determination r^2 (the square of the correlation coefficient) tells us what proportion of variation in y-values is attributed to its linear dependence on x-values (and vice versa). Thus, larger the r^2, the better the linear regression model.

Under both the regression model and the bivariate normal model, the error random variable $\epsilon = y - (\alpha + \beta x)$ is normally distributed. However, the two models differ with respect to the associated variance: In the regression model, error variance is a constant (with respect to x); but in the bivariate normal model, the error variance decreases as x moves away from $\bar{x}$ in either direction. Therefore, the larger (solid) summary ellipse of Figure 1(b) captures 98% of data under the bivariate normal distribution, but not under

the regression model. However, it still can be used to flag potential x-, y- or regression outliers.

3. RECOVERING BIVARIATE STATISTICS FROM THE TWO REGRESSION LINES

When both $\hat{y}$- and $\hat{x}$-regression lines are superimposed on a scatter plot their point of intersection is the mean vector $I = (\bar{x}, \bar{y})$. However, the two regression lines together also depict the correlation coefficient r and the coefficient of determination r^2: Combining (3) and (4), we see that r^2 is the ratio of the slope of the less steep regression line $\hat{y}$ to the slope of the steeper regression line $\hat{x}$. Therefore, as shown in Figure 2, if we draw a horizontal line *IH* of any length and then draw its perpendicular through *H* intersecting $\hat{y}$ at *P* and $\hat{x}$ at *Q*, then

$$r^2 = HP/HQ \tag{5}$$

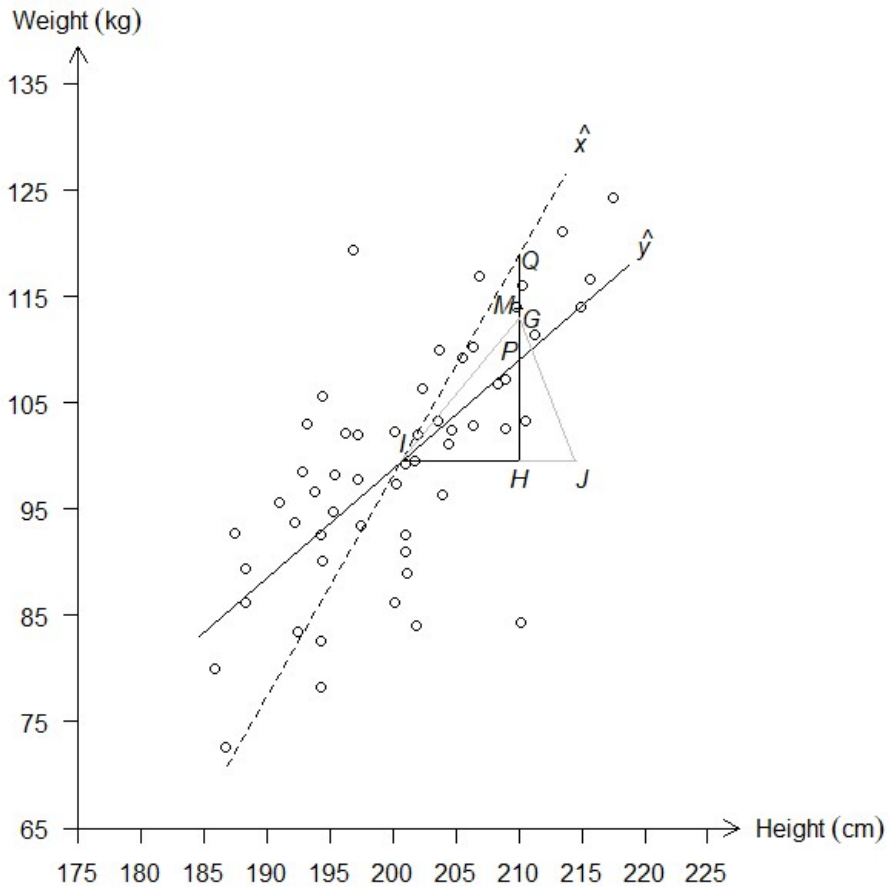

Figure 2. Given the two regression lines, the coefficient of determination r^2 equals the ratio *HP*:*HQ*, where *H* is any point on a horizontal line through *I* and *HQ* is orthogonal to *IH*, intersecting the $\hat{y}$ line at *P* and the $\hat{x}$ line at *Q*. Also, the correlation coefficient is the ratio $HG: HQ = HP/HG$, where G on *HQ* is such that *HG* is the geometric mean of *HP* and *HQ*. The line IG with slope s_y/s_x is called the SD-line.

Having obtained $r^2 = HP/HQ$, we can find the correlation coefficient r as follows: Let *M* be the midpoint of *PQ*. Extend *IH* to a point *J* such that $HJ = MQ = (HQ - HP)/2$.

Next, find G on PQ such that $JG = HM = (HQ + HP)/2$. Then $HG = \sqrt{HP \cdot HQ}$ is the geometric mean between HP and HQ, and

$$r = \sqrt{HP/HQ} = HG/HQ = HP/HG \tag{6}$$

with sign the same as that of the slope of either regression line. Also, the line *IG* with slope s_y/s_x is called the SD-line. Thus, given the two regression lines, we can compute the ratio of the two SDs as the slope of the SD-line *IG*; however, we cannot compute the SDs themselves.

4. RECOVERING BIVARIATE STATISTICS FROM THE TWO SUMMARY ELLIPSES

Suppose that the larger summary ellipse (with a solid boundary), given by (1) with $c^2 = qf(.98, 2, n-2)$, is enclosed by the smallest rectangle whose alternate two sides are horizontal and the other two sides are vertical. Let the four points of tangency between the ellipse and the covering rectangle's right, top, left, and bottom sides be P, R, P', R', respectively. See Figure 3.

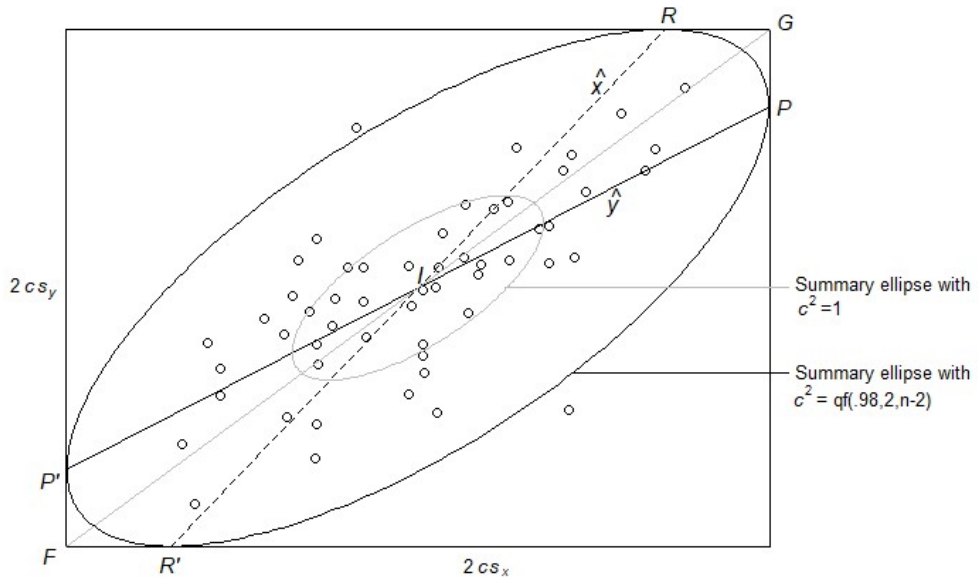

Figure 3. Enclose the larger summary ellipse by a rectangle with sides horizontal and vertical with points of tangency P, R, P', R'. Then the slope of the diagonal FG equals s_y/s_x, PP' is the $\hat{y}$-line, and RR' is the $\hat{x}$-line.

Lemma 1: The (positive) slope of the southwest to northeast diagonal FG of the covering rectangle equals the ratio s_y/s_x, $P'P$ is the $\hat{y}$-line, and $R'R$ is the $\hat{x}$-line.

Proof. Define the standardized variables $z = (x - \bar{x})/(cs_x)$ and $w = (y - \bar{y})/(cs_y)$. Then the ellipses given in (1) can be rewritten as

$$(z, w): \quad z^2 - 2rzw + w^2 = (1 - r^2) \tag{7}$$

Differentiating (7) implicitly with respect to z, we note that $dw/dz = 0$, if and only if $(z, w) = (r, 1)$ or $(-r, -1)$. Hence, the points of tangency R and R' are given by

R: $(z = r, w = 1)$, or equivalently, $(x = \bar{x} + rcs_x, y = \bar{y} + cs_y)$

R': $(z = -r, w = -1)$, or equivalently, $(x = \bar{x} - rcs_x, y = \bar{y} - cs_y)$

Likewise, differentiating (7) implicitly with respect to w, we note that $dz/dw = 0$, if and only if $(z, w) = (1, r)$ or $(-1, -r)$. Hence, the points of tangency P and P' are given by

P: $(z = 1, w = r)$, or equivalently, $(x = \bar{x} + cs_x, y = \bar{y} + rcs_y)$

P': $(z = -1, w = -r)$, or equivalently, $(x = \bar{x} - cs_x, y = \bar{y} - rcs_y)$

Also, the point G is given by $(z, w) = (1, 1)$, or equivalently, $(x = \bar{x} + cs_x, y = \bar{y} + cs_y)$; and the point F is given by $(z, w) = (-1, -1)$, or equivalently, $(x = \bar{x} - cs_x, y = \bar{y} - cs_y)$.

The three claims of the lemma follow. □

In the previous section we have already explained how starting from the two regression lines we obtain the correlation coefficient r and the coefficient of determination r^2. To obtain the two SDs, we enclose the smaller ellipse (with a gray boundary and $c = 1$) by the smallest rectangle whose sides are horizontal and vertical. Then the horizontal sides of the enclosing rectangle are of length $2s_x$ and the vertical sides are of length $2s_y$. Thus, the smaller ellipse yields the exact values of the two SDs; hence, the GI plots.

5. SOME IMPLICATIONS

Let us now explain the uses of summary ellipses for bivariate data coming from one or more subgroups.

The directions of the two axes of the summary ellipse are the same and are given by

$$\text{Major axis:} \quad y - \bar{y} = \frac{\sqrt{\left(s_x^2 - s_y^2\right)^2 + \left(2rs_xs_y\right)^2} - \left(s_x^2 - s_y^2\right)}{2rs_xs_y}(x - \bar{x}) \tag{8}$$

$$\text{Minor axis:} \quad y - \bar{y} = -\frac{\sqrt{\left(s_x^2 - s_y^2\right)^2 + \left(2rs_xs_y\right)^2} + \left(s_x^2 - s_y^2\right)}{2rs_xs_y}(x - \bar{x})$$

Only in the special cases when either $s_x = s_y$ or $r = \pm 1$, the major/minor axis agrees with the SD-line. The major axis is the direction in which the scatter points must be projected to maximize the SD of the projections. Likewise, the minor axis is the direction in which the projected scatter points are most densely packed.

If the (x, y) data belong to two groups coming from two bivariate normal distributions that differ in the mean vectors (but agree on the dispersion matrix; that is, on the SDs and

the correlation), then the summary ellipses for the two groups are location shifts of each other. Their SD lines are parallel, the major (minor) axes are parallel, and the line joining the points of intersection between corresponding ellipses of the two groups is the linear discrimination boundary between the two groups. See Figure 4.

If the (x, y) data belong to two groups coming from two bivariate normal distributions that differ both in the mean vectors and in the dispersion matrices, then the summary ellipses for the two groups differ in both centers and orientations. To understand the discrimination boundary between the two groups, consider two ellipses drawn with the same choice of c^2 varying over $(0, \infty)$. As c^2 increases, the two ellipses will eventually touch each other. Then as c^2 increases further, the two ellipses will intersect each other at two points. The locus of the points of intersection is the quadratic discrimination boundary. We leave to the reader to draw this boundary.

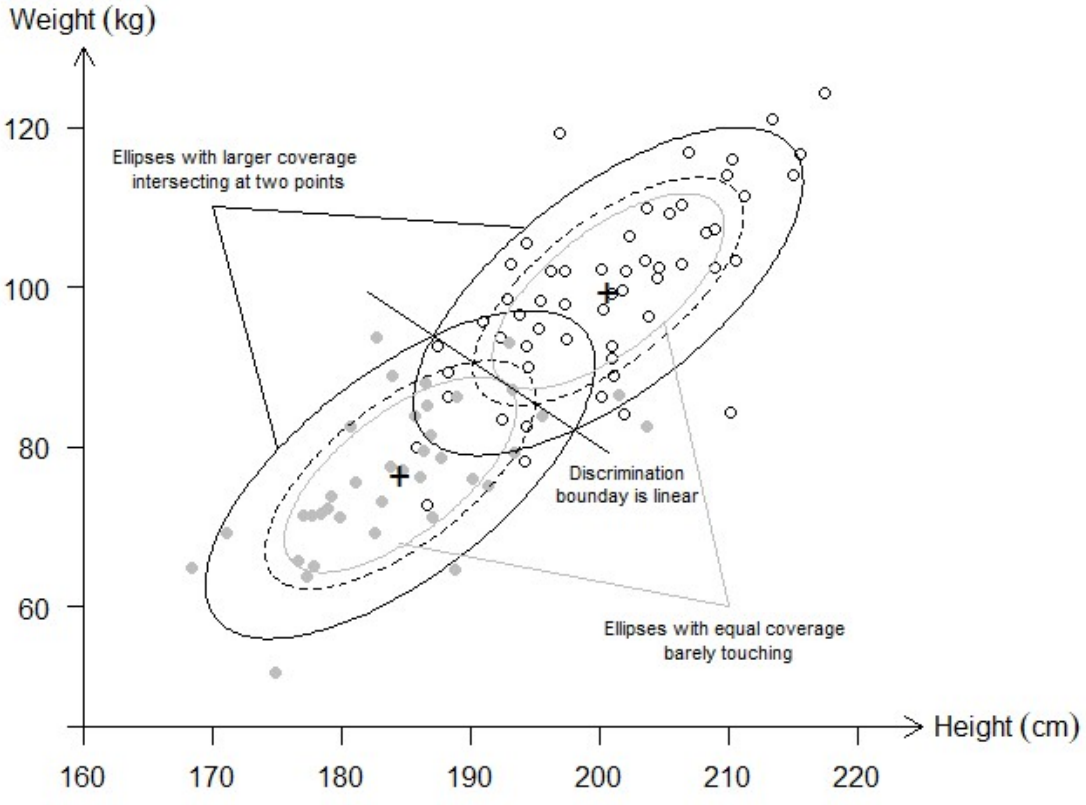

Figure 4. Women's NBA and Men's NBA players' stature and a linear discrimination boundary between the two gender groups.

When the (x, y) data belong to multiple groups, the superimposed 98%-coverage ellipses give an overview of the summary statistics (mean, SD, correlation, etc.) for the different groups as well as a general idea of the degree of overlap among the groups. Figure 5 illustrates it for the celebrated Fisher's Iris data for three types of iris flower in terms of petal length and petal width. One type can be fully separated from the other two which overlap substantially.

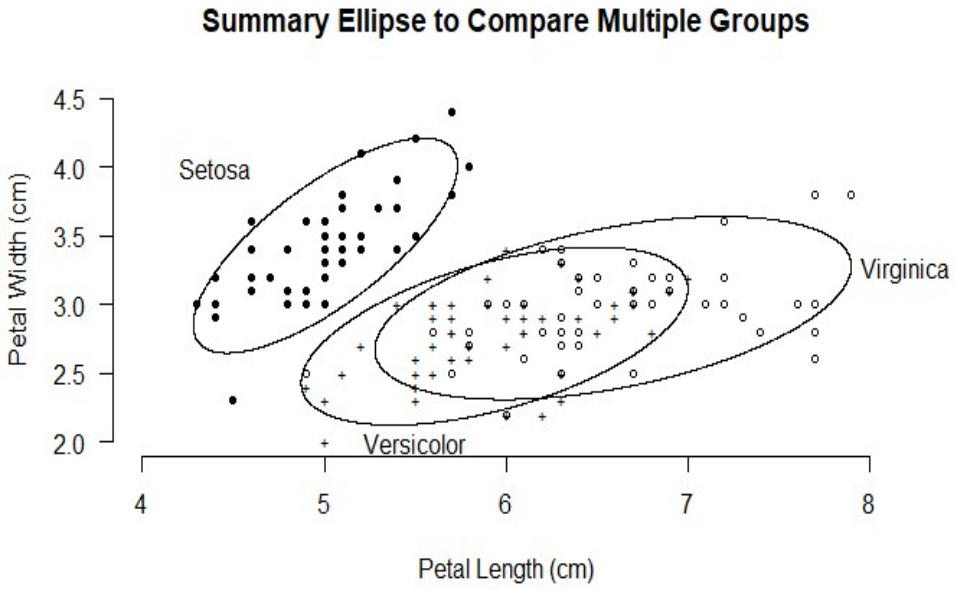

Figure 5. Petal length and petal width of three types of iris flowers: One type can be fully separated from the other two which overlap substantially.

6. R PACKAGES

We have demonstrated that when a scatter plot is superimposed by the summary ellipses, we can extract all summary statistics of a linear regression model or a bivariate normal model. We conclude this paper by drawing the readers' attention to the following paper:

An R package called GIplot (see Phuyal et al. 2021) allows users to draw GI plot for a quantitative variable (against a grouping variable) and to superimpose GI plots of x and y on their scatter plot.

A third R package called summary.ellipse is being developed. It will allow users to draw the summary ellipses for any data set consisting of two quantitative variables.

Our proposal to superimpose on a scatter plot two summary ellipses is only a small modification to the standard scatter plot that require no additional space or text. However, it conveniently summarizes all bivariate statistics in a simple linear regression model with normally distributed error or in a bivariate normal distribution model. These summary ellipses contain a wealth of information, which we have explained in this paper how to extract. Also, any scatter point outside the larger summary ellipse is a potential x-, y-, regression- or bivariate outlier deserving special attention. Furthermore, the degree of overlap among multiple groups can be demonstrated visually.

Acknowledgements

We thank our colleagues and students for playing the game of guessing the coefficient of determination from plain scatter plots, scatter plots superimposed with the two regression lines and scatter plots superimposed with two summary ellipses.

References

[1] Cleveland, W. (1993). Visualizing data. Murray Hill, N.J. Summit, N.J: At & T Bell Laboratories Published by Hobart Press. ISBN 978-0963488404.

[2] Johnson, R. A., and Wichern, D. W. (2007), "Applied multivariate statistical analysis (6th ed)," Pearson Prentice Hall.

[3] Sarkar, J., and Rashid, M. (2016), "Visualizing Mean, Median, Mean Deviation, and Standard Deviation of a Set of Numbers," The American Statistician, **70**:3, 304-312, DOI: 10.1080/00031305.2016.1165734

[4] Sarkar, J., and Rashid, M. (2021), "IVY Plot and Gaussian Interval Plot," Teaching Statistics, 1–6. https://doi.org/10.1111/test.12257

[5] Phuyal, S., Rashid, M., and Sarkar, J. (2021b), "GIplot: The R Gaussian Interval Plot Package," Version 0.1.0. https://CRAN.R-project.org/package=GIplot

[6] R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/.

[7] Spear, M. E. (1952), Charting Statistics, McGraw-Hill, New York, NY.

[8] Spear, M. E. (1969), Practical Charting Techniques, McGraw-Hill, New York, NY.

[9] Wilkinson, L. (1999), "Dot plot," The American Statistician, 53(3), 276–281.