

# Comparison of regression lines based on individual and averaged data

A. R. de LEON<sup>\*,1</sup>, B. WU<sup>1</sup> and J. R. PUNZALAN<sup>2</sup>

<sup>1</sup>*Department of Mathematics and Statistics  
University of Calgary  
Calgary, Alberta T2N1N4, Canada*

<sup>2</sup>*School of Statistics  
University of the Philippines  
Diliman, Quezon City, Philippines*

## SUMMARY

Regression analysis sometimes uses averages that combine information from many individuals. Because correlations based on averages over many individuals are usually stronger than corresponding correlations based on data for individuals, results of a regression analysis on averaged data may be misleading. This note compares regression lines based on averaged and individual data and studies the relationship between their least-squares estimates.

**Keywords:**  $F$ -ratio; Lack-of-fit; Least squares estimates; Weighted least squares.

## 1. Introduction

Because averages vary less than data on individuals do, scatterplots of averages over groups tend to show less variability than we would see if we measured the same variable on individuals. Correlations based on averages over many individuals, or so-called ecological correlations (Freedman, Pisani and Purves, 2007, p. 148], are generally stronger than corresponding correlations based on individual data. Plotting the average heights of young children against their age in months, for example, may show a strong positive correlation between height and age, suggesting that a regression line can be used to adequately predict height from age in months. However, children at the same age vary greatly in height, as indicated by the greater scatter of points on a scatterplot of height against age for individual children. A regression line of children's average heights on their ages used for predicting an individual child's height can thus give a misleadingly good prediction. Elementary statistics textbooks by Moore, McCabe and Craig (2009, p. 135) and De Veaux, Velleman and Bock (2008, p. 237), among others, have discussed this phenomenon, albeit briefly, in the context of simple linear regression. They caution that regression lines based on averaged data should be

---

\*email: adeleon@math.ucalgary.ca

applied carefully in practice, as they tend to give a more optimistic picture of relationships between dependent and independent variables than what actually exist—an example of a common mistake in the interpretation of statistical data known as ecological fallacy, which assumes that what holds true for the group also holds true for individuals.

In this article, we attempt to thresh out this phenomenon concerning regression lines calculated from averaged and individual data. In the process, specific relationships between relevant quantities based on averaged and individual data are established.

## 2. Results

Suppose for each level  $x_i$  of independent variable  $x$ ,  $n_i$  values  $y_{i1}, \dots, y_{in_i}$  of dependent variable  $y$  are observed,  $i = 1, \dots, K$ , where  $N = \sum_{i=1}^K n_i$ . These values of  $y$  are referred to as repeats or replicates and are used to check for lack-of-fit of a postulated regression line (see, e.g., Montgomery, Peck and Vining, 2006, p. 145). The regression model is

$$y_{ij} = \alpha + \beta x_i + \delta_{ij}, \quad (1)$$

where the  $\delta_{ij}$ s are independent errors with common mean 0 and common variance  $\sigma^2$ . If only averaged data  $\bar{y}_i = \sum_j y_{ij}/n_i$ ,  $i = 1, \dots, K$ , are provided, then model (1) yields the average-data regression model

$$\bar{y}_i = \alpha + \beta x_i + \bar{\delta}_i, \quad (2)$$

where  $\bar{\delta}_1 = \sum_{j=1}^{n_1} \delta_{1j}/n_1, \dots, \bar{\delta}_K = \sum_{j=1}^{n_K} \delta_{Kj}/n_K$ , are independent with common mean 0 and respective variances  $\sigma^2/n_1, \dots, \sigma^2/n_K$ . Denoting the least-squares estimate of  $\beta$  based on individual data as  $\hat{\beta}_o$  and that based on averaged data as  $\hat{\beta}_a$ , note that  $\hat{\beta}_a$ , while still unbiased for the regression coefficients, is not optimal anymore; Montgomery, Peck and Vining (2006, p. 180) discuss this issue in the context of weighted least squares estimation. We have the following result.

**Result 1** Define  $\bar{x}_o = \sum_{i=1}^K n_i x_i / N$ ,  $\bar{y}_o = \sum_{i=1}^K n_i \bar{y}_i / N$ ,  $\bar{x}_a = \sum_{i=1}^K x_i / K$ ,  $\bar{y}_a = \sum_{i=1}^K \bar{y}_i / K$ ,

$$s_{x_o}^2 = \frac{1}{N-1} \sum_{i=1}^K n_i (x_i - \bar{x}_o)^2, \quad s_{y_o}^2 = \frac{1}{N-1} \sum_{i=1}^K \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_o)^2$$

$$s_{x_a}^2 = \frac{1}{K-1} \sum_{i=1}^K (x_i - \bar{x}_a)^2, \quad s_{y_a}^2 = \frac{1}{K-1} \sum_{i=1}^K (\bar{y}_i - \bar{y}_a)^2$$

and  $R_{xy} = \sum_{i=1}^K n_i (x_i - \bar{x}_a)(\bar{y}_i - \bar{y}_a) / \sum_{i=1}^K (x_i - \bar{x}_a)(\bar{y}_i - \bar{y}_a)$ . Then

(i) the following relationship holds between individual- and average-data correlations  $r_o$  and  $r_a$ :

$$r_o = \frac{R_{xy}(K-1)s_{x_a}s_{y_a}}{(N-1)s_{x_o}s_{y_o}} \left\{ r_a - \frac{N(\bar{x}_a - \bar{x}_o)(\bar{y}_a - \bar{y}_o)}{R_{xy}(K-1)s_{x_a}s_{y_a}} \right\}; \quad (3)$$

(ii) the relationship between  $\hat{\beta}_o$  and  $\hat{\beta}_a$  is given by

$$\hat{\beta}_o = \frac{R_{xy}(K-1)s_{x_a}^2}{(N-1)s_{x_o}^2} \left\{ \hat{\beta}_a - \frac{N(\bar{x}_a - \bar{x}_o)(\bar{y}_a - \bar{y}_o)}{R_{xy}(K-1)s_{x_a}^2} \right\}. \quad (4)$$

**Proof:** Noting that

$$r_o = \sum_{i=1}^K \sum_{j=1}^{n_i} \frac{(x_i - \bar{x}_o)(y_{ij} - \bar{y}_o)}{(N-1)s_{x_o}s_{y_o}}, \quad r_a = \sum_{i=1}^K \frac{(x_i - \bar{x}_a)(\bar{y}_i - \bar{y}_a)}{(K-1)s_{x_a}s_{y_a}}, \quad (5)$$

(3) is straightforward; (4) follows easily from  $\hat{\beta}_o = r_o s_{y_o}/s_{x_o}$  and  $\hat{\beta}_a = r_a s_{y_a}/s_{x_a}$ .  $\square$

It is clear from (4) that  $\hat{\beta}_o$  based on individual data can differ enormously from  $\hat{\beta}_a$  based on averaged data. The following example illustrates that it is possible for  $\hat{\beta}_o > \hat{\beta}_a$ .

**Example 2** Consider the following data from Example 4.8 of Montgomery, Peck and Vining (2006, p. 348):

|     |       |      |       |       |       |       |       |       |       |
|-----|-------|------|-------|-------|-------|-------|-------|-------|-------|
| $x$ | 1     | 1    | 2     | 3.3   | 3.3   | 4     | 4     | 4     | 4.7   |
| $y$ | 10.84 | 9.3  | 16.35 | 22.88 | 24.35 | 24.56 | 25.86 | 29.16 | 24.59 |
| $x$ | 5     | 5.6  | 5.6   | 5.6   | 6     | 6     | 6.5   | 6.9   |       |
| $y$ | 22.25 | 25.9 | 27.2  | 25.61 | 25.45 | 26.56 | 21.03 | 21.46 |       |

From (3) and (4), we get  $r_o = 0.7 > r_a = 0.66$  and  $\hat{\beta}_o = 2.13 > \hat{\beta}_a = 1.77$ . The next example shows that  $\hat{\beta}_o < \hat{\beta}_a$  is also possible.

**Example 3** Consider the following data from Table 2.1 of Draper and Smith (1998, p. 51):

|     |     |     |     |     |     |     |     |     |     |     |     |     |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| $x$ | 1.3 | 1.3 | 2   | 2   | 2.7 | 3.3 | 3.3 | 3.7 | 3.7 | 4   | 4   | 4   |
| $y$ | 2.3 | 1.8 | 2.8 | 1.5 | 2.2 | 3.8 | 1.8 | 3.7 | 1.7 | 2.8 | 2.8 | 2.2 |
| $x$ | 4.7 | 4.7 | 5   | 5.3 | 5.3 | 5.3 | 5.7 | 6   | 6   | 6.3 | 6.7 |     |
| $y$ | 3.2 | 1.9 | 1.8 | 3.5 | 2.8 | 2.1 | 3.4 | 3.2 | 3   | 3   | 5.9 |     |

For these data, we easily obtain  $r_o = 0.51 < r_a = 0.63$  and  $\hat{\beta}_o = 0.316 < \hat{\beta}_a = 0.382$ . Figure 1 displays individual- and average-data regression lines for Examples 2 and 3.

When do the two regression lines coincide? The answer is provided by the following result, which follows from Result 1.

**Result 4** Consider the balanced case of equal numbers of repeats at levels  $x_1, \dots, x_K$  of  $x$  (i.e.,  $n_1 = \dots = n_K = n$ ), so that  $\bar{x}_o = \bar{x}_a = \bar{x}$  and  $\bar{y}_o = \bar{y}_a = \bar{y}$ . Then

(i) the following relationship holds between  $r_o$  and  $r_a$ :

$$r_o = \frac{r_a}{\sqrt{1 + \left(\frac{N-K}{K-1}\right) F_y^{-1}}}, \quad (6)$$

where  $F_y = \{(N-K)n \sum_{i=1}^K (\bar{y}_i - \bar{y})^2\} / \{(K-1) \sum_{i=1}^K \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2\}$  is the balanced one-way ANOVA  $F$ -ratio for testing equality of means for  $y$  grouped according to levels  $x_1, \dots, x_K$  of  $x$ ;

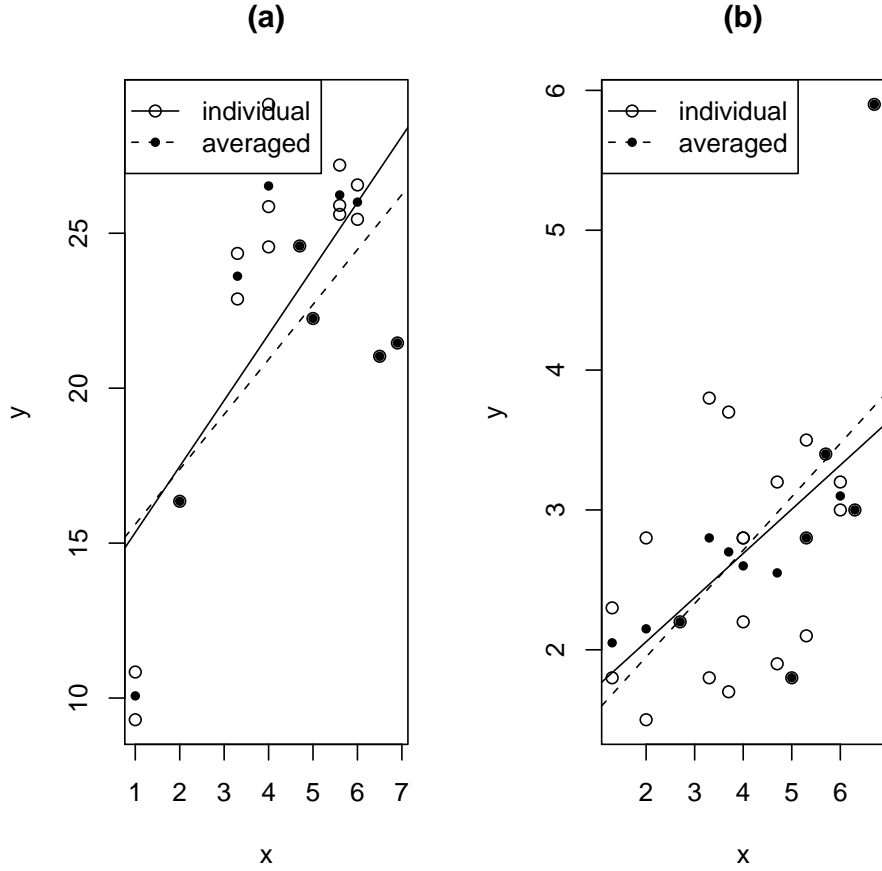


Figure 1: Comparison of regression lines from averaged (dashed) and individual (solid) data from (a) Example 2 and (b) Example 3. Averaged data are denoted by  $\bullet$ .

(ii) *the two regression lines based on averaged and individual data coincide, i.e.,  $\hat{\beta}_o = \hat{\beta}_a$ , hence,  $\hat{\alpha}_o = \hat{\alpha}_a$ .*

Note that  $F_y$  is also the  $F$ -ratio for testing lack of fit, where  $SS_{pe} = \sum_{i=1}^K \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2$  and  $SS_{lof} = n \sum_{i=1}^K (\bar{y}_i - \bar{y})^2$  are respectively the pure-error and lack-of-fit sums of squares. It is clear from (3) that while  $r_a$  can be less than, equal to, or greater than  $r_o$ , (6) indicates that the latter can never be greater than the former in the balanced case. This confirms the common notion that averaged data, having less spread than individual data, typically exhibit stronger correlations than do individual data.

Observe that  $r_a \rightarrow r_o$  as  $F_y \rightarrow \infty$  in (6); that is,  $r_a$  will be close to  $r_o$  for large  $F_y$ , in which case  $SS_{pe}$  is considerably smaller than  $SS_{lof}$ . This is the case whenever  $y_{ij} \approx \bar{y}_i$  for all  $i, j$ , i.e., the replicates are approximately equal, which happens when the variances of replicates are uniformly very small.

Conditions abound on the equality of ordinary and weighted least-squares estimates (e.g., Puntanen and Styan, 1989). Result 4 indicates that a necessary and sufficient condition for  $\widehat{\beta}_o = \widehat{\beta}_a$  is that  $n_1 = \cdots = n_K = n$ . Noting that this implies  $\bar{d}_i$ s are independent and have equal variances, this agrees with McElroy's (1967) necessary and sufficient condition.

It is interesting that while  $r_a \geq r_o$  in the balanced case, the two regression lines based on averaged and individual data coincide. This may be explained by the fact that correlation incorporates the spread in the linear relationship in both the  $y$ - and  $x$ -directions while the regression line of  $y$  on  $x$  explains only the variation in  $y$  at a fixed  $x$ .

## Acknowledgment

This work was partially supported by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

## References

- De Veaux, R. D., Velleman, P. F., and Bock, D. E. (2008), *Intro Stats*, 3rd ed., Pearson Addison-Wesley.
- Draper, N. R. and Smith, H. (1998), *Applied Regression Analysis*, 3rd ed., Wiley.
- Freedman, D., Pisani, R., and Purves, R. (2007), *Statistics*, 4th ed., W.W. Norton.
- McElroy, F. W. (1967), "A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased," *Journal of the American Statistical Association*, **62**, 1302–1304.
- Montgomery, D. C., Peck, E. A., and Vining, G. G. (2006), *Introduction to Linear Regression Analysis*, 4th ed., Wiley.
- Moore, D. S., McCabe, G. P., and Craig, B. A. (2009), *Introduction to the Practice of Statistics*, 6th ed., W. H. Freeman & Co.
- Puntanen, S. and Styan, G. P. H. (1989), "The equality of the ordinary least-squares estimator and the best linear unbiased estimator," *The American Statistician*, **43**, 153–164.