fortune Journals

Research Article

JOURNAL OF BIOINFORMATICS AND SYSTEMS BIOLOGY

ISSN: 2688-5107

Linear Regression of Sampling Distributions of the Mean

David J Torres*, Ana Vasilic, Jose Pacheco

Abstract

We show that the simple and multiple linear regression coefficients and the coefficient of determination R^2 computed from sampling distributions of the mean (with or without replacement) are equal to the regression coefficients and coefficient of determination computed with individual data. Moreover, the standard error of estimate is reduced by the square root of the group size for sampling distributions of the mean. The result has applications when formulating a distance measure between two genes in a hierarchical clustering algorithm. We show that the Pearson *R* coefficient can measure how differential expression in one gene correlates with differential expression in a second gene.

Keywords: Linear regression; Pearson R; Sampling distributions of the mean.

Introduction

Linear regression coefficients and the Pearson R correlation have long been used to quantify the relationship between dependent and independent variables [1]. However, the "ecological fallacy" has shown that linear regression and correlation coefficients based on group averages cannot be used to estimate linear regression and correlation coefficients based on individual scores [2, 3].

It may not be well known that if all possible groups are considered, in the case of sampling distributions of the mean, the Pearson R coefficient computed from the group averages is equal to the Pearson R coefficient computed from the original individual scores for one independent variable [4, 5]. We extend this result and show that the linear regression coefficients (for simple and multiple regression) and the coefficient of determination R^2 computed from sampling distributions of the mean (with or without replacement) are the same as the coefficient of determination and linear regression coefficients computed with the original individual data. The sampling distributions of the mean can also be constructed using differences between two groups of different size. The result has implications for hierarchical clustering of genes. Specifically, the Pearson R coefficient can be used to measure how differential expression in one gene correlates with differential expression in a second gene.

The standard error of estimate is a measure of accuracy for the surface of regression [6]. Using the coefficient of determination, we show that the standard error of estimate is reduced by the square root of the group size for sampling distributions of the mean.

In Section 1, we recall and reformulate the system of equations that are solved to determine the linear regression coefficients for individual scores. In Section 2, we prove the assertion that the same system of equations needs to be solved for sampling

Affiliation:

Department of Mathematics and Physical Science, Northern New Mexico College, Española, New Mexico, USA

***Corresponding author:**
David J Torres. Department of Mathematics and Physical Science, Northern New Mexico College, Española, New Mexico, USA

Citation: David J Torres, Ana Vasilic, Jose Pacheco. Department of Mathematics and Physical Science, Northern New Mexico College, Española, New Mexico, USA. Journal of Bioinformatics and Systems Biology. 7 (2024): 63-80.

Received: December 12, 2023
Accepted: December 19, 2023
Published: March 04 2024

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

distributions of the mean with and without replacement or differences between two groups of sampling distributions. In Section 3, we show that the coefficient of determination is the same whether it is calculated using individual scores or all possible group averages from sampling distributions. Section 4 shows that the standard error of estimate is reduced by the square root of the group size for sampling distributions of the mean. Section 5 performs numerical simulations to illustrate these principles. Section 6 applies these results and shows that the Pearson R coefficient can be used to measure how differential expression in one gene correlates with differential expression in a second gene when the z-statistic is used.

1 Computing regression coefficients

Multiple regression requires one to compute the coefficients $\{\beta_j^*, j = 0, ..., K\}$ that minimize the sum of squares

$$\min_{\beta_j^*} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{K} \beta_j^* x_i^{(j)} - \beta_0^* \right)^2 \tag{1}$$

where $x^{(1)}$, $x^{(2)}$, ..., $x^{(K)}$ represent K different independent variables, y is the dependent variable, and the i^{th} realization of variables $x^{(j)}$ and y are $x_i^{(j)}$ and y_i, respectively, $1 \leq i \leq N$. Note that for simple linear regression, $K = 1$. We recast the sum (1) in the form

$$\sum_{i=1}^{N} \left((y_i - \bar{y}) - \sum_{j=1}^{K} \beta_j (x_i^{(j)} - \bar{x}^{(j)}) - \beta_0 \right)^2 \tag{2}$$

where

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}, \quad \bar{x}^{(j)} = \frac{\sum_{i=1}^{N} x_i^{(j)}}{N}, \; j = 1, 2, ...K \tag{3}$$

and the coefficients are related by

$$\beta_0 = \beta_0^* - \bar{y} + \sum_{j=1}^{K} \beta_j^* \bar{x}^{(j)}, \tag{4}$$

$$\beta_j = \beta_j^*, \qquad j = 1, 2, ...K. \tag{5}$$

To solve for β_0, we set the partial derivative of (2) with respect to β_0 to zero which yields

$$\sum_{i=1}^{N} (y_i - \bar{y}) - \sum_{j=1}^{K} \beta_j \left(\sum_{i=1}^{N} (x_i^{(j)} - \bar{x}^{(j)}) \right) = N\beta_0. \tag{6}$$

However

$$\sum_{i=1}^{N} (y_i - \bar{y}) = 0, \quad \sum_{i=1}^{N} (x_i^{(j)} - \bar{x}^{(j)}) = 0 \;\; \forall j, \;\; 1 \leq j \leq K \tag{7}$$

which implies that $\beta_0 = 0$. Thus one can redefine the problem of computing multiple regression coefficients (1) to be selection of the coefficients $\{\beta_j, j = 1, ..., K\}$ that minimizes the sum of squares

$$\min_{\beta_j} \sum_{i=1}^{N} \left((y_i - \bar{y}) - \sum_{j=1}^{K} \beta_j (x_i^{(j)} - \bar{x}^{(j)}) \right)^2. \tag{8}$$

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

In the matrix approach to minimizing the sum of squares which can be derived by setting the partial derivatives $\frac{\partial}{\partial \beta_j}$ of (8) to zero, the system of equations in (8) is written in matrix form [8]

$$\mathbf{y} - \bar{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} \tag{9}$$

where

$$\mathbf{y} - \bar{\mathbf{y}} = \begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ y_3 - \bar{y} \\ \vdots \\ y_N - \bar{y} \end{pmatrix}_{N\times 1}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{pmatrix}_{K\times 1},$$

and $\mathbf{X}$ is a N by K matrix whose entries are:

$$\mathbf{X} = \begin{pmatrix} x_1^{(1)} - \bar{x}^{(1)} & x_1^{(2)} - \bar{x}^{(2)} & \cdots & x_1^{(K)} - \bar{x}^{(K)} \\ x_2^{(1)} - \bar{x}^{(1)} & x_2^{(2)} - \bar{x}^{(2)} & \cdots & x_2^{(K)} - \bar{x}^{(K)} \\ \vdots & \vdots & \vdots & \vdots \\ x_N^{(1)} - \bar{x}^{(1)} & x_N^{(2)} - \bar{x}^{(2)} & \cdots & x_N^{(K)} - \bar{x}^{(K)} \end{pmatrix}_{N\times K}.$$

One can solve for the multiple regression coefficients in the vector $\boldsymbol{\beta}$ by left multiplying (9) by the transpose $\mathbf{X}^T$ and solving the linear system

$$\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} = \mathbf{X}^T\left(\mathbf{y} - \bar{\mathbf{y}}\right). \tag{10}$$

The elements in the square K by K matrix $\mathbf{X}^T\mathbf{X}$ will be sums of the form

$$\left(\mathbf{X}^T\mathbf{X}\right)_{ij} = \sum_{p=1}^{N}(x_p^{(i)} - \bar{x}^{(i)})(x_p^{(j)} - \bar{x}^{(j)}). \tag{11}$$

Similarly the entries in the K by 1 vector $\mathbf{b} \equiv \mathbf{X}^T\left(\mathbf{y} - \bar{\mathbf{y}}\right)$ will be sums of the form

$$b_i = \sum_{p=1}^{N}(x_p^{(i)} - \bar{x}^{(i)})(y_p - \bar{y}), \quad i = 1, 2, ..., K. \tag{12}$$

It should be noted that for each pair of fixed indices i and j, the sum in either expression (11) or (12) can be represented using a sum of the form

$$S = \sum_{p=1}^{N}\left(u_p - \bar{u}\right)\left(v_p - \bar{v}\right). \tag{13}$$

In the following section, we show that if the variables $x^{(1)}, x^{(2)}, ..., x^{(k)}$, and y are replaced with all the elements from the sampling distributions of the mean, the system (14) is obtained

$$\alpha\left(\mathbf{X}^T\mathbf{X}\right)\boldsymbol{\beta} = \alpha\mathbf{X}^T\left(\mathbf{y} - \bar{\mathbf{y}}\right) \tag{14}$$

for some constant α. Moreover, we obtain a closed form for the constant α. If m is the group size and we account for order, the size of the matrix $\mathbf{X}$ will be $N^m \times K$ for selections with replacement and $\frac{(N-m+1)!}{(N-m)!} \times K$ for selections without replacement. However, the resulting system (14) will still be a $K \times K$ system. Since the system (14) is equivalent to the system (10), the regression coefficients for the sampling distributions of the mean will be the same as the regression coefficients computed from the original data according to (5) for $1 \leq j \leq K$. The equivalence of β_0^* follows from

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

$\beta_0 = 0$, equation (4), and the fact that the means of the original data ($\bar{y}$ and $\bar{x}^{(j)}$) are the same as the means computed using all the elements from the sampling distributions of the mean (with or without replacement). If we assume that there are N_p elements in the sampling distribution, this can be stated mathematically as

$$\frac{\sum_{\mathcal{P}} \left(\frac{w_{p_1} + w_{p_2} + \ldots + w_{p_m}}{m} \right)}{N_p} = \bar{w} \tag{15}$$

or

$$\sum_{\mathcal{P}} \left(\frac{w_{p_1} + w_{p_2} + \ldots + w_{p_m}}{m} - \bar{w} \right) = 0 \tag{16}$$

where w can represent y or $x^{(j)}$ and $\bar{w} = \left(\sum_{i=1}^{N} w_i \right) / N$. The sum $\sum_{\mathcal{P}}$ is a sum over all possible index values in the sampling distribution.

2 Regression with averages

Let us create elements from the sampling distribution of the mean using elements chosen from the groups

$$U = \{u_p \mid p = 1, 2, ..., N\}, \quad V = \{v_p \mid p = 1, 2, ..., N\} \tag{17}$$

by averaging all possible groups of size m_1 and size m_2

$$\left(\frac{u_{p_1} + u_{p_2} + \ldots + u_{p_{m_1}}}{m_1} \right), \quad \left(\frac{v_{p_1} + v_{p_2} + \ldots + v_{p_{m_2}}}{m_2} \right) \tag{18}$$

chosen from the sets U and V. We assume without loss of generality that $m_1 \geq m_2$. The first m_2 choices are paired

$$\{(u_{p_1}, v_{p_1}), (u_{p_2}, v_{p_2}) \ldots, (u_{p_{m_2}}, v_{p_{m_2}})\} \subset \{(u_1, v_1), (u_2, v_2) \ldots, (u_N, v_N)\},$$

while the remaining choices remain unpaired

$$u_i \in \{u_1, u_2, \ldots u_N\}, \quad i = m_2 + 1, \ldots m_1.$$

If the selections are done without replacement, $p_r \neq p_s$ if $r \neq s$. However, if the selections are formed with replacement, p_r can equal p_s.

Let us now replace u_p and v_p in (13) with all possible averages of m_1 and m_2 elements as shown in (18),

$$\begin{aligned} S_{\bar{u},\bar{v}} := \sum_{\mathcal{P}} & \left[\left(\frac{u_{p_1} + u_{p_2} + \ldots + u_{p_{m_1}}}{m_1} \right) - \bar{u} \right] \\ & \cdot \left[\left(\frac{v_{p_1} + v_{p_2} + \ldots + v_{p_{m_2}}}{m_2} \right) - \bar{v} \right], \end{aligned} \tag{19}$$

where

$$\sum_{\mathcal{P}} = \sum_{p_1=1}^{N} \sum_{p_2=1}^{N} \cdots \sum_{p_{m_1}=1}^{N} \tag{20}$$

for selections with replacement and

$$\sum_{\mathcal{P}} = \sum_{p_1=1}^{N} \sum_{\substack{p_2=1 \\ \mathcal{E}_2^p}}^{N} \cdots \sum_{\substack{p_{m_1}=1 \\ \mathcal{E}_{m_1}^p}}^{N} \tag{21}$$

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

for selections without replacement where $\mathcal{E}_l^p$ is used to denote the exclusion of previously chosen indices

$$\mathcal{E}_l^p := \{p_l \mid p_l \neq p_k, k = 1, 2, ..., l-1\}. \tag{22}$$

The means of the original scores $\bar{u} = \left(\sum_{i=1}^N u_i\right)/N$, $\bar{v} = \left(\sum_{i=1}^N v_i\right)/N$ are used in (19) since they are equal to the means of the sampling distributions by (15). Note that order matters in the way the sums are written in (20-21). For example, (u_1, u_2, u_3) is considered a different choice then (u_3, u_2, u_1). Disregarding order would lead to $m_1!$ fewer terms in (21). However the same system (14) would be generated if order was not considered for sampling distributions without replacement.

Factoring out $\frac{1}{m_1 m_2}$ from (19) yields

$$S_{\bar{u},\bar{v}} := \frac{1}{m_1 m_2} \sum_{\mathcal{P}} \left[(u_{p_1} - \bar{u}) + (u_{p_2} - \bar{u}) + ... + (u_{p_{m_1}} - \bar{u})\right] \cdot \left[(v_{p_1} - \bar{v}) + (v_{p_2} - \bar{v}) + ... + (v_{p_{m_2}} - \bar{v})\right]. \tag{23}$$

Sections 2.1 and 2.2 show that $S_{\bar{u},\bar{v}}$ will be equal to a factor α times S as defined by (13) for sampling distributions with and without replacement respectively. Section 2.3 generalizes these results to differences of two groups of sampling distibutions. In all cases, the elements of the matrix $\left(\mathbf{X}^T\mathbf{X}\right)$ and the vector $\mathbf{X}^T\left(\mathbf{y} - \bar{\mathbf{y}}\right)$ will be multiplied by the same factor α when elements from the sampling distributions of the mean are used.

2.1 Sampling distributions with replacement

Start with $S_{\bar{u},\bar{v}}$ as defined by (23). Since we are considering sampling distribution with replacement, the values chosen for summation indices p_i do not need to be different. We will show that

$$S_{\bar{u},\bar{v}} = \frac{N^{m_1-1}}{m_1} \sum_{p=1}^{N} (u_p - \bar{u})(v_p - \bar{v}). \tag{24}$$

If we distribute the sums inside the parentheses in (23), two types of terms are formed. The first type of term takes the form

$$(u_{p_i} - \bar{u})(v_{p_i} - \bar{v}), \quad i = 1, ..., m_2 \tag{25}$$

where the same summation index p_i is used for u_{p_i} and v_{p_i}. The second type of term takes the form

$$(u_{p_i} - \bar{u})(v_{p_j} - \bar{v}), \quad i \neq j, \quad i = 1, ..., m_1, \ j = 1, ..., m_2 \tag{26}$$

where different summation indices, p_i and p_j are used.

All the terms of the form shown in (26) are zero since when the sums $\sum_{p_i=1}^N$ and $\sum_{p_j=1}^N$ from $\sum_{\mathcal{P}}$ are moved to apply directly to $(u_{p_i} - \bar{u})(v_{p_j} - \bar{v})$, each term can be summed independently

$$\sum_{p_i=1}^{N} \sum_{p_j=1}^{N} (u_{p_i} - \bar{u})(v_{p_j} - \bar{v}) = \sum_{p_i=1}^{N} (u_{p_i} - \bar{u}) \sum_{p_j=1}^{N} (v_{p_j} - \bar{v}), \quad i \neq j.$$

However

$$\sum_{p_i=1}^{N} (u_{p_i} - \bar{u}) = 0, \quad \sum_{p_j=1}^{N} (v_{p_j} - \bar{v}) = 0 \tag{27}$$

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

as noted by equation (7). Thus we must only consider terms of the form (25). The sum (20) acting on $(u_{p_i} - \bar{u})(v_{p_i} - \bar{v})$ can be rearranged as

$$\sum_{\mathcal{P}} (u_{p_i} - \bar{u})(v_{p_i} - \bar{v}) = \sum_{p_1=1}^{N} \sum_{p_2=1}^{N} \cdots \sum_{p_{i-1}=1}^{N} \sum_{p_{i+1}=1}^{N} \cdots \sum_{p_{m_1}=1}^{N} \sum_{p_i=1}^{N} (u_{p_i} - \bar{u})(v_{p_i} - \bar{v})$$

and simplified to

$$\sum_{\mathcal{P}} (u_{p_i} - \bar{u})(v_{p_i} - \bar{v}) = N^{m_1-1} \sum_{p_i=1}^{N} (u_{p_i} - \bar{u})(v_{p_i} - \bar{v}), \quad i = 1, ..., m_2, \tag{28}$$

since each sum $\sum_{p_1=1}^{N} \sum_{p_2=1}^{N} \cdots \sum_{p_{i-1}=1}^{N} \sum_{p_{i+1}=1}^{N} \cdots \sum_{p_{m_1}=1}^{N}$ contributes a factor of N. Multiplying the right side of (28) by m_2 to ensure all summation indices $p_i, i = 1, ..., m_2$ are accounted for and multiplying by the factor $\frac{1}{m_1 m_2}$ present in (23) yields (24). One can also derive (24) using random variables and expected values.

Since $S_{\bar{u},\bar{v}}$ is a multiple of S defined by (13), the system of equations (14) will be formed where $\alpha = \frac{N^{m-1}}{m}$ when we set $m = m_1 = m_2$. Thus the multiple regression coefficients $\boldsymbol{\beta}$ computed from sampling distributions of the mean with replacement will be equal to the multiple regression coefficients computed from the original scores.

2.2 Sampling distribution without replacement

We will show that

$$S_{\bar{u},\bar{v}} = \frac{(N-2)!}{m_1(N-m_1-1)!} \sum_{p=1}^{N} (u_p - \bar{u})(v_p - \bar{v}) \tag{29}$$

for sampling distributions created without replacement. If we distribute the sums inside the parentheses of (23), we again distinguish between two types of terms: terms of the form $(u_{p_i} - \bar{u})(v_{p_i} - \bar{v}),\ i = 1, ..., m_2$ and terms of the form $(u_{p_i} - \bar{u})(v_{p_j} - \bar{v})$, $i \neq j,\ \ i = 1, ..., m_1,\ j = 1, ..., m_2$.

Choose a summation index p_i. The sum (21) applied to $(u_{p_i} - \bar{u})(v_{p_i} - \bar{v})$ can be written as

$$\sum_{\mathcal{P}} (u_{p_i} - \bar{u})(v_{p_i} - \bar{v}) =$$
$$\sum_{p_i=1}^{N} \sum_{\substack{p_1=1 \\ \mathcal{E}^p_{1,i}}}^{N} \sum_{\substack{p_2=1 \\ \mathcal{E}^p_{2,i}}}^{N} \cdots \sum_{\substack{p_{i-1}=1 \\ \mathcal{E}^p_{i-1,i}}}^{N} \sum_{\substack{p_{i+1}=1 \\ \mathcal{E}^p_{i+1,i}}}^{N} \cdots \sum_{\substack{p_{m_1}=1 \\ \mathcal{E}^p_{m_1,i}}}^{N} (u_{p_i} - \bar{u})(v_{p_i} - \bar{v}) \tag{30}$$

where the sum $\sum_{p_i=1}^{N}$ with summation index p_i is placed first and the term

$$\mathcal{E}^p_{l,i} := \{p_l \mid p_l \neq p_k, k = 1, 2, ..., l-1\} \backslash \{p_i\}$$

excludes previously chosen index values and the index value chosen for p_i. The right side of (30) can be simplified to

$$\frac{(N-1)!}{(N-m_1)!} \sum_{p_i=1}^{N} (u_{p_i} - \bar{u})(v_{p_i} - \bar{v})$$

since the choice made for p_i in the first sum $\sum_{p_i=1}^{N}$ leaves $N-1$ choices for the second sum, $N-2$ choices for the third sum, up to $N-(m_1-1)$ choices for the last sum.

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

Moreover there are m_2 terms similar to (30) for each summation index, p_i. Thus the terms of the form $(u_{p_i} - \bar{u})(v_{p_i} - \bar{v})$ contribute

$$m_2 \frac{(N-1)!}{(N-m_1)!} \sum_{p=1}^{N} (u_p - \bar{u})(v_p - \bar{v}) \tag{31}$$

to the sum $S_{\bar{u},\bar{v}}$.

We now consider terms of the form $(u_{p_i} - \bar{u})(v_{p_j} - \bar{v})$ with two different summation indices p_i and p_j. The sum (21) applied to $(u_{p_i} - \bar{u})(v_{p_j} - \bar{v}), i \neq j$ can be written as

$$\sum_{\mathcal{P}} (u_{p_i} - \bar{u})(v_{p_j} - \bar{v}) =$$

$$\sum_{p_i=1}^{N} \sum_{\substack{p_j=1 \\ p_j \neq p_i}}^{N} \sum_{\substack{p_1=1 \\ \mathcal{E}^p_{1,i,j}}}^{N} \sum_{\substack{p_2=1 \\ \mathcal{E}^p_{2,i,j}}}^{N} \cdots \sum_{\substack{p_{m_1}=1 \\ \mathcal{E}^p_{m_1,i,j}}}^{N} (u_{p_i} - \bar{u})(v_{p_j} - \bar{v}) \tag{32}$$

where the sums $\sum_{p_i=1}^{N}$, $\sum_{\substack{p_j=1 \\ p_j \neq p_i}}^{N}$ with summation indices p_i and p_j are placed first and the term

$$\mathcal{E}^p_{l,i,j} := \{p_l \mid p_l \neq p_k, k = 1, 2, ..., l-1\} \backslash \{p_i, p_j\}$$

excludes previously chosen index values and the index values chosen for p_i and p_j. Equation (32) can be simplified to

$$\frac{(N-2)!}{(N-m_1)!} \sum_{p_i=1}^{N} \sum_{\substack{p_j=1 \\ p_j \neq p_i}}^{N} (u_{p_i} - \bar{u})(v_{p_j} - \bar{v})$$

since the choice made for p_i in the first sum $\sum_{p_i=1}^{N}$ and p_j in the second sum $\sum_{\substack{p_j=1, \\ p_j \neq p_i}}^{N}$ leaves $N-2$ choices for the third sum, $N-3$ choices for the fourth sum, up to $N-(m_1-1)$ choices for the last sum. Moreover there are $m_2(m_1-1)$ sums of the form (32) that can be identified when the terms in (23) are distributed. Therefore the terms of the form $(u_{p_i} - \bar{u})(v_{p_j} - \bar{v})$ contribute

$$m_2(m_1-1) \frac{(N-2)!}{(N-m_1)!} \sum_{p=1}^{N} \sum_{\substack{q=1 \\ q \neq p}}^{N} (u_p - \bar{u})(v_q - \bar{v}) \tag{33}$$

to the sum $S_{\bar{u},\bar{v}}$. Remove $m_2(m_1-1)\frac{(N-2)!}{(N-m_1)!}$ terms of the form $\sum_{p=1}^{N}(u_p - \bar{u})(v_p - \bar{v})$ from (31) and add them to (33) to form

$$m_2(m_1-1) \frac{(N-2)!}{(N-m_1)!} \sum_{p=1}^{N} \sum_{q=1}^{N} (u_p - \bar{u})(v_q - \bar{v}) =$$

$$m_2(m_1-1) \frac{(N-2)!}{(N-m_1)!} \sum_{p=1}^{N} (u_p - \bar{u}) \sum_{q=1}^{N} (v_q - \bar{v})$$

which is zero by (27). This leaves

$$\left[m_2 \frac{(N-1)!}{(N-m_1)!} - m_2(m_1-1) \frac{(N-2)!}{(N-m_1)!} \right] \sum_{p=1}^{N} (u_p - \bar{u})(v_p - \bar{v})$$

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

remaining terms from (31) which simplifies to (29) after multiplying by the factor $\frac{1}{m_1 m_2}$ present in (23).

Since $S_{\bar{u},\bar{v}}$ is a multiple of S as defined by (13), the system of equations (14) will be formed where $\alpha = \frac{(N-2)!}{m(N-m-1)!}$ when we set $m = m_1 = m_2$. Thus the multiple regression coefficients $\boldsymbol{\beta}$ computed from sampling distributions of the mean without replacement will be equal to the multiple regression coefficients computed from the original scores.

2.3 Difference between two groups

The results in Sections 2.1 and 2.2 generalize to a difference of two groups of sampling distributions. Let m_1 be the size of Group 1 and m_2 be the size of Group 2. The two groups can be composed to allow or exclude common elements.

2.3.1 Group 1 and Group 2 can share elements

Consider the expression S_d shown in (34). The sum $\sum_{\mathcal{Q}}$ composed of m_2 iterated sums is essentially the same sum shown in either (20) or (21) except that the indexing is done with q instead of p. We first examine the case where Group 1 and Group 2 can share elements: i.e. p_i may be equal to q_j.

$$S_d := \sum_{\mathcal{Q}} \sum_{\mathcal{P}} \left[\left(\frac{u_{p_1} + u_{p_2} + ... + u_{p_{m_1}}}{m_1} \right) - \left(\frac{u_{q_1} + u_{q_2} + ... + u_{q_{m_2}}}{m_2} \right) \right] \cdot \left[\left(\frac{v_{p_1} + v_{p_2} + ... + v_{p_{m_1}}}{m_1} \right) - \left(\frac{v_{q_1} + v_{q_2} + ... + v_{q_{m_2}}}{m_2} \right) \right]. \tag{34}$$

S_d can be written in the form

$$S_d = \sum_{\mathcal{Q}} \sum_{\mathcal{P}} \left[\left(\frac{u_{p_1} + u_{p_2} + ... + u_{p_{m_1}}}{m_1} - \bar{u} \right) - \left(\frac{u_{q_1} + u_{q_2} + ... + u_{q_{m_2}}}{m_2} - \bar{u} \right) \right] \cdot \left[\left(\frac{v_{p_1} + v_{p_2} + ... + v_{p_{m_1}}}{m_1} - \bar{v} \right) - \left(\frac{v_{q_1} + v_{q_2} + ... + v_{q_{m_2}}}{m_2} - \bar{v} \right) \right]. \tag{35}$$

Distributing gives

$$\begin{aligned} S_d = & \sum_{\mathcal{Q}} \sum_{\mathcal{P}} \left[\left(\frac{u_{p_1} + u_{p_2} + ... + u_{p_{m_1}}}{m_1} - \bar{u} \right) \left(\frac{v_{p_1} + v_{p_2} + ... + v_{p_{m_1}}}{m_1} - \bar{v} \right) \right] \\ & - \sum_{\mathcal{Q}} \sum_{\mathcal{P}} \left[\left(\frac{u_{q_1} + u_{q_2} + ... + u_{q_{m_2}}}{m_2} - \bar{u} \right) \left(\frac{v_{p_1} + v_{p_2} + ... + v_{p_{m_1}}}{m_1} - \bar{v} \right) \right] \\ & - \sum_{\mathcal{Q}} \sum_{\mathcal{P}} \left[\left(\frac{u_{p_1} + u_{p_2} + ... + u_{p_{m_1}}}{m_1} - \bar{u} \right) \left(\frac{v_{q_1} + v_{q_2} + ... + v_{q_{m_2}}}{m_2} - \bar{v} \right) \right] \\ & + \sum_{\mathcal{Q}} \sum_{\mathcal{P}} \left[\left(\frac{u_{q_1} + u_{q_2} + ... + u_{q_{m_2}}}{m_2} - \bar{u} \right) \left(\frac{v_{q_1} + v_{q_2} + ... + v_{q_{m_2}}}{m_2} - \bar{v} \right) \right]. \end{aligned} \tag{36}$$

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

After one accounts for the sums $\sum_{\mathcal{Q}}$ in the first term and $\sum_{\mathcal{P}}$ in the fourth term, one can write

$$\begin{aligned} S_d = A\sum_{\mathcal{P}}\left[\left(\frac{u_{p_1}+u_{p_2}+...+u_{p_{m_1}}}{m_1}-\bar{u}\right)\left(\frac{v_{p_1}+v_{p_2}+...+v_{p_{m_1}}}{m_1}-\bar{v}\right)\right] \\ -\sum_{\mathcal{P}}\left(\frac{v_{p_1}+v_{p_2}+...+v_{p_{m_1}}}{m_1}-\bar{v}\right)\sum_{\mathcal{Q}}\left(\frac{u_{q_1}+u_{q_2}+...+u_{q_{m_2}}}{m_2}-\bar{u}\right) \\ -\sum_{\mathcal{P}}\left(\frac{u_{p_1}+u_{p_2}+...+u_{p_{m_1}}}{m_1}-\bar{u}\right)\sum_{\mathcal{Q}}\left(\frac{v_{q_1}+v_{q_2}+...+v_{q_{m_2}}}{m_2}-\bar{v}\right) \\ +B\sum_{\mathcal{Q}}\left[\left(\frac{u_{q_1}+u_{q_2}+...+u_{q_{m_2}}}{m_2}-\bar{u}\right)\left(\frac{v_{q_1}+v_{q_2}+...+v_{q_{m_2}}}{m_2}-\bar{v}\right)\right] \end{aligned} \tag{37}$$

where

$$A=\begin{cases} N^{m_2} & \text{with replacement} \\ \frac{N!}{(N-m_2)!} & \text{without replacement} \end{cases}$$

and

$$B=\begin{cases} N^{m_1} & \text{with replacement} \\ \frac{N!}{(N-m_1)!} & \text{without replacement.} \end{cases}$$

The expression for A can be derived by recognizing that N choices are available for each of the m_2 sums in $\sum_{\mathcal{Q}}$ when the selections are made with replacement. When the selections are made without replacement, there are N choices for the first sum, $N-1$ choices for the second sum, up to $N-(m_2-1)$ for the m_2'th sum. The same reasoning can be used to derive the expression for B. By (16), the second and third terms in (37) are zero and can be eliminated. Using equations (24) and (29), one can simplify (37) to

$$S_d = C\sum_{p=1}^{N}(u_p-\bar{u})(v_p-\bar{v}), \tag{38}$$

where

$$C=\frac{N^{m_2}N^{m_1}}{N}\left(\frac{1}{m_1}+\frac{1}{m_2}\right)$$

when selections are made with replacement and

$$C=\frac{N!(N-2)!}{(N-m_1-1)!(N-m_2-1)!}\left[\frac{1}{m_1(N-m_2)}+\frac{1}{m_2(N-m_1)}\right]$$

when selections are made without replacement.

Note that the mean of a difference of two groups of sampling distributions of the mean is zero. When $\bar{u}$ and $\bar{v}$ are set to zero in (19) and a difference of two groups of sampling distributions are used, it is evident that S_d is similar in format to (19). Thus the system of equations (14) will be formed where $\alpha = C$. Thus the multiple regression coefficients $\boldsymbol{\beta}$ computed from a difference of two groups of sampling distributions of the mean will be equal to the multiple regression coefficients computed from the original scores.

2.3.2 Group 1 and Group 2 do not simultaneously share elements

We also consider the case where Group 1 and Group 2 do not simultaneously share any elements. We assume the selections are done without replacement. Under these

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

restrictions, one can write (36) as

$$\begin{aligned}\tilde{S}_d = \tilde{A}\sum_{\mathcal{P}}\left[\left(\frac{u_{p_1}+u_{p_2}+...+u_{p_{m_1}}}{m_1}-\bar{u}\right)\left(\frac{v_{p_1}+v_{p_2}+...+v_{p_{m_1}}}{m_1}-\bar{v}\right)\right]\\ -\sum_{\mathcal{P}}\left(\frac{v_{p_1}+v_{p_2}+...+v_{p_{m_1}}}{m_1}-\bar{v}\right)\sum_{\substack{\mathcal{Q}\\ \mathcal{Q}\neq\mathcal{P}}}\left(\frac{u_{q_1}+u_{q_2}+...+u_{q_{m_2}}}{m_2}-\bar{u}\right)\\ -\sum_{\mathcal{P}}\left(\frac{u_{p_1}+u_{p_2}+...+u_{p_{m_1}}}{m_1}-\bar{u}\right)\sum_{\substack{\mathcal{Q}\\ \mathcal{Q}\neq\mathcal{P}}}\left(\frac{v_{q_1}+v_{q_2}+...+v_{q_{m_2}}}{m_2}-\bar{v}\right)\\ +\tilde{B}\sum_{\mathcal{Q}}\left[\left(\frac{u_{q_1}+u_{q_2}+...+u_{q_{m_2}}}{m_2}-\bar{u}\right)\left(\frac{v_{q_1}+v_{q_2}+...+v_{q_{m_2}}}{m_2}-\bar{v}\right)\right]\end{aligned} \tag{39}$$

where

$$\tilde{A}=\frac{(N-m_1)!}{(N-m_1-m_2)!}\ \text{ without replacement.}$$

and

$$\tilde{B}=\frac{(N-m_2)!}{(N-m_1-m_2)!}\ \text{ without replacement.}$$

The notation $\mathcal{Q}\neq\mathcal{P}$ is used to exclude any elements in the sum $\mathcal{Q}$ from indices previously selected in the sum $\mathcal{P}$. The $\tilde{A}$ coefficient can be derived by noting that for the distinct m_1 indices $\{p_1,p_2,\ldots,p_{m_1}\}$ chosen in $\sum_{\mathcal{P}}$, there remain $N-m_1$ choices for the first sum in $\sum_{\mathcal{Q}}$, $N-m_1-1$ choices for the second sum, and so on up to $N-m_1-(m_2-1)$ choices for the m_2'th sum. Similarly, the $\tilde{B}$ coefficient can be derived by noting that for the distinct m_2 indices $\{q_1,q_2,\ldots,q_{m_2}\}$ chosen in $\sum_{\mathcal{Q}}$, there remain $N-m_2$ choices for the first sum in $\sum_{\mathcal{P}}$, $N-m_2-1$ choices for the second sum, and so on up to $N-m_2-(m_1-1)$ choices for the m_1'th sum. Turning to the second half of the second term of (39), which we define to be

$$S_u:=\sum_{\substack{\mathcal{Q}\\ \mathcal{Q}\neq\mathcal{P}}}\left(\frac{u_{q_1}+u_{q_2}+...+u_{q_{m_2}}}{m_2}-\bar{u}\right)=$$

$$\frac{1}{m_2}\sum_{\substack{q_1=1\\ \mathcal{E}_1^{q,p}}}^{N}\sum_{\substack{q_2=1\\ \mathcal{E}_2^{q,p}}}^{N}\cdots\sum_{\substack{q_{m_2}=1\\ \mathcal{E}_{m_2}^{q,p}}}^{N}\left((u_{q_1}-\bar{u})+(u_{q_2}-\bar{u})+...+(u_{q_{m_2}}-\bar{u})\right) \tag{40}$$

where

$$\mathcal{E}_l^{q,p}:=\{q_l\mid q_l\neq q_k,k=1,2,...,l-1\}\backslash\{p_1,p_2,\ldots,p_{m_1}\}$$

excludes previously chosen indices in the $\sum_{\mathcal{Q}}$ sum and any previously chosen indices $\{p_1,p_2,\ldots,p_{m_1}\}$ selected from the $\sum_{\mathcal{P}}$ sum. Applying the sum to the specific term $(u_{q_i}-\bar{u})$, the sum $\sum_{\mathcal{Q}}$ can be rearranged as

$$S_{u_i}:=\frac{1}{m_2}\sum_{\substack{q_i=1\\ \mathcal{E}^{q_i,p}}}^{N}\sum_{\substack{q_1=1\\ \mathcal{E}_{1,i}^{q,p}}}^{N}\sum_{\substack{q_2=1\\ \mathcal{E}_{2,i}^{q,p}}}^{N}\cdots\sum_{\substack{q_{i-1}=1\\ \mathcal{E}_{i-1,i}^{q,p}}}^{N}\sum_{\substack{q_{i+1}=1\\ \mathcal{E}_{i+1,i}^{q,p}}}^{N}\cdots\sum_{\substack{q_{m_2}=1\\ \mathcal{E}_{m_2,i}^{q,p}}}^{N}(u_{q_i}-\bar{u}),$$

where the term

$$\mathcal{E}^{q_i,p}:=\{q_i\mid q_i\neq p_k,k=1,2,...,m_1\} \tag{41}$$

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

excludes previously chosen indices $\{p_1, p_2, \ldots, p_{m_1}\}$ selected from the $\sum_{\mathcal{P}}$ sum and where the term

$$\mathcal{E}^{q,p}_{l,i} := \{q_l \mid q_l \neq q_k, k = 1, 2, ..., l-1\} \backslash \{q_i, p_1, p_2, \ldots, p_{m_1}\}$$

excludes previously chosen indices in the $\sum_{\mathcal{Q}}$ sum, the index q_i chosen in the $\sum_{\substack{q_i=1 \\ \mathcal{E}^{q_i,p}}}^{N}$ sum, and any previously chosen indices $\{p_1, p_2, \ldots, p_{m_1}\}$ selected from the $\sum_{\mathcal{P}}$ sum. Bear in mind that the sums

$$\sum_{\substack{q_1=1 \\ \mathcal{E}^{q,p}_{1,i}}}^{N} \sum_{\substack{q_2=1 \\ \mathcal{E}^{q,p}_{2,i}}}^{N} \cdots \sum_{\substack{q_{i-1}=1 \\ \mathcal{E}^{q,p}_{i-1,i}}}^{N} \sum_{\substack{q_{i+1}=1 \\ \mathcal{E}^{q,p}_{i+1,i}}}^{N} \cdots \sum_{\substack{q_{m_2}=1 \\ \mathcal{E}^{q,p}_{m_2,i}}}^{N} (u_{q_i} - \bar{u}) \tag{42}$$

will contribute the same factor to $(u_{q_i} - \bar{u})$ regardless of the selected value for the summation index q_i. Taking care to avoid selecting a index that has been already chosen, we note that m_1 choices have already been made for the set $\{p_1, p_2, \ldots, p_{m_1}\}$. In addition, for each choice of q_i in $\sum_{\substack{q_i=1 \\ \mathcal{E}^{q_i,p}}}^{N}$ there remain $N - m_1 - 1$ choices left for the first sum $\sum_{\substack{q_1=1 \\ \mathcal{E}^{q,p}_{1,i}}}^{N}$, $N - m_1 - 2$ choices left for the second sum, up to $N - m_1 - (m_2 - 1)$ choices for the $(m_2 - 1)$'th sum or

$$\frac{(N - m_1 - 1)!}{(N - m_1 - m_2)!}$$

total choices. Thus

$$S_{u_i} = \frac{1}{m_2} \frac{(N - m_1 - 1)!}{(N - m_1 - m_2)!} \sum_{\substack{q_i=1 \\ \mathcal{E}^{q_i,p}}}^{N} (u_{q_i} - \bar{u}) \,. \tag{43}$$

Using the definition of the excluded terms $\mathcal{E}^{q_i,p}$ (41) in the sum $\sum_{\substack{q_i=1 \\ \mathcal{E}^{q_i,p}}}^{N} (u_{q_i} - \bar{u})$,

$$\sum_{\substack{q_i=1 \\ \mathcal{E}^{q_i,p}}}^{N} (u_{q_i} - \bar{u}) = \sum_{q_i=1}^{N} (u_{q_i} - \bar{u}) - \sum_{j=1}^{m_1} (u_{p_j} - \bar{u}), \tag{44}$$

one can replace $\sum_{\substack{q_i=1 \\ q_i \neq p_j}}^{N} (u_{q_i} - \bar{u})$ in (43) with the right hand side of (44) to yield,

$$S_{u_i} = \frac{1}{m_2} \frac{(N - m_1 - 1)!}{(N - m_1 - m_2)!} \left[\sum_{q_i=1}^{N} (u_{q_i} - \bar{u}) - \sum_{j=1}^{m_1} (u_{p_j} - \bar{u}) \right].$$

The sum $\sum_{q_i=1}^{N} (u_{q_i} - \bar{u})$ is zero by (27). Since there are m_2 terms of the form $(u_{q_i} - \bar{u})$ in (40), S_u can be written as

$$S_u = -\frac{(N - m_1 - 1)!}{(N - m_1 - m_2)!} \sum_{j=1}^{m_1} \left(u_{p_j} - \bar{u}\right)$$

or

$$S_u = -m_1 \frac{(N - m_1 - 1)!}{(N - m_1 - m_2)!} \left(\frac{u_{p_1} + u_{p_2} + \ldots + u_{p_{m_1}}}{m_1} - \bar{u} \right).$$

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

We can apply the same steps to the second half of the third term of (39), which we define to be S_v

$$S_v := \sum_{\substack{\mathcal{Q} \\ \mathcal{Q}\neq\mathcal{P}}} \left(\frac{v_{q_1} + v_{q_2} + ... + v_{q_{m_2}}}{m_2} - \bar{v} \right)$$

to show that

$$S_v = -m_1 \frac{(N - m_1 - 1)!}{(N - m_1 - m_2)!} \left(\frac{v_{p_1} + v_{p_2} + ... + v_{p_{m_1}}}{m_1} - \bar{v} \right).$$

Using these results in (39),

$$\begin{aligned} \tilde{S}_d = \tilde{A} \sum_{\mathcal{P}} \left[\left(\frac{u_{p_1} + u_{p_2} + ... + u_{p_{m_1}}}{m_1} - \bar{u} \right) \left(\frac{v_{p_1} + v_{p_2} + ... + v_{p_{m_1}}}{m_1} - \bar{v} \right) \right] \\ + 2\tilde{D} \sum_{\mathcal{P}} \left(\frac{u_{p_1} + u_{p_2} + ... + u_{p_{m_1}}}{m_1} - \bar{u} \right) \left(\frac{v_1 + v_2 + \ldots + v_{p_{m_1}}}{m_1} - \bar{v} \right) \\ + \tilde{B} \sum_{\mathcal{Q}} \left[\left(\frac{u_{q_1} + u_{q_2} + ... + u_{q_{m_2}}}{m_2} - \bar{u} \right) \left(\frac{v_{q_1} + v_{q_2} + ... + v_{q_{m_2}}}{m_2} - \bar{v} \right) \right] \end{aligned} \tag{45}$$

where

$$\tilde{D} = m_1 \frac{(N - m_1 - 1)!}{(N - m_1 - m_2)!}.$$

Using equation (29), (45) simplifies to

$$\tilde{S}_d = \tilde{C} \sum_{p=1}^{N} (u_p - \bar{u})(v_p - \bar{v}), \tag{46}$$

where

$$\tilde{C} = \frac{(N-2)!}{(N - m_1 - m_2)!} \left(\frac{(N - m_1)!}{m_1(N - m_1 - 1)!} + \frac{(N - m_2)!}{m_2(N - m_2 - 1)!} + 2 \right),$$

keeping in mind that the selections are made without replacement. Again $\tilde{S}_d$ is a multiple of S. Therefore the system of equations (14) will be formed where $\alpha = \tilde{C}$.

3 Coefficient of determination

The coefficient of determination R^2 is the proportion of variability in the dependent variable that can be accounted for by the independent variables [6]. It is defined using

$$R^2 = \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{47}$$

where $\hat{y}_i$ is the prediction provided by the surface of regression

$$(\hat{y}_i - \bar{y}) = \sum_{j=1}^{K} \beta_j (x_i^{(j)} - \bar{x}^{(j)}). \tag{48}$$

Substituting (48) into (47),

$$R^2 = \frac{\sum_{i=1}^{N} \left(\sum_{j=1}^{K} \beta_j (x_i^{(j)} - \bar{x}^{(j)}) \right)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}, \tag{49}$$

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

$$R^2 = \frac{\sum_{i=1}^{N}\left(\sum_{j=1}^{K}\sum_{j'=1}^{K}\beta_j\beta_{j'}(x_i^{(j)}-\bar{x}^{(j)})(x_i^{(j')}-\bar{x}^{(j')})\right)}{\sum_{i=1}^{N}(y_i-\bar{y})^2}, \tag{50}$$

$$R^2 = \frac{\sum_{j=1}^{K}\sum_{j'=1}^{K}\beta_j\beta_{j'}\sum_{i=1}^{N}(x_i^{(j)}-\bar{x}^{(j)})(x_i^{(j')}-\bar{x}^{(j')})}{\sum_{i=1}^{N}(y_i-\bar{y})^2}. \tag{51}$$

Again we see the presence of sums $\sum_{i=1}^{N}(x_i^{(j)}-\bar{x}^{(j)})(x_i^{(j')}-\bar{x}^{(j')})$, $\sum_{i=1}^{N}(y_i-\bar{y})^2$ of the form (13). Both numerator and denominator will be multiplied by the same constant according to (24), (29), (38), and (46) leaving the coefficient of determination R^2 unchanged when elements of the sampling distribution of the mean or differences of two groups of sampling distributions of the mean are used. For one independent variable, R will have the same sign for sampling distributions of the mean and individual scores since R shares the same sign as the linear regression slope.

4 Standard error of estimate

The standard error of estimate is a measure of accuracy for the surface of regression [7]. In this section, we show the standard error of estimate is reduced by the factor $\frac{1}{\sqrt{m}}$ where m is the group size for sampling distributions with replacement and by

$$\frac{1}{\sqrt{m}}\sqrt{\frac{N-m}{N-1}}$$

for sampling distributions without replacement.

The sum of squares error SSE is defined to be

$$SSE = \sum_{i=1}^{N}(y_i-\hat{y}_i)^2. \tag{52}$$

Given this definition, the standard error of estimate s_e can be defined

$$s_e = \sqrt{\frac{SSE}{N-2}}. \tag{53}$$

Now by [8]

$$SST = SSR + SSE \tag{54}$$

where SST is the total variation and SSR is the sum of squared regression,

$$SST = \sum_{i=1}^{N}(y_i-\bar{y})^2, \quad SSR = \sum_{i=1}^{N}(\hat{y}_i-\bar{y})^2. \tag{55}$$

With these definitions, the coefficient of determination (47) can also be written as

$$R^2 = \frac{SSR}{SST} = \frac{SST-SSE}{SST}. \tag{56}$$

Solving (56) for SSE and dividing by N

$$\frac{SSE}{N} = \frac{SST}{N}\left(1-R^2\right) = \sigma^2\left(1-R^2\right) \tag{57}$$

where

$$\sigma^2 = \frac{SST}{N} \tag{58}$$

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

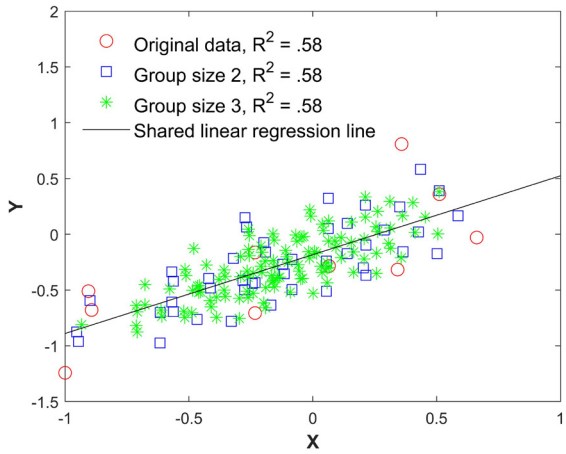

Figure 1. Original data (x_i, y_i), all elements from the sampling distributions of the mean, and the shared linear regression line. The red circles are the original 15 points, the blue squares are the averaged data of size $m = 2$, and the green asterisks are the averaged data of size $m = 3$ without replacement.

is the population variance. Now by (53) $SSE = (N-2)s_e^2$. Replacing SSE in (57) with $(N-2)s_e^2$ and solving for s_e yields

$$s_e = \sigma\sqrt{1-R^2}\sqrt{\frac{N}{N-2}}. \tag{59}$$

When sampling distributions of the mean are used, R remains the same, but σ is replaced by $\sigma_{\bar{Y}}$ where

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{m}} \tag{60}$$

for sampling distributions with replacement and

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{m}}\sqrt{\frac{N-m}{N-1}} \tag{61}$$

for sampling distributions without replacement [7]. Thus for selections made with replacement and for selections made without replacement (if $N >> m$), s_e will be reduced by $\frac{1}{\sqrt{m}}$ when sampling distributions of the mean are used. This result is analogous to the reduction of the standard deviation by $\frac{1}{\sqrt{m}}$ when using sampling distributions of the mean for one variable.

5 Numerical simulations

Figure 1 plots the original data $\{(x_i, y_i), 1 \leq i \leq 10\}$ in red and all elements from the sampling distribution of the mean generated without replacement for groups of size $m = 2$ in blue and $m = 3$ in green for $N = 10$ original points. The original data and elements from the sampling distribution of the mean share the same regression line and coefficient of determination R^2. The elements of the sampling distribution of the mean are clustered more closely about the regression line compared to the original data which is consistent with (59) and (61).

Figure 2 plots the original data $\{(x_i, y_i, z_i), 1 \leq i \leq 15\}$ in red and all elements from the sampling distribution of the mean generated without replacement for groups of size

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

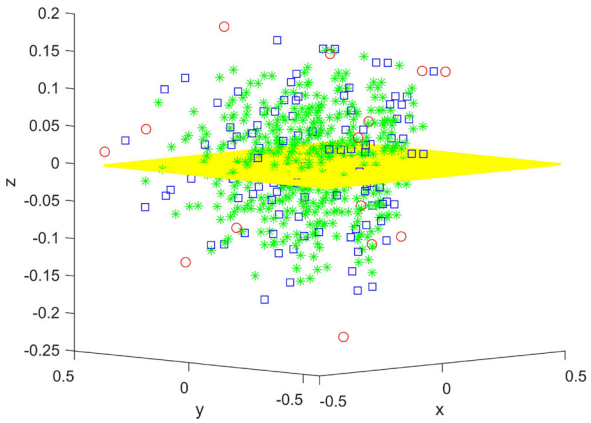

Figure 2. Original data $(x_i, y_i, z_i), i = 1, 15$ in red and all elements from the sampling distribution of the mean for $m = 2$ (blue) and $m = 3$ (green), and the shared linear regression plane.

$m = 2$ in blue and $m = 3$ in green for $N = 15$ original points. The original data and elements from the sampling distribution of the mean share the same regression plane $z = .653x - .712y$ and coefficient of determination $R^2 = 0.76$. For visualization purposes, the normal distance to the plane is plotted as the z-coordinate and the multiple regression plane is aligned with the $z = 0$ plane.

Figure 3 shows the convergence of sampling distributions of the mean for $\{(x_i, y_i), i = 1, 2, ..., N\}$, $N = 11$ scores with Pearson correlation coefficient $R = 0.35$ and linear regression slope $\beta_1 = 0.27$. In the first simulation shown in black and red, elements from the sampling distributions of the mean are created using groups of size $m = 5$ without replacement. In the second simulation shown in blue and green, elements from the sampling distributions of the mean are created using differences of two groups of size $m_1 = 4$ and $m_2 = 2$ without replacement. The horizontal axis plots the fraction of total selections used in the sampling distributions. There are $11^5 = 161,051$ total selections for the first simulation and $11!/5! = 332,640$ total selections for the second simulation. The vertical axis plots the base 10 logarithm of the absolute difference. The absolute difference can be between either the Pearson R = 0.35 based on individual scores and the Pearson R computed from a fraction of the elements from the sampling distributions of the mean (black and blue graphs), **or** between the linear regression slope $\beta_1 = .27$ based on the original scores and the slope computed using a fraction of the elements from the sampling distributions of the mean (red and green graphs). While not entirely obvious due to the density of points, all differences decrease from approximately 10^{-6} to less than 10^{-13} in the last 0.001% of the total selections. In addition, the differences do not always decrease monotonically as the fraction of total selections increase, and the differences decrease to very small values (less than 10^{-5}) at certain points during the course of the convergence as noted by the downward spikes.

6 Gene expression and distance between genes

A useful way of organizing the data obtained from microarrays or RNA-seq data is to group together genes that exhibit similar expression patterns through hierarchical clustering. A hierarchical clustering algorithm generates a dendrogram (tree diagram). However, the algorithm requires that a distance be defined to quantify similarities in expression between two individual genes.

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

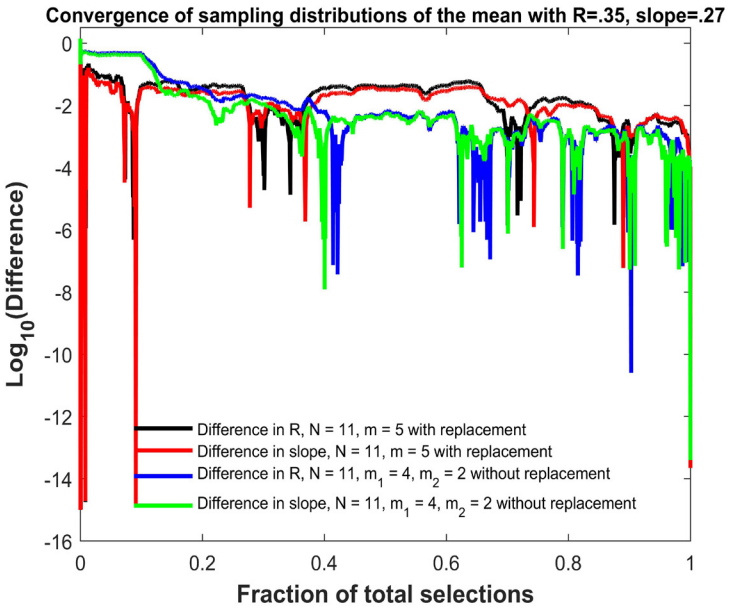

Fig 3. Convergence of the Pearson R and linear regression slope for sampling distributions of the mean.

Let A_i denote the expression level of gene A for patient i and let B_i denote the expression level for gene B for patient i, $1 \leq i \leq N$. Distances between genes can be computed using many metrics [9], but two common ones are the Euclidean distance

$$D_E = \sum_{i=1}^{N} \sqrt{(A_i - B_i)^2}, \tag{62}$$

and the Manhattan distance,

$$D_M = \sum_{i=1}^{N} |A_i - B_i|.$$

Correlation coefficients [10] are also used to measure the similarities between two genes. One measure of distance using the Pearson R is

$$D_R = 1 - |R|, \tag{63}$$

$$R = \frac{\sum_{i=1}^{N}(A_i - \bar{A})(B_i - \bar{B})}{\sqrt{\sum_{i=1}^{N}(A_i - \bar{A})^2 \sum_{i=1}^{N}(B_i - \bar{B})^2}} \tag{64}$$

or $D'_R = 1 - R^2$ [11] if the sign of R is not important. If R is close to 1 or -1, the distances D_R, D'_R will be close to zero.

The purpose of the next section is to propose a new distance based on the differential expression of two genes. We then show the new measure of distance is the same as the Pearson R coefficient computed from the original scores (64), thus lending support to the use of the Pearson R coefficient in measuring the distance between two genes.

6.1 Formulating a new distance between two genes

Let us formulate a new distance based on differential expression. Select $m_1 \leq \frac{N}{2}$ distinct random patients and their expression levels for gene A and assign them to

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

Group 1. Select a second group of $m_2 \leq \frac{N}{2}$ distinct (and different from Group 1) random patients and assign their expression levels to Group 2. Repeat the process using the same selections for gene B. Since both groups are sampled from a population with a known variance σ^2, the z-statistic [12] for two independent samples can be used to measure differential expression for gene A

$$z_A = \frac{\bar{A}_1 - \bar{A}_2}{\sqrt{\frac{\sigma^2}{m_1} + \frac{\sigma^2}{m_2}}} \tag{65}$$

which if $m_1, m_2 \geq 30$ will be approximately normally distributed. Let z_B be the z-statistic for gene B for the same selection of patients using the same equation (65). This process can be repeated multiple times giving a set of ordered pairs (z_A^k, z_B^k) for each different selection (k) of groups. The Pearson R value, R_t can then be computed from these ordered pairs using all possible selections K

$$R_t = \frac{\sum_{k=1}^{K}(z_A^k - \bar{z}_A)(z_B^k - \bar{z}_B)}{\sqrt{\sum_{k=1}^{K}(z_A^k - \bar{z}_A)^2 \sum_{k=1}^{K}(z_B^k - \bar{z}_B)^2}}. \tag{66}$$

The new distance will now be defined as D_T or alternatively D'_T

$$D_T = 1 - |R_t|,\ D'_T = 1 - R_t^2. \tag{67}$$

Given N total patients, there exist $K = \frac{N!}{(N-m_1-m_2)!}$ total selections. Computing all selections is prohibitive for large N. However, we know from the analysis in Section 2.3.2, and the fact that the Pearson R coefficient is not affected by the multiplicative factor $\sqrt{\frac{\sigma^2}{m_1} + \frac{\sigma^2}{m_2}}$ in (65), that the distance D_R will be equal D_T and D'_R will be equal D'_T.

7 Conclusion

We have shown that the linear regression coefficients (simple and multiple) and the coefficient of determination R^2 computed from sampling distributions of the mean (with or without replacement) are equal to the regression coefficients and coefficient of determination computed with the original data. This result also applies to a difference of two groups of sampling distributions of the mean. Moreover, the standard error of estimate is reduced by the square root of the group size for sampling distributions of the mean.

The result has implications for the construction of hierarchical clustering trees or heat maps which visualize the relationship between many genes. These processes require one to define a distance between two genes using their expression levels. We developed a new measure of distance based on how differential expression in one gene correlates with differential expression in a second gene using the z-statistic. We showed that the new measure is equivalent to the Pearson R coefficient computed from the original scores, thus lending support to the use of the Pearson R coefficient for measuring a distance between two genes.

Funding: This is research is supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103451. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

Torres DJ, et al., J Bioinform Syst Biol 2024
DOI:10.26502/jbsb.5107079

References

1. Yan X. Linear regression analysis: Theory and computing. Singapore: World Scientific; 2009.

2. Robinson WS. Ecological correlations and the behavior of individuals. Am Sociol Rev. 1950; 15(3):351-357, https://doi.org/10.2307/2087176.

3. Goodman LA. Ecological regressions and the behavior of individuals. Am Sociol Rev. 1953; 18:663-664.

4. Kenney JF. Mathematics of statistics, Part Two. New York: D. Van Nostrand Company, Inc.; 1947.

5. Pearson K. On the probable errors of frequency constants. Biometrika. 1913; 9:1-10.

6. Pagano R. Understanding statistics in the behavioral sciences, 10th ed. Belmont, California: Wadsworth Publishing; 2012.

7. Sullivan M III. Fundamentals of statistics, 3rd ed. New York: Pearson; 2010.

8. Walpole RE and Myers RH. Probability and statistics for engineers and scientists, 3rd ed. New York: MacMillan Publishing Company; 1985.

9. D'haeseleer P. How does gene expression clustering work? Nat Biotechnol. 2005; 23:1499-1501.

10. Quackenbush J. Computational analysis of microarray data. Nat Rev Genet. 2001; 2(6):418-427.

11. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA. 1998; 95(25):14863-14868.

12. Triola MF. Elementary statistics using the TI-83/84 Plus calculator, 3rd ed. New York: Addison-Wesley; 2011.