Pivot Property in Weighted Least Regression Based on Single Repeated Observations

Yuancheng Si¹

Received: 13 September 2019 / Revised: 1 April 2020 / Accepted: 9 April 2020 / Published online: 6 May 2020
© The Author(s) 2020

Abstract

How genuine repeated observations influence the result of linear regression could be an interesting topic in regression analysis. In this paper, we discuss the pivot behavior in weighted linear regression based on certain data pattern and give an explanation about the pivot behavior.

Keywords Least square regression · Duplicated observations · Weighted least square · Weight scheme · Repeated data

1 Introduction

Repetition of identical observations is an important phenomenon in a variety of domains such as clinical trials, categorical survey, traffic data [1, 2] and how genuine repeated observations influence the result of linear regression could be an interesting topic in regression analysis. As [3] stated, 'It is important to understand that repeated runs must be genuine repeats and not just repetitions of the same reading'. This kind of data with genuine repeats could be easily found in categorical data analysis and obtaining estimators for unknown parameters within the framework of statistics always comes down to a problem of optimization using the least square method and that could be the origin of our study in this paper. In [4], the author investigates how single repeating data point (x_i, y_i) affects the ordinary least squares regression and the result shows that the regression lines will pivot at some certain point under specific condition. i.e. $\sum_{j \neq i}^n x_j \neq (n-1)x_i$ where the pivot point P has coordinate (x_p, y_p) as

$$x_p = x_i + \frac{\sum_{j \neq i}^n (x_j - x_i)^2}{\sum_{j \neq i}^n (x_j - x_i)} \quad (1)$$

✉ Yuancheng Si
yuancheng.si@manchester.ac.uk

¹ School of Mathematics, University of Manchester, Manchester M13 9PL, UK

$$y_P = y_i + \frac{\sum_{j \neq i}^n (x_j - x_i)(y_j - y_i)}{\sum_{j \neq i}^n (x_j - x_i)} \quad (2)$$

The study background of [4] is based on predicting the result of sports games using matches information, which makes the explanatory variable and the response variable are both precise positive integer numbers (like rank or number of games) where the occurrences of repeated observation always exist. In [4], we find that if certain data point was repeated several times and we employ simple linear regression based on updated dataset each time, we could find all the simple linear regression lines would intersect at some pivot point P (Fig. 1).

However, once repeated data point (for simplicity, we denote it as (x_1, y_1)) is not identical with P . We would find the series of squares of residuals after regression for repeated points (x_1, y_1) will converge to 0. The series of squares of residuals for non-repeated data points in datasets $\hat{\epsilon}_2^k, \hat{\epsilon}_3^k \dots \hat{\epsilon}_n^k$ will converge to a fixed non-zero constant as the number of repeated times k increases, which could lead to heteroscedasticity and does not satisfy Gauss–Markov theorem through hypothesis testing which leads to failure of the best linear unbiased estimator (BLUE) of the coefficients given by Ordinary Least Squares regression.

Fig. 1 The pivot properties by repeated each single data point 20 times based on given dataset

For example, we use the datasets given from [4] (1, 6), (2, 2), (5, 1), (7, 2), if we repeat (2, 2) for additional 19 times and do regression using OLS method, we could find that the simple linear model is no longer valid based on the result of either t-test or F-test, we can also exert heteroscedasticity tests using Breusch-Pagan test or Whites General Test to detect this issue.

```

1 Linear regression model:
2 y ~ 1 + x1
3
4 Estimated Coefficients:
5 Estimate      SE      tStat      pValue
6 -----
7 (Intercept)    2.6726    0.36703    7.2817    2.713e-07
8 x1             -0.23894   0.1426    -1.6756
9
10
11
12 Number of observations: 24, Error degrees of freedom: 22
13 Root Mean Squared Error: 0.819
14 R-squared: 0.113, Adjusted R-Squared 0.0729
15 F-statistic vs. constant model: 2.81, p-value = 0.108

```

We should mention that as the sample size increases, there could be a trade-off between invalidity of the model through p value and sample size [5]. For a more precise test, we could use alternative and exact solution of power and sample size calculations for model validation in linear regression analysis. [6, 7]

Meanwhile, bringing back to real-world cases, different observations have a different reliable factor; for example, we cannot treat the result of friendly match and finals match with equal importance. Thus we should treat various observations with different importance factor respectively, which inspire us to investigate whether the pivot behaviour exists in more robust linear models. In this paper, we mainly focus on the pivot behaviour based on Weighted least square(WLS) method.

2 Regression Lines and Pivot Property Under Simple Linear Regression Using WLS Method

To be more generalized, supposing we start with the points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ and we consistently use (x_1, y_1) as the repeated point R in following work. Note that we repeat (x_1, y_1) for additional k times so we have design matrix $X \in \mathbb{R}^{(n+k) \times 2}$ as

$$k + 1 \text{ times} \left\{ \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \right\}_{(n+k) \times 2}$$

Assume we have prior information about the weight matrix $W \in \mathbb{R}^{(n+k) \times (n+k)}$ as

$$\begin{bmatrix} w_{1,1} & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & w_{1,k+1} & 0 & \dots & 0 \\ 0 & \dots & 0 & w_2 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & w_n \end{bmatrix}_{(n+k) \times (n+k)}$$

$w_{1,1}, \dots, w_{1,k+1} \in \mathbb{R}^+$ are corresponding weights for total $k + 1$ repeated points (x_1, y_1) respectively and $w_2, \dots, w_n \in \mathbb{R}^+$ are weights for non-repeated points. Note that we are not engaged in discussing how to determine the optimal weight scheme for each observation in this article, it depends on personal judgement on the precision of measurement or specific weight scheme which could make estimator for unknown parameters significant under WLS method. Meanwhile, we should notice the weights $w_{1,1}, \dots, w_{1,k+1}$ for total $k + 1$ times repeated data points (x_1, y_1) do not need to be identical and their sum do not need to add up to 1 under our parameterization.

Suppose the model under consideration is

$$Y = X\beta + \epsilon \quad (3)$$

$$\beta = \begin{bmatrix} m_{k+1} \\ b_{k+1} \end{bmatrix} \quad (4)$$

$$\epsilon \sim N(0, W\sigma^2) \quad (5)$$

where the intercept is b_{k+1} is the value of $E(Y|X = x)$ when x equals zero and the slope m_{k+1} is the rate of change in $E(Y|X = x)$ for a unit change in X both given under $k + 1$ times repeated data points (x_1, y_1) .

$$Y = \left\{ \begin{bmatrix} y_1 \\ \vdots \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right\}_{k+1 \text{ times}}$$

The normal equation

$$X^T W X \beta = X^T W Y \quad (6)$$

becomes

$$\begin{aligned} & \begin{bmatrix} x_1^2 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i x_i^2, & x_1 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i x_i \\ x_1 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i x_i, & \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i \end{bmatrix} \begin{bmatrix} m_k \\ b_k \end{bmatrix} \\ &= \begin{bmatrix} x_1 y_1 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i x_i y_i \\ y_1 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i y_i \end{bmatrix} \end{aligned} \quad (7)$$

From second component of normal equation we have

$$\frac{x_1 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i x_i}{\sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i} m_k + b_k = \frac{y_1 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i y_i}{\sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i} \quad (8)$$

which means each regression line includes weighted average means of the x - and y - coordinates of total $n + k$ data points, we denote it as $M_k(\mu_{x;k}, \mu_{y;k})$ where

$$\mu_{x;k} = \frac{x_1 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i x_i}{\sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i} \quad (9)$$

$$\mu_{y;k} = \frac{y_1 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i y_i}{\sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i} \quad (10)$$

$(\mu_{x;k}, \mu_{y;k})$ turns to $(\bar{x}, \bar{y})$ as we assign equal weight for each observation, i.e. $w_{1,1} = \dots = w_{1,k+1} = w_2 = \dots = w_n$, which is accord with our knowledge that ordinary least square regression line must go through the mean point of dataset. Meanwhile, the limiting position of $\mu_{x;k}$ and $\mu_{y;k}$ as $k \rightarrow +\infty$ depends on the convergence of $\sum_{i=1}^{k+1} w_{1,i}$, that is, if $\lim_{k \rightarrow +\infty} \sum_{i=1}^{k+1} w_{1,i} = +\infty$ we have

$$\lim_{k \rightarrow +\infty} \mu_{x;k} = x_1 \quad (11)$$

$$\lim_{k \rightarrow +\infty} \mu_{y;k} = y_1, \quad (12)$$

i.e. $M_k \rightarrow P(x_1, y_1)$ as k grows.

Whereas if $\lim_{k \rightarrow +\infty} \sum_{i=1}^{k+1} w_{1,i} = c \in \mathbb{R} \geq 0$ we have

$$\lim_{k \rightarrow +\infty} \mu_{x;k} = \frac{x_1 c + \sum_{i=2}^n w_i x_i}{c + \sum_{i=2}^n w_i} \quad (13)$$

$$\lim_{k \rightarrow +\infty} \mu_{y;k} = \frac{y_1 c + \sum_{i=2}^n w_i y_i}{c + \sum_{i=2}^n w_i} \quad (14)$$

which is not approach to repeated point but a new point R_{skew} .

Note that the slope of the line segment connecting M_k and M_{k+1} is given by:

$$\frac{\mu_{y;(k+1)} - \mu_{y;k}}{\mu_{x;(k+1)} - \mu_{x;k}} = \frac{y_1(\sum_{i=2}^n w_i) - \sum_{i=2}^n w_i y_i}{x_1(\sum_{i=2}^n w_i) - \sum_{i=2}^n w_i x_i} \quad (15)$$

if (15) keeps invariant for a different choice of k , we can conclude that the weighted average points of data set is collinear for the quotient is independent of k . In fact, it mainly relies on our weight scheme setup for updated dataset. For example, if we assume that new added repeated point (x_1, y_1) will not influence the weights of non-repeated points $w_2, \dots, w_n$, (15) will be independent of k . On the contrary case, if we assume that the newly added repeated point will increase/decrease the original weights for non-repeated points with a linear pattern, for example, $w_i = w_i + c_i k$, the collinear property for weighted average points $M_0, M_1 \dots M_k$ could disappear for different selection of c_i .

2.1 Finding Pivot Point

Enlightened by method we used above, we tried to apply it to first component of normal equation (7) to find the pivot behavior. When $x_1 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i x_i \neq 0$, we have

$$\frac{x_1^2 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i x_i^2}{x_1 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i x_i} m + b = \frac{x_1 y_1 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i x_i y_i}{x_1 \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i x_i} \quad (16)$$

solving by coordinate system exchange from Cartesian coordinate to R -centric coordinate ($\chi_i = x_i - x_1, \gamma_i = y_i - y_1$), we express (16) as

$$\frac{\sum_{i=2}^n w_i \chi_i^2}{\sum_{i=2}^n w_i \chi_i} m_k + b_k = \frac{\sum_{i=2}^n w_i \chi_i \gamma_i}{\sum_{i=2}^n w_i \chi_i} \quad (17)$$

From (17) we clearly find that for any choice of k , all regression lines will intersect at point P under R -centric coordinate with proper weight scheme

$$\left(\frac{\sum_{i=2}^n w_i \chi_i^2}{\sum_{i=2}^n w_i \chi_i}, \frac{\sum_{i=2}^n w_i \chi_i \gamma_i}{\sum_{i=2}^n w_i \chi_i} \right) \quad (18)$$

or

$$\left(x_1 + \frac{\sum_{i=2}^n w_i (x_i - x_1)^2}{\sum_{i=2}^n w_i (x_i - x_1)}, y_1 + \frac{\sum_{i=2}^n w_i (x_i - x_1)(y_i - y_1)}{\sum_{i=2}^n w_i (x_i - x_1)} \right) \quad (19)$$

under Cartesian coordinate.

We should notice that the regression lines pivot at P with two prerequisite conditions.

1. $\sum_{i=2}^n w_i \chi_i \neq 0$ under R -centric coordiante.If $\sum_{i=2}^n w_i \chi_i = 0$, (7) could be written as

$$\begin{bmatrix} \sum_{i=2}^n w_i \chi_i^2, & 0 \\ 0, & \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i \end{bmatrix} \begin{bmatrix} m_k \\ b_k \end{bmatrix} = \begin{bmatrix} \sum_{i=2}^n w_i \chi_i \gamma_i \\ \sum_{i=2}^n w_i \gamma_i \end{bmatrix} \quad (20)$$

where we find that we have a cloud of parallel regression lines with identical slope

$$m_k = \frac{\sum_{i=2}^n w_i \chi_i \gamma_i}{\sum_{i=2}^n w_i \chi_i^2}$$

but different intercepts

$$b_k = \frac{\sum_{i=2}^n w_i \gamma_i}{\sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i}.$$

We can find that b is not free from k The figure below depicts the case when $\sum_{i=2}^n w_i \chi_i = 0$ holds for all equal weight for repeated and non-repeated data points. We amend the datasets as (6, 6), (2, 2), (5, 1), (7, 2), if we repeat the third data point (5, 1), as it satisfies the condition $6 + 2 + 7 = 5 * (4 - 1)$, we have a set of parallel regression lines without even a single intersection between any two regression line (Fig. 2).

2. Still, the weight scheme plays a latent but vital role in determining whether the pivot point exist. Actually, even the newly added repeated point affects the weights of non-repeated data points, provided that (19) keeps invariant, we can always find a pivot behavior. For example, if $w_i \propto k$ or $w_i = c_i/k$ for $i = 2, \dots, n$ we will still find a pivot point for regression lines.

Fig. 2 Repeat (5, 1) in (6, 6), (2, 2), (5, 1), (7, 2) dataset

2.2 Futher About Pivot Point

Let us trace back to the origin of pivot point, in order to simplify our calculation, we express the following calculation in R -centric coordinate. The normal equation under R -centric coordinate system is given below:

$$\begin{bmatrix} \sum_{i=2}^n w_i \chi_i^2, & \sum_{i=2}^n w_i \chi_i \\ \sum_{i=2}^n w_i \chi_i, & \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i \end{bmatrix} \begin{bmatrix} m_k \\ b_k \end{bmatrix} = \begin{bmatrix} \sum_{i=2}^n w_i \chi_i \gamma_i \\ \sum_{i=2}^n w_i \gamma_i \end{bmatrix} \quad (21)$$

If left hand side matrix $X^T W X$ is non-singular, we could solve this normal equation with specific repeated time k

$$\begin{aligned} m_k &= \frac{(\sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i)(\sum_{i=2}^n \chi_i \gamma_i w_i) - (\sum_{i=2}^n \chi_i w_i)(\sum_{i=2}^n \gamma_i w_i)}{(\sum_{i=2}^n \chi_i^2 w_i)(\sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i) - (\sum_{i=2}^n \chi_i w_i)^2} \\ b_k &= \frac{(\sum_{i=2}^n \chi_i^2 w_i)(\sum_{i=2}^n \gamma_i w_i) - (\sum_{i=2}^n \chi_i \gamma_i w_i)(\sum_{i=2}^n \chi_i w_i)}{(\sum_{i=2}^n \chi_i^2 w_i)(\sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i) - (\sum_{i=2}^n \chi_i w_i)^2} \end{aligned} \quad (22)$$

In order to control variations between weights, we assume that newly added repeated point will not affect the weights of $n + k - 1$ data points i.e. $w_{1,1}, \dots, w_{1,k}, w_2, \dots, w_n$ keep the same when $w_{1,k+1}$ added.

2.3 Results Under Some Certain Weight Schemes

The followings we discussed are results as we apply different weight scheme to datasets:

1. if $\lim_{k \rightarrow +\infty} \sum_{i=1}^{k+1} w_{1,i} = +\infty$ In this case we have

$$m = \lim_{k \rightarrow +\infty} m_k = \frac{\sum_{i=2}^n \chi_i \gamma_i w_i}{\sum_{i=2}^n \chi_i^2 w_i} \quad (23)$$

$$b = \lim_{k \rightarrow +\infty} b_k = 0 \quad (24)$$

Based on this result and (11), we can conclude that limiting position of regression line must go through origin (0, 0) and the pivot point (19). Actually, we can calculate the slope of limiting position of regression line directly through (19)

$$\frac{\sum_{i=2}^n w_i \chi_i \gamma_i}{\sum_{i=2}^n w_i \chi_i} / \frac{\sum_{i=2}^n w_i \chi_i^2}{\sum_{i=2}^n w_i \chi_i}$$

the result is accord with (23).

2. if $\lim_{k \rightarrow +\infty} \sum_{i=1}^{k+1} w_{1,i} = c \in \mathbb{R}$ where $c \geq 0$

$$\begin{aligned}
 m &= \lim_{k \rightarrow +\infty} m_k = \frac{(c + \sum_{i=2}^n w_i)(\sum_{i=2}^n x_i y_i w_i) - (\sum_{i=2}^n x_i w_i)(\sum_{i=2}^n y_i w_i)}{(\sum_{i=2}^n x_i^2 w_i)(c + \sum_{i=2}^n w_i) - (\sum_{i=2}^n x_i w_i)^2} \\
 b &= \lim_{k \rightarrow +\infty} b_k = \frac{(\sum_{i=2}^n x_i^2 w_i)(\sum_{i=2}^n y_i w_i) - (\sum_{i=2}^n x_i y_i w_i)(\sum_{i=2}^n x_i w_i)}{(\sum_{i=2}^n x_i^2 w_i)(c + \sum_{i=2}^n w_i) - (\sum_{i=2}^n x_i w_i)^2}
 \end{aligned} \quad (25)$$

Same as above condition, we can also use coordinate of R_{skew} ((13) converted to R -centric coordinate system first) and the pivot point P to calculate the slope of limiting position of regression line.

The figure below shows that we assign weight to k -th newly added (x_1, y_1) (total repeated 100 times) with weight $w_{1,k} = \frac{1}{2^k}$ where the remaining non-repeated points $(x_2, y_2), (x_3, y_3), (x_4, y_4)$ with invariant weights $w_2 = 2, w_3 = 3, w_4 = 4$ respectively. We can still find a pivot point for regression lines but will no longer approach to (x_1, y_1) no matter how many repeated times added, which could be compared to Fig. 1 even we only repeat (x_1, y_1) for 20 times (Fig. 3).

From the figure, we notice that the limiting position of regression line could not go through repeated point R as $k \rightarrow +\infty$ (or the limiting line may not go through origin $(0, 0)$ in R -centric coordinate system.) for we have sum of weights of repeated point as finite constant $c \in \mathbb{R}$ as $k \rightarrow +\infty$.

3. Each weight is identical and free from k , i.e. $w_{1,1} = w_{1,2} = \dots = w_{1,(k+1)} = w_2 = \dots = w_n = w \in \mathbb{R}^+$, which can be viewed as OLS regression and it is a special case for case 1, we have

$$m = \lim_{k \rightarrow +\infty} m_k = \frac{\sum_{i=2}^n x_i y_i}{\sum_{i=2}^n x_i^2} \quad (26)$$

Fig. 3 Repeat $(1, 6)$ in $(1, 6), (2, 2), (5, 1), (7, 2)$ for 100 times

$$b = \lim_{k \rightarrow +\infty} b_k = 0. \quad (27)$$

2.4 Explanations About Pivot Point

We denote $\tilde{l} = mx + b$ as the limiting position of these regression lines. I am going to show that for any given repeated time k , the line $\gamma_k = m_k \chi + b_k$ is a weighted average of $l_k : \tilde{l}$ and l_0 with weighted w_k , which is equivalent to prove:

$$\begin{aligned} b_k &= w_k b_0 + (1 - w_k) b \\ m_k &= w_k m_0 + (1 - w_k) m \end{aligned} \quad (28)$$

where

$$m_0 = \frac{(w_{1,1} + \sum_{i=2}^n w_i)(\sum_{i=2}^n \chi_i \gamma_i w_i) - (\sum_{i=2}^n \chi_i w_i)(\sum_{i=2}^n \gamma_i w_i)}{(\sum_{i=2}^n \chi_i^2 w_i)(w_{1,1} + \sum_{i=2}^n w_i) - (\sum_{i=2}^n \chi_i w_i)^2} \quad (29)$$

$$b_0 = \frac{(\sum_{i=2}^n \chi_i^2 w_i)(\sum_{i=2}^n \gamma_i w_i) - (\sum_{i=2}^n \chi_i \gamma_i w_i)(\sum_{i=2}^n \chi_i w_i)}{(\sum_{i=2}^n \chi_i^2 w_i)(w_{1,1} + \sum_{i=2}^n w_i) - (\sum_{i=2}^n \chi_i w_i)^2} \quad (30)$$

We have got the expression for m_k and b_k from (22) and m and b from (25) and (22).

We verified that:

$$w_k = \frac{(\sum_{i=2}^n \chi_i^2 w_i)(w_{1,1} + \sum_{i=2}^n w_i) - (\sum_{i=2}^n \chi_i w_i)^2}{(\sum_{i=2}^n \chi_i^2 w_i)(\sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i) - (\sum_{i=2}^n \chi_i w_i)^2} \quad (31)$$

holds (28) under case 1 in Sect. 2.3.

$$w_k = \frac{(\sum_{i=2}^n \chi_i^2 w_i)(c + \sum_{i=2}^n w_i) - (\sum_{i=2}^n \chi_i w_i)^2}{(\sum_{i=2}^n \chi_i^2 w_i)(\sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i) - (\sum_{i=2}^n \chi_i w_i)^2} \quad (32)$$

holds (28) under case 2 in Sect. 2.3.

That is, each l_k is a weighted average of l_0 and $\tilde{l}$. So the intersection between l_0 and $\tilde{l}$ must appear on each weighted line between l_0 and $\tilde{l}$, which is exactly the Pivot point we find.

3 The Existence of Pivot Hyperplane in Higher Dimensional Cases

In this section, we consider pivot behavior under multiple regression cases, (the number of regressors is m where $m \geq 1 \in \mathbb{N}^+$) Supposing we start with the n unique observations $(x_{11}, x_{12}, \dots, x_{1m}, y_1), (x_{21}, x_{22}, \dots, x_{2m}, y_2), \dots, (x_{n1}, x_{n2}, \dots, x_{nm}, y_n)$ and we consistently use $(x_{11}, x_{12}, \dots, x_{1m}, y_1)$ as repeated observations R for updated dataset. Note that we repeat $(x_{11}, x_{12}, \dots, x_{1m}, y_1)$ for additional k times so we have design matrix $X_k \in \mathbb{R}^{(n+k) \times m+1}$ as

$$k + 1 \text{ times} \left\{ \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ x_{11} & x_{12} & \dots & x_{1m} & 1 \\ x_{21} & x_{22} & \dots & x_{2m} & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} & 1 \end{bmatrix} \right\}_{(n+k) \times (m+1)}$$

and we have prior information about the weight matrix $W_k \in \mathbb{R}^{(n+k) \times (n+k)}$ as

$$W_k = \begin{bmatrix} w_{1,1} & \dots & 0 & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & w_{1,k+1} & 0 & \dots & 0 \\ 0 & \dots & 0 & w_2 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & \dots & w_n \end{bmatrix}_{(n+k) \times (n+k)}$$

$w_{1,1}, \dots, w_{1,k+1} \in \mathbb{R}^+$ are corresponding weights for total $k + 1$ repeated observations $(x_{11}, x_{12}, \dots, x_{1m}, y_1)$ respectively and $w_2, \dots, w_n \in \mathbb{R}^+$ are weights for non-repeated points. Same as we discussed in section 2 we should notice the weights $w_{1,1}, \dots, w_{1,k+1}$ for total $k + 1$ times repeated observations do not need to be identical and their sum $\sum_{i=1}^{k+1} w_{1,k+1} + \sum_{i=2}^n w_i$ do not need to add up to 1 under our parameterization.

Suppose the sample regression model under consideration can be written in the form

$$Y_k = X_k \beta_k + \epsilon_k \quad (33)$$

where in general we have

1. Y_k is an $(n + k) * 1$ vector of observations,

$$Y_k = \left\{ \begin{bmatrix} y_1 \\ \vdots \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right\}_{k+1 \text{ times}}$$

2. X_k is an $(n + k) * (m + 1)$ matrix of known form we mentioned above

$$X_k = k + 1 \text{ times } \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} & 1 \\ \vdots & \vdots & & \vdots & \\ x_{11} & x_{12} & \dots & x_{1m} & 1 \\ x_{21} & x_{22} & \dots & x_{2m} & 1 \\ \vdots & \vdots & & \vdots & \\ x_{n1} & x_{n2} & \dots & x_{nm} & 1 \end{bmatrix}_{(n+k) \times (m+1)}$$

3. β_k is a $(m+1) \times 1$ vector of parameters, $\beta_i^k, i = 1, 2, \dots, m$ is parameters for coefficient of regressors x_i , b_m is parameter for constant.

$$\beta_k = \begin{bmatrix} \beta_1^k \\ \beta_2^k \\ \vdots \\ \beta_m^k \\ b_k \end{bmatrix}_{(m+1) \times 1} \quad (34)$$

4. ϵ_k is an $(n+k) \times 1$ vectors of errors.

$$\epsilon_k \sim N(0, W_k \sigma^2) \quad (35)$$

To minimize residual sum of squares through weighted least square method, the normal equations become:

$$X_k^T W_k X_k \beta = X_k^T W_k Y_k \quad (36)$$

Using the same method in Sect. 2.1, normal equation 36 can be expressed under R -centric coordinate where :

$$\begin{bmatrix} \sum_{i=2}^n w_i \chi_{i1}^2 & \sum_{i=2}^n w_i \chi_{i1} \chi_{i2} & \dots & \sum_{i=2}^n w_i \chi_{i1} \chi_{im} & \sum_{i=2}^n w_i \chi_{i1} \\ \sum_{i=2}^n w_i \chi_{i1} \chi_{i2} & \sum_{i=2}^n w_i \chi_{i2}^2 & \dots & \sum_{i=2}^n w_i \chi_{i2} \chi_{im} & \sum_{i=2}^n w_i \chi_{i2} \\ \vdots & \vdots & & \vdots & \vdots \\ \sum_{i=2}^n w_i \chi_{i1} \chi_{im} & \sum_{i=2}^n w_i \chi_{i2} \chi_{im} & \dots & \sum_{i=2}^n w_i \chi_{im}^2 & \sum_{i=2}^n w_i \chi_{im} \\ \sum_{i=2}^n w_i \chi_{i1} & \sum_{i=2}^n w_i \chi_{i2} & \dots & \sum_{i=2}^n w_i \chi_{im} & \sum_{i=1}^{k+1} w_{1,i} + \sum_{i=2}^n w_i \end{bmatrix} \begin{bmatrix} \beta_1^k \\ \beta_2^k \\ \vdots \\ \beta_m^k \\ b_k \end{bmatrix} \\ = \begin{bmatrix} \sum_{i=2}^n w_i \chi_{i1} \chi_i \\ \sum_{i=2}^n w_i \chi_{i2} \chi_i \\ \vdots \\ \sum_{i=2}^n w_i \chi_{im} \chi_i \\ \sum_{i=2}^n w_i \chi_i \end{bmatrix} \quad (37)$$

where

$$\begin{aligned} 2 \leq i \in N^+ \leq n, 1 \leq j \in N^+ \leq m \\ \chi_{ij} = x_{ij} - x_{1j} \end{aligned} \quad (38)$$

$$\gamma_i = y_i - y_1 \quad (39)$$

From first m equations in 37, we notice that the WLS multiple regression lines

$$Y_k = X_k \beta_k \quad (40)$$

would include m specific points, namely:

$$\begin{aligned} P_1 &= \left(\sum_{i=2}^n w_i \chi_{i1}^2, \sum_{i=2}^n w_i \chi_{i1} \chi_{i2}, \dots, \sum_{i=2}^n w_i \chi_{i1} \chi_{im}, \sum_{i=2}^n w_i \chi_{i1} \right) \\ P_2 &= \left(\sum_{i=2}^n w_i \chi_{i1} \chi_{i2}, \sum_{i=2}^n w_i \chi_{i2}^2, \dots, \sum_{i=2}^n w_i \chi_{i2} \chi_{im}, \sum_{i=2}^n w_i \chi_{i2} \right) \\ &\vdots \\ P_m &= \left(\sum_{i=2}^n w_i \chi_{i1} \chi_{im}, \sum_{i=2}^n w_i \chi_{i2} \chi_{im}, \dots, \sum_{i=2}^n w_i \chi_{im}^2, \sum_{i=2}^n w_i \chi_{im} \right) \end{aligned}$$

in $m + 1$ dimensional space regardless any selection of repeated times k under R -centric coordinate.

Once the R -centric coordinate matrix has full rank, i.e.

$$\det \left(\begin{bmatrix} \sum_{i=2}^n w_i \chi_{i1}^2 & \sum_{i=2}^n w_i \chi_{i1} \chi_{i2} & \dots & \sum_{i=2}^n w_i \chi_{i1} \chi_{im} & \sum_{i=2}^n w_i \chi_{i1} \\ \sum_{i=2}^n w_i \chi_{i1} \chi_{i2} & \sum_{i=2}^n w_i \chi_{i2}^2 & \dots & \sum_{i=2}^n w_i \chi_{i2} \chi_{im} & \sum_{i=2}^n w_i \chi_{i2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sum_{i=2}^n w_i \chi_{i1} \chi_{im} & \sum_{i=2}^n w_i \chi_{i2} \chi_{im} & \dots & \sum_{i=2}^n w_i \chi_{im}^2 & \sum_{i=2}^n w_i \chi_{im} \end{bmatrix}_{m \times m} \right) \neq 0. \quad (41)$$

We can conclude that there exists a pivot hyperplane with dimension m in $m + 1$ variable space, which is consistent with our conclusion in 2.1.

4 Simulation Study of Pivot Behavior

In this section, I will try to find the pivot behavior in three dimensional data cases under equal weight scheme. Suppose we want to pass a plane $y = k_1 x_1 + k_2 x_2 + b$ through the points $\underbrace{(2, -1, 4) \dots (2, -1, 4)}_{k+1 \text{ times}}, (-1, 3, -2), (0, 2, 3), \text{ and } (-1, -2, 0)$ using

OLS method. We have total $k + 4$ data points, $k + 1$ of them are $(2, -1, 4)$ with 1 original point and k repetitions. We can select any single points to repeat, here for simplicity we repeat $(x_{11}, x_{21}, y_1) = (2, -1, 4)$. We centralize the whole datasets with respect to repeated point by setting $(x_{11}, x_{21}, y_1) = (2, -1, 4)$ as original point $(0, 0, 0)$, shifting in space, so the dataset turn out to be $\underbrace{(0, 0, 0) \dots (0, 0, 0)}_{k+1 \text{ times}}, (-3, 4, -6), (-2, 3, -1), \text{ and } (-3, -1, -4)$.

More generally, a point with Cartesian coordinates (x_{i1}, x_{i2}, y_i) has R -centric coordinates $\chi_{i1} = x_{i1} - x_{11}$, $\chi_{i2} = x_{i2} - x_{12}$, $v_i = y_i - y_1$. Fitting the plane $v = k_1\chi_1 + k_2\chi_2 + b$ to the data is equivalent to

$$k + 1 \text{ times } \begin{bmatrix} 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \\ -3 & 4 & 1 \\ -2 & 3 & 1 \\ -3 & -1 & 1 \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ b \end{bmatrix} + \epsilon = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ -6 \\ -1 \\ -4 \end{bmatrix}$$

with associated normal equation

$$\begin{bmatrix} \sum_{i=2}^4 \chi_{i1}^2 & \sum_{i=2}^4 \chi_{i1} \chi_{i2} & \sum_{i=2}^4 \chi_{i1} \\ \sum_{i=2}^4 \chi_{i1} \chi_{i2} & \sum_{i=2}^4 \chi_{i2}^2 & \sum_{i=2}^4 \chi_{i2} \\ \sum_{i=2}^4 \chi_{i1} \chi_{i1} & \sum_{i=2}^4 \chi_{i1} \chi_{i2} & k + n \end{bmatrix} \begin{bmatrix} k_1 \\ k_2 \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=2}^4 \chi_{i1} v_i \\ \sum_{i=2}^4 \chi_{i2} v_i \\ \sum_{i=2}^4 v_i \end{bmatrix}$$

By comparing the first 2 equations from the matrix equation above, we have

$$\sum_{i=2}^4 \chi_{i1}^2 k_1 + \sum_{i=2}^4 \chi_{i1} \chi_{i2} k_2 + \sum_{i=2}^4 \chi_{i1} b = \sum_{i=2}^4 \chi_{i1} v_i \quad (42)$$

$$\sum_{i=2}^4 \chi_{i1} \chi_{i2} k_1 + \sum_{i=2}^4 \chi_{i2}^2 k_2 + \sum_{i=2}^4 \chi_{i2} b = \sum_{i=2}^4 \chi_{i2} v_i \quad (43)$$

When $\sum_{i=2}^4 \chi_{i1} \neq 0$ and $\sum_{i=2}^4 \chi_{i2} \neq 0$, we can divide to find corresponding points on the plane $v = k_1\chi_1 + k_2\chi_2 + b$, namely

$$\left(\frac{\sum_{i=2}^4 \chi_{i1}^2}{\sum_{i=2}^4 \chi_{i1}}, \frac{\sum_{i=2}^4 \chi_{i1} \chi_{i2}}{\sum_{i=2}^4 \chi_{i1}}, \frac{\sum_{i=2}^4 \chi_{i1} v_i}{\sum_{i=2}^4 \chi_{i1}} \right) \quad (44)$$

$$\left(\frac{\sum_{i=2}^4 \chi_{i1} \chi_{i2}}{\sum_{i=2}^4 \chi_{i2}}, \frac{\sum_{i=2}^4 \chi_{i2}^2}{\sum_{i=2}^4 \chi_{i2}}, \frac{\sum_{i=2}^4 \chi_{i2} v_i}{\sum_{i=2}^4 \chi_{i2}} \right)$$

notice that (44) is absent from repeated times k , which means no matter how many times we repeated, the fitting plane should contain points (44). If (44) are not identical, we can easily find the pivot line goes through (44), which is given by symmetric equations of the line and the graph below provide an illustration for pivot behavior in three dimensional case (Fig. 4):

Fig. 4 The pivot behavior given by repeated $(2, -1, 4)$ $k = 10$ times

$$\begin{aligned}
 & x_1 - \frac{\sum_{i=2}^4 X_{i1}^2}{\sum_{i=2}^4 X_{i1}} \\
 & \frac{\sum_{i=2}^4 X_{i1}^2}{\sum_{i=2}^4 X_{i1}} - \frac{\sum_{i=2}^4 X_{i1} X_{i2}}{\sum_{i=2}^4 X_{i2}} \\
 & x_2 - \frac{\sum_{i=2}^4 X_{i1} X_{i2}}{\sum_{i=2}^4 X_{i1}} \\
 & = \frac{\sum_{i=2}^4 X_{i1} X_{i2}}{\sum_{i=2}^4 X_{i1}} - \frac{\sum_{i=2}^4 X_{i2}^2}{\sum_{i=2}^4 X_{i2}} \\
 & y - \frac{\sum_{i=2}^4 X_{i1} v_i}{\sum_{i=2}^4 X_{i1}} \\
 & = \frac{\sum_{i=2}^4 X_{i1} v_i}{\sum_{i=2}^4 X_{i1}} - \frac{\sum_{i=2}^4 X_{i2} v_i}{\sum_{i=2}^4 X_{i2}}
 \end{aligned} \tag{45}$$

5 Conclusion

We study the pivot behaviour in regression using weighted least square method and give a plausible explanation for it in this paper. As we can see, the weighting scheme plays a vital role throughout the whole study procedure. The dependency between weight scheme and pivot behaviour would be our future research topic, [8] could be a helpful tool for determining corresponding weights for observations and the efficiency loss for estimators in regression analysis under repeated observations could be estimated simulated using Monte Carlo method [9]. Meanwhile, the explanation of pivot behaviour could be applied in different scientific areas to give further explanations of impact coming from repeated data points according to the knowledge of

their specific domain. Meanwhile, whether longitudinal data would inherit this pivot behaviour, this could be another research topic.

Acknowledgements The authors would like to thank the Editor and the referee for careful reading and comments which greatly improved the paper.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Zhijiang W, Hao Z, Jiajin H, Haiyan Z, Jie Y, Zhong N (2019) Single-channel EEG-based machine learning method for prescreening major depressive disorder. *Int J Inf Technol Decis Mak* 18(05):1579–1603
2. Tang WM, Yiu KFC, Wong H (2020) Subset Selection Using Frequency Decomposition with Applications. *Int J Inf Technol Decis Mak* 19(01):195–220
3. Draper NR, Smith H (1998) *Applied regression analysis*, 3rd edn. Wiley, Hoboken
4. Lutzer CV (2017) A curious feature of regression. *College Math J* 48(3):189–198. <https://doi.org/10.4169/college.math.j.48.3.189>
5. Eugene D (2016) The p-value you cant buy? *Am Stat*. <https://doi.org/10.1080/00031305.2015.1069760>
6. Jan S, Shieh G (2019) Sample size calculations for model validation in linear regression analysis. *BMC Med Res Methodol* 19:54. <https://doi.org/10.1186/s12874-019-0697-9>
7. Austin Peter C, Steyerberg Ewout W (2015) The number of subjects per variable required in linear regression analyses. *J Clin Epidemiol*. <https://doi.org/10.1016/j.jclinepi.2014.12.014>
8. Jiahua C, Jun S (1993) Iterative weighted least squares estimators. *Ann Stat* 6(2):1071–1092. <https://doi.org/10.1214/aos/1176349165>
9. Francesco S, Malgorzata M (2017) Bias and efficiency loss in regression estimates due to duplicated observations: a Monte Carlo simulation. *Surv Res Methods*. <https://doi.org/10.18148/srm/2017.v1i1i.7149>
10. Radhakrishna Rao C, et al (2008) *Linear models and generalizations: least squares and alternatives*, Third Extended Edition, Springer, Berlin

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.