The SSR Plot: A Graphical Representation for Regression

CHONG SUN HONG\(^1\) AND JONG-MIN KIM\(^2\)

\(^1\)Department of Statistics, Sungkyunkwan University, Seoul, Korea
\(^2\)Statistics, Division of Science and Mathematics, University of Minnesota–Morris, Morris, Minnesota

An alternative graphical method, called the SSR plot, is proposed for use with a multiple regression model. The new method uses the fact that the sum of squares for regression (SSR) of two explanatory variables can be partitioned into the SSR of one variable and the increment in SSR due to the addition of the second variable. The SSR plot represents each explanatory variable as a vector in a half circle. Our proposed SSR plot explains that the explanatory variables corresponding to the vectors located closer to the horizontal axis have stronger effects on the response variable. Furthermore, for a regression model with two explanatory variables, the magnitude of the angle between two vectors can be used to identify suppression.

**Keywords** Coefficient of determination; Correlation; Geometry; SSR; Suppression.

**Mathematics Subject Classification** 62Jxx.

1. Introduction

The graphical method describing multiple regression models is a useful tool for statistics education. Authors like Box et al. (1978), Margolis (1979), Herr (1980), Draper and Smith (1981), Bryant (1984), and others have used graphical methods to describe regression models. Hamilton (1987, 1988) showed examples of the inadequacy of simple measures of association and \(x - y\) scatterplots in a multiple regression model while incorrectly stating that correlated variables are always redundant, \(R^2 \leq r_{x_1}^2 + r_{x_2}^2\). Hamilton (1987) derived the necessary and sufficient conditions for \(R^2 > r_{x_1}^2 + r_{x_2}^2\), as \(r_{x_1x_2} \{r_{x_1}^2 - 2r_{x_1x_2}r_{x_2}/(r_{x_1}^2 + r_{x_2}^2)\} > 0\). Mitra (1988), Freund (1988), and Schey (1993) also attempted to explain graphically the concept of suppression by comparing the relative magnitudes of the sums of squares for regression as well as correlation coefficients among response and explanatory variables. Cohen et al. (2003) defined the term suppression as a situation in which...
a relationship between independent variables is hiding or suppressing their real relationships with a dependent variable $Y$, which would be larger or possibly of opposite sign if the independent variable were not correlated. So the signs of the correlation coefficient and the regression coefficient differ when suppression is present. Sharpe and Roberts (1997), Friedman and Wall (2005), and Christensen (2006) also studied the conditions under which suppression can occur.

Gabriel (1971) introduced the biplot as a method for displaying the elements of a matrix as the inner products of vectors corresponding to the rows and columns of a matrix. Extending from this concept of the biplot, Corsten and Gabriel (1976) introduced the $h$ plot and the expression of correlation coefficients as angles. Gower and Hand (1996) have extended and generalized the ideas of Gabriel. Trosset (2005) later proposed a correlation diagram using the cosine function of the angles. The correlation diagram represents a correlation coefficient matrix using a set of points on a unit circle. Biplot approaches can simultaneously display genes and samples in low-dimensional graphs and thus can be used to represent the relationships between genes and samples. Pittelkow and Wilson (2005) developed the GE biplot (gene expression biplot), which is a method for exploring gene expression data that has the major advantage of being able to aid the interpretation of both the samples and the genes relative to each other. Park et al. (2008) compared the performance of a PCA (principal components analysis) biplot, FA (factor analysis) biplot, MDS (multidimensional scaling) biplot, and CA (correspondence analysis) biplot by analyzing various types of gene expression data.

Among the geometrical methods for expressing a multiple regression, Schey (1993) and others presented the relationship between the sums of squares for regression (SSR) as a cosine function. In this process, the correlation coefficients for the correlation diagram of Trosset (2005) are replaced with the SSR from the regression models, and an alternative graphical method that can be represented on a half circle is proposed to describe a multiple regression. This method, called the SSR plot, can be used to explore to what degree each explanatory variable explains a response variable in a multiple regression model. This is useful in that it can be used to show that some explanatory variables are more effective in explaining a response variable than others. Section 2 explains the relationship of the SSR plot with the correlation diagram of Trosset (2005) and its properties. Two SSR plots for multiple regression models are introduced, and the properties of these models using the SSR plots are explained.

When a regression model consisting of two explanatory variables is shown as an SSR plot, we find that the angle on the SSR plot is related with both the sum of squares for regression and the extra sum of squares for regression. The extra sum of squares for regression allows us to compare two models for the same response, where one model (the full model) contains all of the predictors in the other model (the reduced model) and more. The existence of suppression can be identified based on the angle in the SSR plot. In particular, one may determine whether or not suppression occurs based on the magnitude of the angle between two vectors corresponding to the two explanatory variables on the SSR plot. The main contribution of this article is that our proposed graphical method, the SSR plot, can be used to help undergraduate statistics students learn how to select variables for a multiple regression model. The most efficient way to teach elementary statistics in the classroom is the graphical approach. It is especially difficult for undergraduate students to understand the concepts of multivariate statistics, such as multiple...
regression correlation coefficient or principal components analysis. If statistics professors use the SSR to explain multivariate data graphically to undergraduate students, then students will experience fewer difficulties in understanding how to select variables for multiple regression models. So in Sec. 3 we describe three regression models and the corresponding SSR plots, and we discuss the suppression with these SSR plots. In Sec. 4 we offer some conclusions.

2. The SSR Plot

Consider a multiple regression model with \( p \) explanatory variables: 

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n.
\]

In statistics, correlation diagrams are generally employed to visualize any correlations that may occur between two sets of observations. The correlation diagram of Trosset (2005) is used to visualize a \( p \times p \) matrix of correlation coefficients \( \{ R \} = (r_{ij}) \) on a unit circle. The vectors correspond to \( p \) points on the unit circle and are associated with the scalar angles \( \theta_1, \theta_2, \ldots, \theta_p \). The angles between pairs of vectors correspond to the correlation of these pairs of vectors. The construction seeks to find \( \Theta = (\theta_1, \theta_2, \ldots, \theta_p) \) for which

\[
\cos(\theta_i - \theta_j) \approx r_{ij}, \quad \text{for all } i \text{ and } j.
\]

Thus, one solves an unconstrained optimization problem for the following objective function:

\[
\min 2 \sum_{i<j} \left[ \cos(\theta_i - \theta_j) - r_{ij} \right]^2.
\] (2.1)


The sum of squares for the regression (SSR) on two explanatory variables \( X_i \) and \( X_j \), SSR\((X_i, X_j)\), is partitioned as SSR\((X_i, X_j)\) = SSR\((X_i)\) + SSR\((X_j | X_i)\), where SSR\((X_i)\) is the SSR on \( X_i \) alone, and SSR\((X_j | X_i)\) is the increment in the SSR resulting from the addition of \( X_j \) to the model already containing \( X_i \); it is also called the extra SSR explained by \( X_j \) given \( X_i \).

From this partition, we create the right triangle in Fig. 1, and one can derive the following equation:

\[
\cos(\theta_i - \theta_j) = \frac{\text{SSR}(X_i)}{\sqrt{\text{SSR}(X_i, X_j)}} \quad \text{for all } i \text{ and } j.
\] (2.2)
Since Eq. (2.2) has a positive value in (0, 1), \(|\theta_i - \theta_j|\) belongs to [0, \pi/2]. For \(0 < \theta_i - \theta_j < \pi/2\), when the angle between the two vectors gets larger, \(\cos(\theta_i - \theta_j)\) becomes smaller. This implies that \(\text{SSR}(X_j | X_i)\) is relatively larger than \(\text{SSR}(X_i)\) for \(0 < \theta_i - \theta_j < \pi/2\). For example, if \(\theta_i - \theta_j = \pi/3\), then \(\cos(\theta_i - \theta_j)^2 = 0.25\). So using Eq. (2.2),

\[
\cos(\theta_i - \theta_j)^2 = \frac{\text{SSR}(X_i)}{\text{SSR}(X_i, X_j)} = \frac{\text{SSR}(X_i)}{\text{SSR}(X_i) + \text{SSR}(X_j | X_i)} = \frac{1}{4},
\]

which implies \(3 \times \text{SSR}(X_i) = \text{SSR}(X_j | X_i)\). One obtains that

\[
\text{SSR}(X_i) < \text{SSR}(X_j | X_i).
\]

Using similar arguments, such as a correlation coefficient matrix for a correlation diagram, one can use an SSR matrix whose \((i, j)\) cell has a value of \(\text{SSR}(X_j | X_i)/\text{SSR}(X_i, X_j)\). From this SSR matrix, the following unconstrained optimization problem may be constructed:

\[
\min \sum_{i=1}^{p} \sum_{j=1}^{p} \left[ \cos(\theta_i - \theta_j) - \sqrt{\frac{\text{SSR}(X_i)}{\text{SSR}(X_i, X_j)}} \right]^2.
\] (2.3)

Then \(\Theta = (\theta_1, \theta_2, \ldots, \theta_p)\) can be obtained by using \textit{nlminb} function in R program as is solved in the correlation diagram.

The correlation coefficient matrix used in Trosset’s correlation diagram is symmetric, but the SSR matrix described above is asymmetric. The angle differences \(\theta_i - \theta_j\) and \(\theta_i - \theta_j\) are related to the square roots of \(\text{SSR}(X_j | X_i)/\text{SSR}(X_i, X_j)\) and \(\text{SSR}(X_i)/\text{SSR}(X_i, X_j)\), respectively. It is found that the absolute difference of the angles \(|\theta_i - \theta_j|\) is related to the mean of the square roots of \(\text{SSR}(X_j | X_i)/\text{SSR}(X_i, X_j)\) and \(\text{SSR}(X_i)/\text{SSR}(X_i, X_j)\). The solution \(\Theta = (\theta_1, \theta_2, \ldots, \theta_p)\) converges and is obtained uniquely because the computation of our proposed graphical representation method is originally from S-Plus function \textit{nlminb}, a quasi-Newtonian algorithm that Trosset (2005) had provided. Our R program is available from the corresponding author upon request.

Working with the values of \(\Theta = (\theta_1, \theta_2, \ldots, \theta_p)\), we now describe our proposed graphical representation method called the SSR plot. Since the range of Pearson’s

<table>
<thead>
<tr>
<th>(X_1)</th>
<th>(X_2)</th>
<th>(X_3)</th>
<th>(X_4)</th>
<th>(X_5)</th>
<th>(X_6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.265824</td>
<td>0.549388</td>
<td>0.509235</td>
<td>0.03592</td>
<td>0.035252</td>
</tr>
<tr>
<td>0.997438</td>
<td>1</td>
<td>0.954508</td>
<td>0.912877</td>
<td>0.128435</td>
<td>0.132398</td>
</tr>
<tr>
<td>0.962289</td>
<td>0.445568</td>
<td>1</td>
<td>0.772768</td>
<td>0.061778</td>
<td>0.055626</td>
</tr>
<tr>
<td>0.996223</td>
<td>0.475947</td>
<td>0.8631</td>
<td>1</td>
<td>0.069263</td>
<td>0.060335</td>
</tr>
<tr>
<td>0.999997</td>
<td>0.952906</td>
<td>0.981894</td>
<td>0.985642</td>
<td>1</td>
<td>0.636078</td>
</tr>
<tr>
<td>0.998584</td>
<td>0.999513</td>
<td>0.899599</td>
<td>0.873627</td>
<td>0.647218</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 2
SSR matrix of the soil characteristic data

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$X_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>1</td>
<td>0.993316</td>
<td>0.78401</td>
<td>0.953897</td>
<td>0.705525</td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.017639</td>
<td>1</td>
<td>0.064671</td>
<td>0.112421</td>
<td>0.640963</td>
</tr>
<tr>
<td>$X_3$</td>
<td>0.199254</td>
<td>0.92555</td>
<td>1</td>
<td>0.995742</td>
<td>0.939024</td>
</tr>
<tr>
<td>$X_4$</td>
<td>0.953897</td>
<td>0.910289</td>
<td>0.563365</td>
<td>1</td>
<td>0.906646</td>
</tr>
<tr>
<td>$X_5$</td>
<td>0.705525</td>
<td>0.985342</td>
<td>0.100865</td>
<td>0.172132</td>
<td>1</td>
</tr>
</tbody>
</table>

correlation coefficients is between $-1$ and $+1$, a correlation diagram by Trosset (2005) is shown as vectors in a unit circle, where the scalar angle of each vector belongs to an interval $[0, 2\pi]$. By contrast, the SSR plot can be represented not on the unit circle but instead on a half circle because the scalar angle of each vector belongs to an interval $[-\pi/2, \pi/2]$. In a half circle, if the scalar angle of each vector belongs to an interval $[-\pi/2, 0]$, then we use an upper quarter circle for the SSR plot. Otherwise, we use either a lower quarter circle for the SSR plot.

In Tables 1 and 2, all the values of the SSR matrix lie between 0 and $+1$, so that the vector's scalar angle of each explanatory variable belongs to either the interval $[-\pi/2, 0]$ for the SSR plot of an upper quarter circle (Fig. 2) or the interval $[0, \pi/2]$ for the SSR plot of a lower quarter circle (Fig. 3).

In particular, one chooses a vector (explanatory variable) that is the most effective in explaining a response variable. The corresponding angle of the vector might be set to 0. For example, if $X_k$ is the variable whose correlation with the response variable has the largest absolute value and the determination coefficient of $X_k$ is the largest among the simple regression models, then the corresponding $\theta_k$ is set to 0, so that the $k$th vector points along the horizontal axis. (Note that one sets $\theta_1 = 0$ in order to remove rotational indeterminacy in the correlation diagram.)

Consider the supervisor performance data of Chatterjee et al. (2000, 53) and the soil characteristic data of Rawlings et al. (1998, 163). These data sets consist of six and five explanatory variables, respectively. Working from these data sets,

![Figure 2. SSR plot of the first data.](image)
For the first data set in Table 1, since $X_1$ has the largest correlation coefficient with a response variable, we set $\theta_1 = 0$. For the second data set in Table 2, we set $\theta_2 = 0$.

From Fig. 2, six explanatory variables can be classified into three clusters: the first cluster, which is close to the horizontal axis, contains three vectors corresponding to $X_1, X_3, X_4$; the second cluster contains only the one vector $X_2$; and the third has two vectors, $X_5, X_6$. One might conclude that the explanatory variables $X_1, X_3, X_4$, including the first cluster, explain a response variable better than any others. From a more generous viewpoint, the six variables can be classified into two clusters: the first cluster contains four vectors corresponding to $X_1, X_2, X_3, X_4$ and the second cluster contains $X_5, X_6$. From this view, the conclusion would be that the explanatory variables $X_1, X_2, X_3, X_4$, including the first cluster, explain a response variable better than any others.

The five variables in Fig. 3 can be classified into two clusters: one cluster close to the horizontal axis that contains two vectors corresponding to $X_2, X_5$ and the other cluster of vectors corresponding to $X_1, X_3, X_4$. One can conclude that the explanatory variables $X_2, X_5$ in the first cluster explain a response variable better than any others in the multiple regression models. The corresponding objective function values, which minimize (2.3) for the two data sets, are equal to 2.7440 and 1.6131, respectively.

Therefore, a regression model with $p$ explanatory variables can be visualized on a right half circle, called the SSR plot, containing $p$ vectors with scalar angles $= (\theta_1, \theta_2, \ldots, \theta_p)$. This SSR plot is designed so that the vectors that more effectively explain a response variable are located closer to the horizontal axis.

3. Suppression

To use the SSR plot for discussing suppression, we consider a regression model with only two explanatory variables. For an asymmetric SSR matrix discussed in Sec. 2, $\cos(\theta_1 - \theta_2) = \sqrt{\text{SSR}(X_1)/\text{SSR}(X_i, X_j)}$ of the $(1, 2)$ cell does not have the
same value as $\cos(\theta_2 - \theta_1) = \sqrt{\text{SSR}(X_2)/\text{SSR}(X_1, X_2)}$ of the $(2, 1)$ cell. However, for
the $(1, 2)$ and $(2, 1)$ cells in the SSR matrix, one obtains the following equation:

$$\cos^2(\theta_1 - \theta_2) + \cos^2(\theta_2 - \theta_1) = \frac{\text{SSR}(X_1) + \text{SSR}(X_2)}{\text{SSR}(X_1, X_2)}$$

$$= 1 - \frac{\text{SSR}(X_1|X_2) - \text{SSR}(X_2)}{\text{SSR}(X_1, X_2)}. \quad (3.1)$$

If the two explanatory variables $X_1$ and $X_2$ are independent, it is well known that
$\text{SSR}(X_2) = \text{SSR}(X_1|X_1)$ and $\text{SSR}(X_1, X_2) = \text{SSR}(X_1) + \text{SSR}(X_2)$. Then one
can get $\cos^2(\theta_1 - \theta_2) + \cos^2(\theta_2 - \theta_1) = 1$ and $|\theta_1 - \theta_2| = \pi/4$ in the SSR plot.
Hence, one can conclude that the angle between the two vectors must be 45 degrees
if the two explanatory variables are uncorrelated.

The variable $X_2$ is defined as a suppressor variable if

$$\text{SSR}(X_2|X_1) > \text{SSR}(X_2) \quad (3.2)$$

or

$$R^2 > r_{12}^2 + r_{13}^2. \quad (3.3)$$

where $R^2$ is the determination coefficient, and $r_{12}^2$ and $r_{13}^2$ are simple correlation
coefficients with a response $Y$. (See Lynn, 2003; Hamilton, 1987; Schey, 1993;
Sharpe and Roberts, 1997; Velicer, 1978 for more detail.) If suppression occurs while
satisfying (3.2), Eq. (3.1) has a value less than 1, and $|\theta_1 - \theta_2|$ has a value greater
than $\pi/4$ in the SSR plot. Therefore, when the SSR plot is explored for regression
data with two explanatory variables, one may conclude that suppression occurs if
the angle between the two vectors is greater than 45 degrees.

Schey (1993) provided several tables of contrived data sets illustrating the
relation between the magnitudes between $\text{SSR}(X_2)$ and $\text{SSR}(X_2|X_1)$ in various cases
for regression of $Y$ on $X_1$ and $X_2$. In this section, we explore the SSR plots from the
data in Tables 1(a)–(c) of Schey (1993, 29) and represented in Figs. 4–6, respectively.
There are two real lines on each SSR plot, one is duplicated with
a horizontal line. The other real line represents a vector corresponding to a less
effective explanatory variable. The dotted line in Figs. 4–6 indicates a reference line
with an angle of 45 degrees.

Figure 4 is drawn for the data in Table 1(a) of Schey (1993), which is the case
where suppression does not occur. One can find that the angle between the two
vectors is 33.2 degrees. Moreover, with values of the determination and correlation
coefficients, we can obtain the following inequality:

$$R^2 = 0.653 < r_{12}^2 + r_{13}^2 = .780^2 + .571^2 = .934.$$

Hence, one might conclude that suppression does not occur in this case since, from
Fig. 4, the angle between the two vectors is less than 45 degrees. Note that the
objective function that minimizes (2.3) has a value of 0.0335.

Figure 5 is based on Table 1(b) of Schey (1993), where a correlation does not
exist between the two explanatory variables. The angle between the two vectors
Figure 4. General case.

is found to be 46.9 degrees, which is much closer to 45 degrees. Thus, one may conclude that the two explanatory variables are independent. Note that

$$R^2 = 0.792 \approx r_{y1}^2 + r_{y2}^2 = .770^2 + .444^2 = .790.$$

and the objective function that minimizes (2.3) has a value of 0.0668.

Figure 6 is for Table 1(c) of Schey (1993), which is a case of suppression. The angle between the two vectors is 55.8 degrees and

$$R^2 = 0.9637 > r_{y1}^2 + r_{y2}^2 = .850^2 + .254^2 = .787.$$

Hence, one might determine that suppression occurs, since, from Fig. 6, the angle between the two vectors is greater than 45 degrees. Note that the objective function that minimizes (2.3) has a value of 0.1843. Based on the inequality (3.2), it is clear that $X_2$ is a suppressor variable. Moreover, Fig. 6 tells us that the vector of $X_1$ lies along the horizontal axis and that $X_2$ is located far away from it. This means that $X_2$ is less effective than $X_1$ in explaining the response variable.

Figure 5. Independent case.
From Table 1 in Hamilton (1987, 131), we can see that $R^2 = 0.9998$ is much larger than $r_{11}^2 + r_{12}^2 = .002^2 + .434^2 = .188$, so suppression occurs. If the SSR plot is used with this data, one can find that the angle between the two vectors is 77.4 degrees, which is much greater than 45 degrees, the vector $X_2$ lies along the horizontal axis, and $X_1$ is located far away from it. Therefore, the suppressor variable is found to be $X_1$. These examples show how the SSR plot can help identify whether suppression occurs and which variable is the suppressor.

4. Conclusion

The SSR plot proposed in this paper, is an alternative geometrical method for exploring a multiple regression. This method uses ideas from the correlation diagram of Trosset (2005) and properties of the sum of squares for regression. The SSR plot is drawn on a right half circle. Each explanatory variable is represented as a vector on the circle after the optimization problem (2.3) is solved. This plot indicates that the explanatory variable corresponding to the vector on a horizontal axis is the most effective in explaining a response variable. Explanatory variables corresponding to vectors located close to a horizontal axis describe the regression data better than others.

In particular, when an SSR plot is used to explore a regression with two explanatory variables, the magnitude of the angle between the two vectors can be identified to indicate whether suppression occurs. If the angle is greater than 45 degrees, then one may conclude that suppression occurs, and the variable corresponding to the vector that does not lie along the horizontal axis is identified as a suppressor variable.

Acknowledgment

This article was supported by Samsung Research Fund, Sungkyunkwan University, 2009.
The SSR Plot

References


Schey, H. M. (1993). The relationship between the magnitudes of $SSR(x_2)$ and $SSR(x_2 | t_{x_1})$: A geometric description. *Am. Stat.* 47:26–3.0

