Multiple R **IS** the Square Root of R^2: Multiple Correlation Coefficient Using Matrix Formulation

T. Mark Beasley
University of Alabama at Birmingham
Birmingham/Atlanta VA Geriatric Research, Education, & Clinical Center

By substituting the matrices necessary to compute Mean Corrected Sums of Squares and Cross-Products into the Pearson correlation scalar formula and utilizing the idempotent properties of partitioned matrices, it is demonstrated that computationally the Multiple R IS the square root of the Model R^2.

While preparing for my preliminary examinations to qualify as a doctoral candidate, I worked through many practice questions. One such question was: "Why is the Multiple Correlation Coefficient, R, always positive?" I was told by fellow students that: "The Multiple Correlation Coefficient is the square root of the Multiple Coefficient of Determination, $R = \sqrt{R^2}$." was not a "good enough" answer. Something more elaborate was needed. A more verbose response would be as follows.

> The Multiple Correlation Coefficient can be defined as the Pearson Correlation between the outcome variable, y, and the predicted values, $\hat{y}$, from regressing y onto a set of x variables. The predicted values, $\hat{y}$, determined from the Ordinary Least Squares (OLS) regression solution are calculated as: $\hat{y} = b_0 + b_1x_1 + \ldots + b_kx_k$. The OLS regression solution determines regression coefficients and predicted values that minimize the Sum of Squared Residuals ($\Sigma(y-\hat{y})^2$), thus yielding the best fit of $\hat{y}$ for y. Therefore, the OLS regression solutions will project $\hat{y}$ into the same space as y. For example, suppose a simple linear regression with one regressor, x_1. If x_1 is negatively (inversely) correlated with y, the regression slope for x_1 (b_1) will be negative. This can be seen in the scalar formula for the slope in simple regression: $b_1 = r_{y1}(S_y/S_1)$; where r_{y1} is the Pearson correlation between y and x_1; S_y and S_1 are the standard deviations of y and x_1, respectively. When the predicted values are calculated, the negative slope reverses x_1 and projects $\hat{y}$ into the same direction as y. In contrast to the negative correlation between y and x_1, the Pearson Correlation between y, and the predicted values, $\hat{y}$, $R_{y\hat{y}}$, will be positive. This property generalizes to multiple regression, where x variables that have negative partial relationships with y, will have negative partial regression coefficients that will project $\hat{y}$ into the same space (direction) as y. Thus, the Multiple Correlation Coefficient is always greater than or equal to zero ($R_{y\hat{y}} \geq 0$).

As wordy and potentially convincing as this may seem, a more mathematical approach using the matrix formulation of multiple regression is used to demonstrate why the Multiple Correlation Coefficient is always positive ($R_{y\hat{y}} \geq 0$) and **IS** equal to the square root of R^2 computationally. Table 1 reports the data for a basic $k = 4$ x variable regression problem with a small sample size of N=10. This sample size to variables ratio is not recommended in applied statistical practice, but rather these data are intended to provide concrete illustrations. In the notation that follows, lower-case bold font denotes vectors or variables (e.g., $\mathbf{x}_1$); upper-case bold font denotes matrices (e.g., $\mathbf{X}_1$; $\mathbf{H}_0$), and italics are used for other statistical terms (r_{y1}).

Matrix Approach to Pearson Correlation

The Pearson Correlation can be defined in many ways. For the following illustrations, the Pearson Correlation will be calculated as the Mean Corrected Cross-Product (numerator) in ratio to the square root of Mean Corrected Sums of Squares for each variable (denominator). As an example, the Pearson Correlation between $\mathbf{y}$ and $\mathbf{x}_1$ yields the following scalar formula:

$$r_{y1} = \frac{\Sigma(\mathbf{y}-\bar{y})(\mathbf{x}_1-\bar{x}_1)}{\sqrt{[\Sigma(\mathbf{y}-\bar{y})^2]}\sqrt{[\Sigma(\mathbf{x}_1-\bar{x}_1)^2}} \quad ; \tag{1}$$

where $\bar{y}$ and $\bar{x}_1$ are means for $\mathbf{y}$ and $\mathbf{x}_1$, respectively.

To Mean Correct the Sums of Squares and Cross-Products, suppose fitting an "intercept-only" regression (i.e., null) model by using an Nx1 vector of ones, $\mathbf{x}_0$, as the design matrix. Under OLS estimation, the intercept (b_0) is simply the mean of $\mathbf{y}$ ($\bar{y}$):

$$\begin{aligned} \mathbf{y} &= b_0\mathbf{x}_0 + \boldsymbol{e}_y \\ \mathbf{y} &= \hat{\mathbf{y}} + \boldsymbol{e}_y \\ \mathbf{y} &= \bar{y} + \boldsymbol{e}_y\,; \end{aligned} \tag{2}$$

which in this case is $\bar{y}$ = 19 (see Table 1).

In scalar notation, the residuals for **y** removing the effect of the mean of **y** (i.e., mean centered deviations) are computed as:

$$\begin{aligned} \boldsymbol{e}_y &= \mathbf{y} - b_0\mathbf{x}_0 \\ \boldsymbol{e}_y &= \mathbf{y} - \hat{\boldsymbol{y}} \qquad (3) \\ \boldsymbol{e}_y &= \mathbf{y} - \bar{y}\ . \end{aligned}$$

Based on the normal equation solution, the OLS regression coefficient in equation (2) is solved as:

$$b_0 = (\mathbf{x_0}'\mathbf{x_0})^{-1}\mathbf{x_0}'\mathbf{y}\ . \qquad (4)$$

The predicted values for model (2) can be solved as:

$$\hat{\boldsymbol{y}} = \mathbf{x}_0 b_0\ .$$

Substituting equation (4):

$$\hat{\boldsymbol{y}} = \mathbf{x}_0(\mathbf{x}_0'\mathbf{x}_0)^{-1}\mathbf{x}_0'\mathbf{y}\ . \qquad (5)$$

The Hat matrix for $\mathbf{x}_0$ is defined as:

$$\mathbf{H_0} = \mathbf{x}_0(\mathbf{x}_0'\mathbf{x}_0)^{-1}\mathbf{x}_0'\ , \qquad (6)$$

which is an NxN matrix with all values equaling $1/N$ (i.e., 1/10). Therefore by substituting (6), the predicted values can be represented as:

$$\hat{\boldsymbol{y}} = \bar{\boldsymbol{y}} = \mathbf{H}_0\mathbf{y}\ , \qquad (7)$$

which is equal to Nx1 vector where every value equals $\bar{y}$. The residuals ($\boldsymbol{e}_y$) are computed as:

$$\boldsymbol{e}_y = \mathbf{y} - \hat{\boldsymbol{y}}\ .$$

Substituting equation (7):

$$\boldsymbol{e}_y = \mathbf{y} - \mathbf{H}_0\mathbf{y} \qquad (8)$$

and using matrix algebra manipulation:

$$\boldsymbol{e}_y = (\mathbf{I} - \mathbf{H}_0)\mathbf{y}\ . \qquad (9)$$

where **I** is an N-dimensional Identity Matrix. Using this formulation, the Mean Corrected Sums of Squares (CSS) for y is equal to:

$$\mathrm{CSS}_y = \sum(\boldsymbol{y} - \bar{y})^2 = \sum \boldsymbol{e}_y^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H}_0)'(\mathbf{I} - \mathbf{H}_0)\mathbf{y}\ . \qquad (10)$$

Both **I** and $\mathbf{H}_0$ are symmetric and idempotent. A symmetric matrix has the property that it is equal to its transpose; $\mathbf{A}' = \mathbf{A}$. An idempotent matrix has the property that it equals its square: $\mathbf{A}^2 = \mathbf{AA} = \mathbf{A}$. From these properties, $(\mathbf{I} - \mathbf{H}_0)$ is also symmetric and idempotent:

$$(\mathbf{I} - \mathbf{H}_0)'(\mathbf{I} - \mathbf{H}_0) = \mathbf{II} - 2\mathbf{I}\,\mathbf{H}_0 + \mathbf{H}_0\mathbf{H}_0 = \mathbf{I} - 2\mathbf{H}_0 + \mathbf{H}_0 = (\mathbf{I} - \mathbf{H}_0)\ . \qquad (11)$$

Thus, equation (10) reduces to:

$$\mathrm{CSS}_y = \sum(\boldsymbol{y} - \bar{y})^2 = \sum \boldsymbol{e}_y^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H}_0)\mathbf{y}\ . \qquad (12)$$

Similarly, the mean centered deviations for x_1 can be defined as:

$$\boldsymbol{e}_\mathbf{1} = (\mathbf{I} - \mathbf{H}_0)\mathbf{x}_1\ . \qquad (13)$$

with Mean Corrected Sums of Squares equal to:

$$\mathrm{CSS}_1 = \sum(\mathbf{x}_1 - \bar{x}_1)^2 = \sum \boldsymbol{e}_1^2 = \mathbf{x}_1'(\mathbf{I} - \mathbf{H}_0)\,\mathbf{x}_1\ . \qquad (14)$$

Substituting the expressions (9), (12), (13), and (14) into equation (1) yields:

$$\begin{aligned} R_{Y.1} &= r_{y\hat{y}_1} = \frac{\boldsymbol{e}_y{}'\boldsymbol{e}_{\hat{y}_1}}{\sqrt{\mathrm{CSS}_y}\sqrt{\mathrm{CSS}_1}} \\ R_{Y.1} &= r_{y1} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H}_0)'(\mathbf{I}-\mathbf{H}_0)\mathbf{x}_1}{\sqrt{[\mathbf{y}'(\mathbf{I}-\mathbf{H}_0)\mathbf{y}][\mathbf{x}_1'(\mathbf{I}-\mathbf{H}_0)\mathbf{x}_1]}}\ . \qquad (15) \end{aligned}$$

With $(\mathbf{I} - \mathbf{H}_0)$ being symmetric and idempotent (see eq. 11), equation (15) reduces to:

$$R_{Y.1} = r_{y1} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H}_0)'\mathbf{x}_1}{\sqrt{[\mathbf{y}'(\mathbf{I}-\mathbf{H}_0)\mathbf{y}][\mathbf{x}_1'(\mathbf{I}-\mathbf{H}_0)\mathbf{x}_1]}}\ . \qquad (16)$$

The output from SAS® PROC CORR in Table 1 shows that the bivariate Pearson correlation between y and x_1 is r_{y1} = -0.15301. It also shows that the Mean Corrected Sums of Squares for y and x_1 are $\mathrm{CSS}_y = \sum(\boldsymbol{y} - \bar{y})^2 = \mathbf{y}'(\mathbf{I}\text{-}\mathbf{H}_0)\mathbf{y}$ =136 and $\mathrm{CSS}_1 = \sum(\mathbf{x}_1 - \bar{x}_1)^2 = \mathbf{x}_1'(\mathbf{I}\text{-}\mathbf{H}_0)\mathbf{x}_1$ = 38, respectively. The Mean Corrected Cross-Product between y and x_1 is $\mathrm{CCP}_{y1} = \sum(\mathbf{y} - \bar{y})(\mathbf{x}_1 - \bar{x}_1) = \mathbf{y}'(\mathbf{I}\text{-}\mathbf{H}_0)\mathbf{x}_1$ = -11. Thus, the Pearson correlation between **y** and $\mathbf{x}_1$ using either equation (1) or (16) is $r_{y1} = -11/(\sqrt{(136)(38)}$ = -0.15301.

Table 1. Data and Descriptive Statistics (Mean, SD, Mean Corrected Sums of Squares and Cross-Products and Pearson Correlation Matrices. Modified output from SAS® PROC CORR).

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
x1	10	14	2.05480	140	11	17
x2	10	14	2.40370	140	10	17
x3	10	21	1.82574	210	18	24
x4	10	16	2.35702	160	12	19
y	10	19	3.88730	190		

CSSCP Matrix

	x1	x2	x3	x4	y
x1	**38**	26	-1	0	**-11**
x2	26	52	-3	22	19
x3	-1	-3	30	18	14
x4	0	22	18	50	38
y	**-11**	19	14	38	**136**

Pearson Correlation Coefficients, N = 10
Prob > |r| under H0: Rho=0

	x1	x2	x3	x4	y
x1	1.00000	0.58490	-0.02962	0.00000	**-0.15301**
		0.0757	0.9353	1.0000	0.6730
x2	0.58490	1.00000	-0.07596	0.43146	0.22593
	0.0757		0.8348	0.2131	0.5302
x3	-0.02962	-0.07596	1.00000	0.46476	0.21918
	0.9353	0.8348		0.1759	0.5429
x4	0.00000	0.43146	0.46476	1.00000	0.46082
	1.0000	0.2131	0.1759		0.1801
y	-0.15301	0.22593	0.21918	0.46082	1.00000
	0.6730	0.5302	0.5429	0.1801	

```
data R2;
input x0   x1    x2   x3     x4     y;
cards;
      1    14    10    20    12    16
      1    11    12    19    14    15
      1    15    13    22    17    12
      1    15    16    21    13    17
      1    17    17    23    18    19
      1    13    16    20    19    21
      1    14    15    22    18    22
      1    17    16    18    15    21
      1    12    14    21    17    23
      1    12    11    24    17    24
;proc corr data=R2 csscp;var x1 x2 x3 x4 y;run;
```

Simple Linear Regression

Suppose regressing $\mathbf{y}$ on to $\mathbf{x}_1$ with the $\mathbf{x}_0$ vector of ones included to estimate an intercept. The Nx2 design matrix is $\mathbf{X}_1 = \mathbf{x}_0|\mathbf{x}_1$; where the | symbol represent horizontal concatenation of the $\mathbf{x}_0$ and $\mathbf{x}_1$ vectors. The regression model is:

$$\begin{aligned}\mathbf{y} &= \mathbf{X}_1\mathbf{b}_1 + \boldsymbol{e}_{y.1}\\ \mathbf{y} &= b_0\mathbf{x}_0 + b_1\mathbf{x}_1 + \boldsymbol{e}_{y.1}\ ,\end{aligned} \tag{17}$$

where $\boldsymbol{e}_{y.1}$ are the residuals for y removing the effect of $\mathbf{x}_1$.

The OLS solution for the regression coefficients is:

$$\mathbf{b}_1 = (\mathbf{X}_1{}'\mathbf{X}_1)^{-1}\mathbf{X}_1{}'\mathbf{y} = \begin{matrix} b_0 \\ b_1 \end{matrix}\ . \tag{18}$$

The predicted values for model (17) are solved as:

$$\hat{\boldsymbol{y}}_1 = \mathbf{X}_1\mathbf{b}_1\ . \tag{19}$$

Substituting equation (18):

$$\hat{\boldsymbol{y}}_1 = \mathbf{X}_1(\mathbf{X}_1{}'\mathbf{X}_1)^{-1}\mathbf{X}_1{}'\mathbf{y}\ . \tag{20}$$

The Hat Matrix for $\mathbf{X}_1$ is formed as:

$$\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1{}'\mathbf{X}_1)^{-1}\mathbf{X}_1{}'. \tag{21}$$

Substituting equation (21):

$$\hat{\boldsymbol{y}}_1 = \mathbf{H}_1\mathbf{y}\ .$$

The residuals ($\boldsymbol{e}_{y.1}$) are computed as:

$$\boldsymbol{e}_{y.1} = \mathbf{y} - \hat{\boldsymbol{y}}_1\ .$$

Substituting equation (21):

$$\begin{aligned}\boldsymbol{e}_{y.1} &= \mathbf{y} - \mathbf{H}_1\mathbf{y}\\ \boldsymbol{e}_{y.1} &= (\mathbf{I} - \mathbf{H}_1)\mathbf{y}\ .\end{aligned} \tag{22}$$

Similar to the results in (11), $(\mathbf{I} - \mathbf{H}_1)$ is symmetric and idempotent, thus, the Residual Sums of Squares equal:

$$\mathrm{SS}_{\mathrm{E.1}} = \sum(\boldsymbol{y} - \hat{\boldsymbol{y}}_1)^2 = \sum \boldsymbol{e}_{y.1}^2 = \mathbf{y}'(\mathbf{I} - \mathbf{H}_1)\mathbf{y}\ . \tag{23}$$

The scalar formula for the Regression Model Sum of Squares ($\mathrm{SS}_{\mathrm{M.1}}$) is Mean-Corrected:

$$\mathrm{SS}_{\mathrm{M.1}} = \sum(\hat{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}})^2\ . \tag{24}$$

Since $\hat{\boldsymbol{y}}_1 = \mathbf{H}_1\mathbf{y}$ and $\bar{\boldsymbol{y}} = \mathbf{H}_0\mathbf{y}$:

$$\begin{aligned}\boldsymbol{e}_{\hat{y}_1} &= (\hat{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}) = \mathbf{H}_1\mathbf{y} - \mathbf{H}_0\mathbf{y}\\ \boldsymbol{e}_{\hat{y}_1} &= (\hat{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}}) = (\mathbf{H}_1 - \mathbf{H}_0)\mathbf{y}\ .\end{aligned}$$

Thus, the Regression Model Sum of Squares ($\mathrm{SS}_{\mathrm{M.1}}$) in matrix form is

$$\mathrm{SS}_{\mathrm{M.1}} = \mathbf{y}'(\mathbf{H}_1 - \mathbf{H}_0)'(\mathbf{H}_1 - \mathbf{H}_0)\mathbf{y}\ . \tag{25}$$

One property of partitioned matrices is that the partition, $\mathbf{X}_m$, multiplied by the Hat matrix of a "fuller" $\mathbf{X}$ matrix ($\mathbf{H}_F$) is equal to the reduced $\mathbf{X}_m$ partition (see Myers & Milton, 1991, Lemma 4.2.2). In this case, $\mathbf{x}_0$ pre-multiplied by $\mathbf{H}_1$ results in: $\mathbf{H}_1\mathbf{x}_0 = \mathbf{x}_0$. This results in the "fuller" Hat matrix ($\mathbf{H}_1$) multiplied by the "reduced" Hat matrix ($\mathbf{H}_0$) being equal to the "reduced" Hat matrix ($\mathbf{H}_0$). Due to this property:

$$\begin{aligned}(\mathbf{H}_1 - \mathbf{H}_0)'(\mathbf{H}_1 - \mathbf{H}_0) &= \mathbf{H}_1\mathbf{H}_1 - 2\,\mathbf{H}_1\mathbf{H}_0 + \mathbf{H}_0\mathbf{H}_0\\ &= \mathbf{H}_1 - 2\mathbf{H}_0 + \mathbf{H}_0 = (\mathbf{H}_1 - \mathbf{H}_0)\ .\end{aligned} \tag{26}$$

Thus, $(\mathbf{H}_1 - \mathbf{H}_0)$ is symmetric and idempotent and the Regression Model Sum of Squares ($\mathrm{SS}_{\mathrm{M.1}}$) in equation (25) reduces to:

$$\mathrm{SS}_{\mathrm{M.1}} = \sum(\hat{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}})^2 = \sum \boldsymbol{e}_{\hat{y}_1}^2 = \mathbf{y}'(\mathbf{H}_1 - \mathbf{H}_0)\mathbf{y}\ . \tag{27}$$

The first panel of Table 2 shows the Analysis of Variance (ANOVA) Source Table for the Sums of Squares in term of the Hat Matrices used to compute each value. The second panel shows code and output from SAS® PROC REG. The regression coefficients from regressing $\mathbf{y}$ onto $\mathbf{x}_1$ is $\hat{y} = 23.05263 - 0.28947\mathbf{x}_1$.

The output also indicates that the Model and Error (Residual) Sums of Squares are $\mathrm{SS}_{\mathrm{M.1}} = \sum(\hat{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}})^2 = \mathbf{y}'(\mathbf{H}_1 - \mathbf{H}_0)\mathbf{y} = 3.18421$ and $\mathrm{SS}_{\mathrm{E.1}} = \sum(\boldsymbol{y} - \hat{\boldsymbol{y}}_1)^2 = \mathbf{y}'(\mathbf{I}\text{-}\mathbf{H}_0)\mathbf{y} = 132.81579$, respectively. Therefore, the Model $R_{Y.1}^2 = (3.18421/136) = 0.02341$ and the Multiple Correlation Coefficient is $R_{Y.1} = 0.15301$.

Table 2. Analysis of Variance (ANOVA) Source Table for the k =1 predictor model and output from SAS PROC REG.

Source		Sum of Squares Scalar	Sum of Squares Matrix	Value	df	Mean-Square	F
SS_{M1234}	Model ($\mathbf{X}_1$)	$\sum(\hat{y}_1 - \bar{y})^2$	$\mathbf{y}'(\mathbf{H}_1 - \mathbf{H}_0)\mathbf{y}$	**3.18421**	**1**	**3.18421**	**0.19180**
$SS_{E.1234}$	Residual	$\sum(y - \hat{y}_1)^2$	$\mathbf{y}'(\mathbf{I} - \mathbf{H}_1)\mathbf{y}$	**132.81579**	**8**	**16.60197**	
SS_{TOTAL}	Total	$\sum(y - \bar{y})^2$	$\mathbf{y}'(\mathbf{I} - \mathbf{H}_0)\mathbf{y}$	**136.00000**	**9**		

Note: Total Sums of Square for this and all other models is equal to the Mean Corrected Sums of 136.

$\mathbf{H}_0$ is the Hat matrix for the "intercept-only" model; $\mathbf{H}_0 = \mathbf{x}_0(\mathbf{x}_0'\mathbf{x}_0)^{-1}\mathbf{x}_0'$.

$\mathbf{H}_1$ is the Hat matrix for the k=1 predictor model with Design Matrix $\mathbf{X}_1 = \mathbf{x}_0|\mathbf{x}_1$; $\mathbf{H}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$.

```
proc reg data=R2; model y = x1 / clb stb;
output out=R2_1 predicted=yhat1 residual=ey_1;run;
```

```
Analysis of Variance
                                    Sum of            Mean
 Source                  DF        Squares          Square    F Value    Pr > F
 Model                    1        3.18421         3.18421       0.19    0.6730
 Error                    8      132.81579        16.60197
 Corrected Total          9      136.00000

              Root MSE              4.07455    R-Square     0.0234
              Dependent Mean       19.00000    Adj R-Sq    -0.0987

             Parameter    Standard                          Standardized
Variable     Estimate        Error     t Value    Pr > |t|      Estimate    95% Confidence Limits
Intercept    23.05263      9.34299        2.47      0.0389             0     1.50766      44.59760
x1           -0.28947      0.66098       -0.44      0.6730      -0.15301    -1.81370       1.23475
```

Table 3. Mean Corrected Sums of Squares and Cross-Products and Pearson Correlation Matrices for **y** and the predicted values ($\hat{\boldsymbol{y}}_1$) and residuals ($\boldsymbol{e}_{y.1}$) from the k = 1 regression model. Modified output from SAS® PROC CORR.

```
proc corr data=R2_1 cov csscp;var yhat1 ey_1 y;run;

                          Simple Statistics
Variable      N      Mean    Std Dev       Sum     Minimum      Maximum
yhat1        10        19    0.59481       190    18.13158     19.86842
ey_1         10         0    3.84152         0    -6.71053      4.42105
y            10        19    3.88730       190    12.00000     24.00000

  CSSCP Matrix
                  yhat1              ey_1                  y
  yhat1       3.1842105         0.0000000          3.1842105
  ey_1        0.0000000       132.8157895        132.8157895
  y           3.1842105       132.8157895        136.0000000

          Pearson Correlation Coefficients, N = 10
                  Prob > |r| under H0: Rho=0

                              yhat1          ey_1             y
yhat1                       1.00000       0.00000       0.15301
Predicted Value of y                       1.0000        0.6730

ey_1                        0.00000       1.00000       0.98822
Residual                     1.0000                       <.0001

y                           0.15301       0.98822       1.00000
                             0.6730        <.0001
```

PROC REG as well most other statistical software allow the user to save the predicted values ($\hat{\boldsymbol{y}}$) and residuals ($\boldsymbol{e}$). The output from SAS® PROC CORR in Table 3 shows that the Mean Corrected Sums of Squares for **y**, $\mathbf{e}_{y.1}$, and $\hat{\boldsymbol{y}}$ are $\text{CSS}_y = \mathbf{y}'(\mathbf{I}\text{-}\mathbf{H}_0)\mathbf{y} = 136$, $\text{SS}_{\text{E.1}} = \mathbf{y}'(\mathbf{I}\text{-}\mathbf{H}_1)\mathbf{y} = 132.81579$, and $\text{CSS}_{\hat{y}} = \sum(\hat{\boldsymbol{y}}_1 - \bar{\boldsymbol{y}})^2 = \mathbf{y}'(\mathbf{H}_1\text{-}\mathbf{H}_0)\mathbf{y} = 3.18421$, respectively. Note that the Mean Corrected Cross-Product between **y** and $\hat{\boldsymbol{y}}$ is $\text{CCP}_{y\hat{y}} = 3.18421$, which is equal to $\text{CSS}_{\hat{y}}$. Thus, the Pearson correlation between **y** and $\hat{\boldsymbol{y}}$ is $r_{y\hat{y}} = R_{y.1} = 3.18421/(\sqrt{(136)(3.18421)} = 0.15301$. In the case of simple regression, the Multiple R is equal to the absolute value of the Pearson Correlation, as well as being equal to the square root of the Model R^2.

Multiple Linear Regression

Regressing **y** on to $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$, the Nx5 design matrix is: $\mathbf{X}_{1234} = \mathbf{x}_0|\mathbf{x}_1|\mathbf{x}_2|\mathbf{x}_3|\mathbf{x}_4$.
The regression model is:

$$\mathbf{y} = \mathbf{X}_{1234}\mathbf{b}_{1234} + \boldsymbol{e}_{y.1234}$$
$$\boldsymbol{y} = b_0\mathbf{x}_0 + b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + b_3\mathbf{x}_3 + b_4\mathbf{x}_4 + \boldsymbol{e}_{y.1234}\,, \tag{28}$$

where $\boldsymbol{e}_{y.1234}$ are the residuals for **y** removing the effects of $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$.
The OLS solution for the regression coefficients is:

$$\mathbf{b}_{1234} = (\mathbf{X}_{1234}'\mathbf{X}_{1234})^{-1}\mathbf{X}_{1234}'\mathbf{y} = [b_0 \quad b_1 \quad b_2 \quad b_3 \quad b_4\,]'\,. \tag{29}$$

By substituting (29), the predicted values for model (28) are solved as:

$$\hat{\boldsymbol{y}}_{1234} = \mathbf{X}_{1234}(\mathbf{X}_{1234}'\mathbf{X}_{1234})^{-1}\mathbf{X}_{1234}'\mathbf{y}\,. \tag{30}$$

The Hat Matrix for $\mathbf{X}_{1234}$ is defined as:

$$\mathbf{H}_{1234} = \mathbf{X}_{1234}(\mathbf{X}_{1234}'\mathbf{X}_{1234})^{-1}\mathbf{X}_{1234}'\,. \tag{31}$$

Substituting equation (31):

$$\hat{\boldsymbol{y}}_{1234} = \mathbf{H}_{1234}\mathbf{y}\,.$$

The residuals ($\boldsymbol{e}_{y.1234}$) are computed as:

$$\boldsymbol{e}_{y.1234} = \mathbf{y} - \hat{\boldsymbol{y}}_{1234} \ .$$

Substituting equation (31):

$$\boldsymbol{e}_{y.1234} = \mathbf{y} - \mathbf{H}_{1234}\mathbf{y}$$

$$\boldsymbol{e}_{y.1234} = (\mathbf{I} - \mathbf{H}_{1234})\mathbf{y} \ . \quad (32)$$

Similar to the results in (11) and (23), ($\mathbf{I} - \mathbf{H}_{1234}$) is symmetric and idempotent with Residual Sums of Squares:

$$\mathrm{SS}_{E.1234} = \sum(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{1234}})^2 = \sum e^2_{y.1234} = \mathbf{y}'(\mathbf{I} - \mathbf{H}_{1234})\mathbf{y} \ . \quad (33)$$

The scalar formula for the Regression Model Sum of Squares ($\mathrm{SS}_{M.1234}$) is:

$$\mathrm{SS}_{M.1234} = \sum(\hat{\boldsymbol{y}}_{1234} - \overline{\boldsymbol{y}})^2 \ . \quad (34)$$

Since $\hat{\boldsymbol{y}}_{1234} = \mathbf{H}_{1234}\mathbf{y}$ and $\overline{\boldsymbol{y}} = \mathbf{H}_0\mathbf{y}$:

$$\boldsymbol{e}_{\hat{y}_{1234}} = (\hat{\boldsymbol{y}}_{1234} - \overline{\boldsymbol{y}}) = \mathbf{H}_{1234}\mathbf{y} - \mathbf{H}_0\mathbf{y}$$

$$\boldsymbol{e}_{\hat{y}_{1231}} = (\hat{\boldsymbol{y}}_{1234} - \overline{\boldsymbol{y}}) = (\mathbf{H}_{1234} - \mathbf{H}_0)\mathbf{y} \ . \quad (35)$$

Thus, the Regression Model Sum of Squares ($\mathrm{SS}_{M.1234}$) in matrix form is:

$$\mathrm{SS}_{M.1} = \mathbf{y}'(\mathbf{H}_{1234} - \mathbf{H}_0)'(\mathbf{H}_{1234} - \mathbf{H}_0)\mathbf{y} \ . \quad (36)$$

Due to idempotent properties of partitioned matrices shown in (26), the "fuller" Hat matrix ($\mathbf{H}_{1234}$) multiplied by the "reduced" Hat matrix ($\mathbf{H}_0$) is equal to the "reduced" Hat matrix ($\mathbf{H}_0$). Thus, ($\mathbf{H}_{1234} - \mathbf{H}_0$) is also symmetric and idempotent:

$$\begin{aligned}(\mathbf{H}_{1234} - \mathbf{H}_0)'(\mathbf{H}_{1234} - \mathbf{H}_0) &= \mathbf{H}_{1234}\mathbf{H}_{1234} - 2\,\mathbf{H}_{1234}\mathbf{H}_0 + \mathbf{H}_0\mathbf{H}_0 \\ &= \mathbf{H}_{1234} - 2\mathbf{H}_0 + \mathbf{H}_0 = (\mathbf{H}_{1234} - \mathbf{H}_0) \quad (37)\end{aligned}$$

Thus, the Regression Model Sum of Squares ($\mathrm{SS}_{M.1}$) in equation (36) reduces to:

$$\mathrm{SS}_{M.1234} = \sum(\hat{\boldsymbol{y}}_{1234} - \overline{\boldsymbol{y}})^2 = \sum \boldsymbol{e}^2_{\hat{y}_{1234}} = \mathbf{y}'(\mathbf{H}_{1234} - \mathbf{H}_0)\mathbf{y} \ . \quad (38)$$

The first panel of Table 4 shows the ANOVA Source Table for the Sums of Squares in term of Hat Matrices. The second panel shows code and output from SAS® PROC REG. The regression coefficients from regressing $\mathbf{y}$ onto $\mathbf{X}_{1234}$ is $\hat{y} = 8.82664 - 0.61858\mathbf{x}_1 + 0.48924\mathbf{x}_2 + 0.21445\mathbf{x}_3 + 0.46753\mathbf{x}_4$. The output also indicates that the Model and Error (Residual) Sums of Squares are $\mathrm{SS}_{M.1234} = 36.86844$ and $\mathrm{SS}_{E.1234} = 99.13156$, respectively. Therefore, the Model $R^2_{Y.1} = (36.86844/136) = 0.27109$ and the Multiple Correlation Coefficient is $R_{Y.1} = 0.52066$.

The output from SAS® PROC CORR in Table 5 shows that the Mean Corrected Sums of Squares for $\mathbf{y}$, $\mathbf{e}_{y.1}$, and $\hat{\boldsymbol{y}}$ are $\mathrm{CSS}_y = 136$, $\mathrm{SS}_{E.1} = 99.13156$, and $\mathrm{CSS}_{\hat{y}} = 36.86844$, respectively. The Mean Corrected Cross-Product between $\mathbf{y}$ and $\hat{\boldsymbol{y}}$ is equal to $\mathrm{CCP}_{y\hat{y}} = 3.18421$, which is equal to $\mathrm{CSS}_{\hat{y}}$. Thus, the Pearson correlation between $\mathbf{y}$ and $\hat{\boldsymbol{y}}$ is $r_{y\hat{y}} = R_{y.1} = 36.86844/(\sqrt{(136)(36.86844)}) = 0.52066$. In both simple and multiple linear regression, the Model Sums of Squares are equal to the cross-product of $\mathbf{y}$ and $\hat{\boldsymbol{y}}$. The previous examples provide concrete examples that demonstrate that the Mean Corrected Cross-Product between $\mathbf{y}$ and $\hat{\boldsymbol{y}}$ ($\mathrm{CCP}_{y\hat{y}}$) is equal to the Mean Corrected Sum of Squares for $\hat{\boldsymbol{y}}$ ($\mathrm{CSS}_{\hat{y}}$). Next matrix formulation will be used to demonstrate why this results in Multiple R equaling the square root of the Model R^2 computationally.

Matrix Approach to Multiple Correlation

Most regression texts point out the Model R^2 can be calculated at the ratio of $\mathrm{SS}_{MODEL}/\mathrm{SS}_{TOTAL}$:

$$R^2_{Y.1234} = \frac{\mathrm{SS}_{M.1234}}{\mathrm{SS}_{\mathrm{Total}}} = \frac{\mathbf{y}'(\mathbf{H}_{1234} - \mathbf{H}_0)\mathbf{y}}{\mathbf{y}'(\mathbf{I} - \mathbf{H}_0)\mathbf{y}} = \frac{36.8684448}{136} = 0.2711. \quad (39)$$

However, there are many matrix representations that can be used to compute the Model R^2 and Multiple R.

Again, the Multiple Correlation is the Pearson Correlation of $\mathbf{y}$ with the predicted value, $\hat{\boldsymbol{y}}$, which will be calculated as the Mean Corrected Cross-Product in ratio to the square root of Mean Corrected Sums of Squares for each variable. For this four-predictor regression model, the scalar formula for the Pearson correlation between $\mathbf{y}$ and $\hat{\boldsymbol{y}}_{1234}$ is:

$$R_{Y.1234} = r_{y\hat{y}_{1234}} = \frac{\sum(\mathbf{y} - \bar{y})(\hat{\boldsymbol{y}}_{1234} - \overline{\boldsymbol{y}}_{1234})}{\sqrt{[\sum(\boldsymbol{y} - \bar{y})^2 \sum(\hat{\boldsymbol{y}}_{1234} - \overline{\boldsymbol{y}}_{1234})^2]}}, \quad (40)$$

where $\bar{\boldsymbol{y}}_{1234}$ is the mean for the predicted values, $\hat{\boldsymbol{y}}_{1234}$. From regression theory, the expected value of the predicted values is equal to the expected value of **y**; thus, the mean of the predicted values equals the mean of **y** (e.g., Draper & Smith, 1998). Thus, formula (40) becomes

$$R_{Y.1234} = r_{y\hat{y}_{1234}} = \frac{\Sigma(\mathbf{y}-\bar{y})(\hat{\boldsymbol{y}}_{1234}-\bar{y})}{\sqrt{[\Sigma(\boldsymbol{y}-\bar{y})^2\,\Sigma(\hat{\boldsymbol{y}}_{1234}-\bar{y})^2]}}, \tag{41}$$

Thus, to Mean Correct $\hat{\boldsymbol{y}}$:

$$\boldsymbol{e}_{\hat{y}_{1234}} = \hat{\boldsymbol{y}}_{1234} - \bar{y}$$

Since $\hat{\boldsymbol{y}}_{1234} = \mathbf{H}_{1234}\mathbf{y}$ and $\bar{y} = \mathbf{H}_0\mathbf{y}$:

$$\boldsymbol{e}_{\hat{y}_{1234}} = \mathbf{H}_{1234}\mathbf{y} - \mathbf{H}_0\mathbf{y}$$

$$\boldsymbol{e}_{\hat{y}_{1234}} = (\mathbf{H}_{1234} - \mathbf{H}_0)\mathbf{y} \tag{42}$$

Thus, the Mean Corrected Sum of Squares for $\hat{\boldsymbol{y}}_{1234}$ is equal to:

$$\text{CSS}_{\hat{y}} = \mathbf{y}'(\mathbf{H}_{1234} - \mathbf{H}_0)'(\mathbf{H}_{1234} - \mathbf{H}_0)\mathbf{y}\ ,$$

which due to idempotent properties shown in (36) is equal to:

$$\text{CSS}_{\hat{y}} = \mathbf{y}'(\mathbf{H}_{1234} - \mathbf{H}_0)\mathbf{y}\ . \tag{43}$$

Thus, the Mean Corrected Sum of Squares for $\hat{\boldsymbol{y}}_{1234}$ is equal to the Regression Model Sum of Squares ($\text{SS}_{\text{M.1234}}$) in equation (38) and the Mean Corrected Cross-Product between **y** and $\hat{\boldsymbol{y}}$ ($\text{CCP}_{y\hat{y}}$). Substituting the matrix formulations for $\boldsymbol{e}_y$ (9), $\boldsymbol{e}_{\hat{y}_{1234}}$ (32), CSS_y (12) and $\text{CSS}_{\hat{y}}$ (43) (or $\text{SS}_{\text{M.1234}}$ (38)):

$$R_{Y.1234} = r_{y\hat{y}_{1234}} = \frac{\boldsymbol{e}_y{}'\boldsymbol{e}_{\hat{y}_{1234}}}{\sqrt{\text{CSS}_{\boldsymbol{y}}}\sqrt{\text{CSS}_{\hat{y}}}}$$

$$R_{Y.1234} = r_{y\hat{y}_{1234}} = \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H_0})'(\mathbf{H_{1234}}-\mathbf{H_0})\mathbf{y}}{\sqrt{\mathbf{y}'(\mathbf{I}-\mathbf{H_0})\mathbf{y}}\sqrt{\mathbf{y}'(\mathbf{H_{1234}}-\mathbf{H_0})\mathbf{y}}} \tag{44}$$

Due to idempotent properties of partitioned matrices (see eqs. 26 & 37), the entity in the middle of the numerator reduces to:

$$\begin{aligned}(\mathbf{I} - \mathbf{H}_0)'(\mathbf{H}_{1234} - \mathbf{H}_0) &= \mathbf{IH}_{1234} - \mathbf{IH}_0 - \mathbf{H}_0\mathbf{H}_{1234} + \mathbf{H}_0\mathbf{H}_0 \\ &= \mathbf{H}_{1234} - \mathbf{H}_0 - \mathbf{H}_0 + \mathbf{H}_0 = (\mathbf{H}_{1234} - \mathbf{H}_0)\end{aligned} \tag{45}$$

Therefore, the Multiple Correlation Coefficient (R) in equation (44) reduces to:

$$\begin{aligned}R_{Y.1234} = r_{y\hat{y}_{1234}} &= \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H_0})'(\mathbf{H_{1234}}-\mathbf{H_0})\mathbf{y}}{\sqrt{\mathbf{y}'(\mathbf{I}-\mathbf{H_0})\mathbf{y}}\sqrt{\mathbf{y}'(\mathbf{H_{1234}}-\mathbf{H_0})\mathbf{y}}} = \frac{\mathbf{y}'(\mathbf{H_{1234}}-\mathbf{H_0})\mathbf{y}}{\sqrt{\mathbf{y}'(\mathbf{I}-\mathbf{H_0})\mathbf{y}}\sqrt{\mathbf{y}'(\mathbf{H_{1234}}-\mathbf{H_0})\mathbf{y}}} \\ &= \frac{\mathbf{36.8684448}}{\sqrt{\mathbf{136}}\sqrt{\mathbf{36.8684448}}} = 0.2711.\end{aligned} \tag{46}$$

Note the numerator **IS** the Sum of Squares for the Regression Model ($\text{SS}_{\text{M.1234}}$), and thus, the Multiple R is always greater than or equal to zero (i.e., always positive). Further, the square root of numerator, $\mathbf{y}'(\mathbf{H}_{1234}\text{–}\mathbf{H}_0)\mathbf{y}$, appears in the denominator, thus reducing equation (46) to:

$$\begin{aligned}R_{Y.1234} = r_{y\hat{y}_{1234}} &= \frac{\mathbf{y}'(\mathbf{H_{1234}}-\mathbf{H_0})\mathbf{y}}{\sqrt{\mathbf{y}'(\mathbf{I}-\mathbf{H_0})\mathbf{y}}\sqrt{\mathbf{y}'(\mathbf{H_{1234}}-\mathbf{H_0})\mathbf{y}}} = \frac{\sqrt{\mathbf{y}'(\mathbf{H_{1234}}-\mathbf{H_0})\mathbf{y}}}{\sqrt{\mathbf{y}'(\mathbf{I}-\mathbf{H_0})\mathbf{y}}} \\ &= \sqrt{\frac{\mathbf{y}'(\mathbf{H_{1234}}-\mathbf{H_0})\mathbf{y}}{\mathbf{y}'(\mathbf{I}-\mathbf{H_0})\mathbf{y}}} = \sqrt{\frac{\mathbf{36.8684448}}{\mathbf{136}}} = \sqrt{0.2711} = \sqrt{R^2_{Y.1234}} = 0.52066\end{aligned} \tag{47}$$

Thus, by substituting the matrices necessary to compute Mean Corrected Sums of Squares and Cross-Products into the Pearson correlation formula (40) and utilizing the idempotent properties of partitioned matrices, it can be shown that computationally the Multiple R **IS** the square root of the Model R^2.

References

Draper, N. R., & Smith, H. (1998). Applied Regression Analysis (3rd ed.). New: York: John Wiley & Sons.
Myers, R. H., & Milton, J. S. (1991). A First Course in the Theory of Linear Models. Boston; PWS-Kent.

Send correspondence to: T. Mark Beasley
University of Alabama at Birmingham
Email: mbeasley@uab.edu

Table 4. Analysis of Variance (ANOVA) Source Table for the k = 4 predictor model and output from SAS PROC REG.

Source		Sum of Squares Scalar	Sum of Squares Matrix	Value	df	Mean-Square	F
SS_{M1234}	Model ($\mathbf{X}_{1234}$)	$\sum(\hat{y}_{1234} - \bar{y})^2$	$\mathbf{y}'(\mathbf{H}_{1234} - \mathbf{H}_0)\mathbf{y}$	**36.86844**	**4**	**3.18421**	**0.46489**
$SS_{E.1234}$	Residual	$\sum(y - \hat{y}_{1234})^2$	$\mathbf{y}'(\mathbf{I} - \mathbf{H}_{1234})\mathbf{y}$	**99.13156**	**5**	**16.60197**	
SS_{TOTAL}	Total	$\sum(y - \bar{y})^2$	$\mathbf{y}'(\mathbf{I} - \mathbf{H}_0)\mathbf{y}$	**136.00000**	**9**		

Note: Total Sums of Square for this and all other models is equal to the Mean Corrected Sums of 136.

$\mathbf{H}_0$ is the Hat matrix for the "intercept-only" model; $\mathbf{H_0} = \mathbf{x}_0(\mathbf{x}_0'\mathbf{x}_0)^{-1}\mathbf{x}_0'$.

$\mathbf{H}_{1234}$ is the Hat matrix for the k=4 predictor model with Design Matrix $\mathbf{X}_{1234} = \mathbf{x}_0|\mathbf{x}_1|\mathbf{x}_2|\mathbf{x}_3|\mathbf{x}_4$; $\mathbf{H}_1 = \mathbf{X}_{1234}(\mathbf{X}_{1234}'\mathbf{X}_{1234})^{-1}\mathbf{X}_{1234}'$.

```
proc reg data=R2; model y = x1 x2 x3 x4/ clb stb scorr2;
output out=R2_1234 predicted=yhat1234 residual=ey_1234;run;
```

```
Analysis of Variance
                                          Sum of            Mean
         Source                   DF     Squares          Square    F Value    Pr > F
         Model                     4    36.86844         9.21711       0.46    0.7610
         Error                     5    99.13156        19.82631
         Corrected Total           9   136.00000

                      Root MSE              4.45267    R-Square     0.2711
                      Dependent Mean       19.00000    Adj R-Sq    -0.3120
                      Coeff Var            23.43513
```

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Squared Semi-partial Corr Type II	95% Confidence Limits	
Intercept	8.82664	20.70312	0.43	0.6876	0	.	-44.39243	62.04571
x1	-0.61858	0.98458	-0.63	0.5574	-0.32698	0.05754	-3.14952	1.91237
x2	0.48924	0.99389	0.49	0.6434	0.30252	0.03532	-2.06562	3.04411
x3	0.21445	1.01558	0.21	0.8411	0.10072	0.00650	-2.39618	2.82509
x4	0.46753	0.92616	0.50	0.6352	0.28348	0.03715	-1.91325	2.84831

Table 5. Mean Corrected Sums of Squares and Cross-Products and Pearson Correlation Matrices for **y**, predicted values ($\hat{\mathbf{y}}_{1234}$), and residuals ($\boldsymbol{e}_{y.1234}$) from the k=4 regression model.
Output from SAS® PROC CORR.

```
proc corr data=R2_1234  csscp;var yhat1234 ey_1234 y;run;
                         Simple Statistics
Variable       N    Mean  Std Dev      Sum     Minimum     Maximum
yhat1234      10     19   2.02398      190    14.95845    21.78520
ey_1234       10      0   3.31883        0    -6.57416     4.11969
y             10     19   3.88730      190    12.00000    24.00000

  CSSCP Matrix
                  yhat1234            ey_1234                 y
    yhat1234    36.8684448          0.0000000        36.8684448
    ey_1234      0.0000000         99.1315552        99.1315552
    y           36.8684448         99.1315552       136.0000000

               Pearson Correlation Coefficients, N = 10
                        Prob > |r| under H0: Rho=0

                          yhat1234       ey_1234              y
yhat1234                   1.00000       0.00000        0.52066
Predicted Value of y                      1.0000         0.1228

ey_1234                    0.00000       1.00000        0.85376
Residual                    1.0000                       0.0017

y                          0.52066       0.85376        1.00000
                            0.1228        0.0017
```