open.michigan

Author(s): Kerby Shedden, Ph.D., 2010

Decomposing Variance

Kerby Shedden

Department of Statistics, University of Michigan

October 20, 2014

Law of total variation

For any regression model involving a response Y and a covariate vector X, we have

$$\mathrm{var}(Y) = \mathrm{var}_X E(Y|X) + E_X \mathrm{var}(Y|X).$$

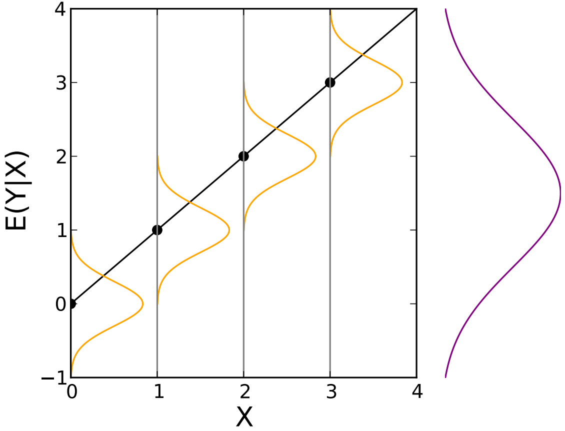
Note that this only makes sense if we treat X as being random.

We often wish to distinguish these two situations:

- The population is homoscedastic: $\mathrm{var}(Y|X)$ does not depend on X, so we can simply write $\mathrm{var}(Y|X) = \sigma^2$, and we get $\mathrm{var}(Y) = \mathrm{var}_X E(Y|X) + \sigma^2$.
- The population is heteroscedastic: $\mathrm{var}(Y|X)$ is a function $\sigma^2(X)$ with expected value $\sigma^2 = E_X \sigma^2(X)$, and again we get $\mathrm{var}(Y) = \mathrm{var}_X E(Y|X) + \sigma^2$.

If we write $Y = f(X) + \epsilon$ with $E(\epsilon|X) = 0$, then $E(Y|X) = f(X)$, and $\mathrm{var}_X E(Y|X)$ summarizes the variation of $f(X)$ over the marginal distribution of X.

Law of total variation



Orange curves: conditional distributions of Y given X
Purple curve: marginal distribution of Y
Black dots: conditional means of Y given X

Pearson correlation

The population Pearson correlation coefficient of two jointly distributed scalar-valued random variables X and Y is

$$\rho_{XY} \equiv \frac{\mathrm{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Given data $Y = (Y_1, \ldots, Y_n)'$ and $X = (X_1, \ldots, X_n)'$, the Pearson correlation coefficient is estimated by

$$\hat{\rho}_{XY} = \frac{\widehat{\mathrm{cov}}(X, Y)}{\hat{\sigma}_X \hat{\sigma}_Y} = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \cdot \sum_i (Y_i - \bar{Y})^2}} = \frac{(X - \bar{X})'(Y - \bar{Y})}{\|X - \bar{X}\| \cdot \|Y - \bar{Y}\|}.$$

When we write $Y - \bar{Y}$ here, this means $Y - \bar{Y} \cdot \mathbf{1}$, where $\mathbf{1}$ is a vector of 1's, and $\bar{Y}$ is a scalar.

Pearson correlation

By the Cauchy-Schwartz inequality,

$$\begin{aligned} -1 &\leq \rho_{XY} \leq 1 \\ -1 &\leq \hat{\rho}_{XY} \leq 1. \end{aligned}$$

The sample correlation coefficient is slightly biased, but the bias is so small that it is usually ignored.

Pearson correlation and simple linear regression slopes

For the simple linear regression model

$$Y = \alpha + \beta X + \epsilon,$$

if we view X as a random variable that is uncorrelated with ϵ, then

$$\mathrm{cov}(X, Y) = \beta \sigma_X^2$$

and the correlation is

$$\rho_{XY} \equiv \mathrm{cor}(X, Y) = \frac{\beta}{\sqrt{\beta^2 + \sigma^2/\sigma_X^2}}.$$

The sample correlation coefficient is related to the least squares slope estimate:

$$\hat{\beta} = \frac{\widehat{\mathrm{cov}}(X, Y)}{\hat{\sigma}_X^2} = \hat{\rho}_{XY} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}.$$

Orthogonality between fitted values and residuals

Recall that the fitted values are

$$\hat{Y} = X\hat{\beta} = PY$$

and the residuals are

$$R = Y - \hat{Y} = (I - P)Y.$$

Since $P(I - P) = 0$ it follows that $\hat{Y}'R = 0$.

since $\bar{R} = 0$, it is equivalent to state that the sample correlation between R and $\hat{Y}$ is zero, i.e.

$$\widehat{\mathrm{cor}}(R, \hat{Y}) = 0.$$

Coefficient of determination

A descriptive summary of the explanatory power of X for Y is given by the coefficient of determination, also known as the proportion of explained variance, or multiple R^2. This is the quantity

$$R^2 \equiv 1 - \frac{\|Y - \hat{Y}\|^2}{\|Y - \bar{Y}\|^2} = \frac{\|\hat{Y} - \bar{Y}\|^2}{\|Y - \bar{Y}\|^2} = \frac{\widehat{\mathrm{var}}(\hat{Y})}{\widehat{\mathrm{var}}(Y)}.$$

The equivalence between the two expressions follows from the identity

$$\begin{aligned} \|Y - \bar{Y}\|^2 &= \|Y - \hat{Y} + \hat{Y} - \bar{Y}\|^2 \\ &= \|Y - \hat{Y}\|^2 + \|\hat{Y} - \bar{Y}\|^2 + 2(Y - \hat{Y})'(\hat{Y} - \bar{Y}) \\ &= \|Y - \hat{Y}\|^2 + \|\hat{Y} - \bar{Y}\|^2, \end{aligned}$$

It should be clear that $R^2 = 0$ iff $\hat{Y} = \bar{Y}$ and $R^2 = 1$ iff $\hat{Y} = Y$.

Coefficient of determination

The coefficient of determination is equal to

$$\widehat{\mathrm{cor}}(\hat{Y}, Y)^2.$$

To see this, note that

$$\begin{aligned}
\widehat{\mathrm{cor}}(\hat{Y}, Y) &= \frac{(\hat{Y} - \bar{Y})'(Y - \bar{Y})}{\|\hat{Y} - \bar{Y}\| \cdot \|Y - \bar{Y}\|} \\
&= \frac{(\hat{Y} - \bar{Y})'(Y - \hat{Y} + \hat{Y} - \bar{Y})}{\|\hat{Y} - \bar{Y}\| \cdot \|Y - \bar{Y}\|} \\
&= \frac{(\hat{Y} - \bar{Y})'(Y - \hat{Y}) + (\hat{Y} - \bar{Y})'(\hat{Y} - \bar{Y})}{\|\hat{Y} - \bar{Y}\| \cdot \|Y - \bar{Y}\|} \\
&= \frac{\|\hat{Y} - \bar{Y}\|}{\|Y - \bar{Y}\|}.
\end{aligned}$$

Coefficient of determination in simple linear regression

In general,

$$R^2 = \widehat{\mathrm{cor}}(Y, \hat{Y})^2 = \frac{\widehat{\mathrm{cov}}(Y, \hat{Y})^2}{\widehat{\mathrm{var}}(Y) \cdot \widehat{\mathrm{var}}(\hat{Y})}.$$

In the case of simple linear regression,

$$\begin{aligned} \widehat{\mathrm{cov}}(Y, \hat{Y}) &= \widehat{\mathrm{cov}}(Y, \hat{\alpha} + \hat{\beta} X) \\ &= \hat{\beta}\, \widehat{\mathrm{cov}}(Y, X), \end{aligned}$$

and

$$\begin{aligned} \widehat{\mathrm{var}}(\hat{Y}) &= \widehat{\mathrm{var}}(\hat{\alpha} + \hat{\beta} X) \\ &= \hat{\beta}^2 \widehat{\mathrm{var}}(X) \end{aligned}$$

Thus for simple linear regression, $R^2 = \widehat{\mathrm{cor}}(Y, X)^2 = \widehat{\mathrm{cor}}(Y, \hat{Y})^2$.

Relationship to the F statistic

The F-statistic for the null hypothesis

$$\beta_1 = \ldots = \beta_p = 0$$

is

$$\frac{\|\hat{Y} - \bar{Y}\|^2}{\|Y - \hat{Y}\|^2} \cdot \frac{n-p-1}{p} = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p},$$

which is an increasing function of R^2.

Adjusted R^2

The sample R^2 is an estimate of the population R^2:

$$1 - \frac{\mathrm{var}(Y|X)}{\mathrm{var}(Y)}.$$

Since it is a ratio, the plug-in estimate R^2 is biased, although the bias is not large unless the sample size is small or the number of covariates is large. The adjusted R^2 is an approximately unbiased estimate of the population R^2:

$$1 - (1 - R^2)\frac{n-1}{n-p-1}.$$

The adjusted R^2 is always less than the unadjusted R^2. The adjusted R^2 is always less than or equal to one, but can be negative.

The unique variation in one covariate

How much "information" about Y is present in a covariate X_k? This question is not straightforward when the covariates are non-orthogonal, since several covariates may contain overlapping information about Y.

Let $X_k^\perp$ be the residual of X_k after regressing it against all other covariates (including the intercept). If P_{-k} is the projection onto $\mathrm{span}(\{X_j, j \neq k\})$, then

$$X_k^\perp = (I - P_{-k})X_k.$$

We could use $\widehat{\mathrm{var}}(X_k^\perp)/\widehat{\mathrm{var}}(X_k)$ to assess how much of the variation in X_k is "unique" in that it is not also captured by other predictors.

But this measure doesn't involve Y, so it can't tell us whether the unique variation in X_k is useful in the regression analysis.

The unique regression information in one covariate

To learn how X_k contributes "uniquely" to the regression, we can consider how introducing X_k to a working regression model affects the R^2.

Let $\hat{Y}_{-k} = P_{-k}Y$ be the fitted values in the model omitting covariate k.

Let R^2 denote the multiple R^2 for the full model, and let R^2_{-k} be the multiple R^2 for the regression omitting covariate X_k. The value of

$$R^2 - R^2_{-k}$$

is a way to quantify how much unique information about Y in X_k is not captured by the other covariates. This is called the semi-partial R^2.

Identity involving norms of fitted values and residuals

Before we continue, we will need a simple identity that is often useful.

In general, if A and B are orthogonal, then $\|A+B\|^2 = \|A\|^2 + \|B\|^2$.

If A and $B - A$ are orthogonal, then

$$\|B\|^2 = \|B - A + A\|^2 = \|B - A\|^2 + \|A\|^2.$$

Thus we have $\|B\|^2 - \|A\|^2 = \|B - A\|^2$.

Applying this fact to regression, we know that the fitted values and residuals are orthogonal. Thus for the regression omitting variable k, $\hat{Y}_{-k}$ and $Y - \hat{Y}_{-k}$ are orthogonal, so

so $\|Y - \hat{Y}_{-k}\|^2 = \|Y\|^2 - \|\hat{Y}_{-k}\|^2$.

By the same argument, $\|Y - \hat{Y}\|^2 = \|Y\|^2 - \|\hat{Y}\|^2$.

Improvement in R^2 due to one covariate

Now we can obtain a simple, direct expression for the semi-partial R^2.

Since $X_k^\perp$ is orthogonal to the other covariates,

$$\hat{Y} = \hat{Y}_{-k} + \frac{\langle Y, X_k^\perp \rangle}{\langle X_k^\perp, X_k^\perp \rangle} X_k^\perp,$$

and

$$\|\hat{Y}\|^2 = \|\hat{Y}_{-k}\|^2 + \langle Y, X_k^\perp \rangle^2 / \|X_k^\perp\|^2.$$

Improvement in R^2 due to one covariate

Thus we have

$$
\begin{aligned}
R^2 &= 1 - \frac{\|Y - \hat{Y}\|^2}{\|Y - \bar{Y}\|^2} \\
&= 1 - \frac{\|Y\|^2 - \|\hat{Y}\|^2}{\|Y - \bar{Y}\|^2} \\
&= 1 - \frac{\|Y\|^2 - \|\hat{Y}_{-k}\|^2 - \langle Y, X_k^\perp \rangle^2 / \|X_k^\perp\|^2}{\|Y - \bar{Y}\|^2} \\
&= 1 - \frac{\|Y - \hat{Y}_{-k}\|^2}{\|Y - \bar{Y}\|^2} + \frac{\langle Y, X_k^\perp \rangle^2 / \|X_k^\perp\|^2}{\|Y - \bar{Y}\|^2} \\
&= R^2_{-k} + \frac{\langle Y, X_k^\perp \rangle^2 / \|X_k^\perp\|^2}{\|Y - \bar{Y}\|^2}.
\end{aligned}
$$

Semi-partial R^2

Thus the semi-partial R^2 is

$$R^2 - R^2_{-k} = \frac{\langle Y, X_k^\perp \rangle^2 / \|X_k^\perp\|^2}{\|Y - \bar{Y}\|^2} = \frac{\langle Y, X_k^\perp / \|X_k^\perp\| \rangle^2}{\|Y - \bar{Y}\|^2}$$

where $\hat{Y}_k$ is the fitted value for regressing Y on $X_k^\perp$.

Since $X_k^\perp / \|X_k^\perp\|$ is centered and has length 1, it follows that

$$R^2 - R^2_{-k} = \widehat{\mathrm{cor}}(Y, X_k^\perp)^2 = \widehat{\mathrm{cor}}(Y, \hat{Y}_k)^2.$$

Thus the semi-partial R^2 for covariate k has two equivalent interpretations:

- It is the improvement in R^2 resulting from including covariate k in a working regression model that already contains the other covariates.
- It is the R^2 for a simple linear regression of Y on $X_k^\perp = (I - P_{-k})X_k$.

Partial R^2

The partial R^2 is

$$\frac{R^2 - R^2_{-k}}{1 - R^2_{-k}} = \frac{\langle Y, X_k^\perp \rangle^2 / \|X_k^\perp\|^2}{\|Y - \hat{Y}_{-k}\|^2}.$$

The partial R^2 for covariate k is the fraction of the maximum possible improvement in R^2 that is contributed by covariate k.

Let $\hat{Y}_{-k}$ be the fitted values for regressing Y on all covariates except X_k.

Since $\hat{Y}'_{-k} X_k^\perp = 0$,

$$\frac{\langle Y, X_k^\perp \rangle^2}{\|Y - \hat{Y}_{-k}\|^2 \cdot \|X_k^\perp\|^2} = \frac{\langle Y - \hat{Y}_{-k}, X_k^\perp \rangle^2}{\|Y - \hat{Y}_{-k}\|^2 \cdot \|X_k^\perp\|^2}$$

The expression on the left is the usual R^2 that would be obtained when regressing $Y - \hat{Y}_{-k}$ on $X_k^\perp$. Thus the partial R^2 is the same as the usual R^2 for $(I - P_{-k})Y$ regressed on $(I - P_{-k})X_k$.

Decomposition of projection matrices

Suppose $P \in \mathcal{R}^{n \times n}$ is a rank-d projection matrix, and U is a $n \times d$ orthogonal matrix whose columns span $\mathrm{col}(P)$. If we partition U by columns

$$U = \begin{pmatrix} | & | & \cdots & | \\ U_1 & U_2 & \cdots & U_d \\ | & | & \cdots & | \end{pmatrix},$$

then $P = UU'$, so we can write

$$P = \sum_{j=1}^{d} U_j U_j'.$$

Note that this representation is not unique, since there are different orthogonal bases for $\mathrm{col}(P)$.

Each summand $U_j U_j' \in \mathcal{R}^{n \times n}$ is a rank-1 projection matrix onto $\langle U_j \rangle$.

Decomposition of R^2

Question: In a multiple regression model, how much of the variance in Y is explained by a particular covariate?

Orthogonal case: If the design matrix X is orthogonal ($X'X = I$), the projection P onto $\mathrm{col}(X)$ can be decomposed as

$$P = \sum_{j=0}^{p} P_j = \frac{11'}{n} + \sum_{j=1}^{p} X_j X_j',$$

where X_j is the j^{th} column of the design matrix (assuming here that the first column of X is an intercept).

Decomposition of R^2 (orthogonal case)

The $n \times n$ rank-1 matrix

$$P_j = X_j X_j'$$

is the projection onto $\mathrm{span}(X_j)$ (and P_0 is the projection onto the span of the vector of 1's). Furthermore, by orthogonality, $P_j P_k = 0$ unless $j = k$. Since

$$\hat{Y} - \bar{Y} = \sum_{j=1}^{p} P_j Y,$$

by orthogonality

$$\|\hat{Y} - \bar{Y}\|^2 = \sum_{j=1}^{p} \|P_j Y\|^2.$$

Here we are using the fact that if $U_1, \ldots, U_m$ are orthogonal, then

$$\|U_1 + \cdots + U_m\|^2 = \|U_1\|^2 + \cdots + \|U_m\|^2.$$

Decomposition of R^2 (orthogonal case)

The R^2 for simple linear regression of Y on X_j is

$$R_j^2 \equiv \|\hat{Y} - \bar{Y}\|^2 / \|Y - \bar{Y}\|^2 = \|P_j Y\|^2 / \|Y - \bar{Y}\|^2,$$

so we see that for orthogonal design matrices,

$$R^2 = \sum_{j=1}^{p} R_j^2.$$

That is, the overall coefficient of determination is the sum of univariate coefficients of determination for all the explanatory variables.

Decomposition of R^2

Non-orthogonal case: If X is not orthogonal, the overall R^2 will not be the sum of single covariate R^2's.

If we let R_j^2 be as above (the R^2 values for regressing Y on each X_j), then there are two different situations: $\sum_j R_j^2 > R^2$, and $\sum_j R_j^2 < R^2$.

Decomposition of R^2

Case 1: $\sum R_j^2 > R^2$

It's not surprising that $\sum_j R_j^2$ can be bigger than R^2. For example, suppose that

$$Y = X_1 + \epsilon$$

is the data generating model, and X_2 is highly correlated with X_1 (but is not part of the data generating model).

For the regression of Y on both X_1 and X_2, the multiple R^2 will be $1 - \sigma^2/\mathrm{var}(Y)$ (since $E(Y|X_1, X_2) = E(Y|X_1) = X_1$).

The R^2 values for Y regressed on either X_1 or X_2 separately will also be approximately $1 - \sigma^2/\mathrm{var}(Y)$.

Thus $R_1^2 + R_2^2 \approx 2R^2$.

Decomposition of R^2

Case 2: $\sum_j R_j^2 < R^2$

This is more surprising, and is sometimes called enhancement.

As an example, suppose the data generating model is

$$Y = Z + \epsilon,$$

but we don't observe Z (for simplicity assume $EZ = 0$). Instead, we observe a value X_1 that satisfies

$$X_1 = Z + X_2,$$

where X_2 has mean 0 and is independent of Z and ϵ.

Since X_2 is independent of Z and ϵ, it is also independent of Y, thus $R_2^2 \approx 0$ for large n.

Decomposition of R^2 (enhancement example)

The multiple R^2 of Y on X_1 and X_2 is approximately $\sigma_Z^2/(\sigma_Z^2 + \sigma^2)$ for large n, since the fitted values will converge to $\hat{Y} = X_1 - X_2 = Z$.

To calculate R_1^2, first note that for the regression of Y on X_1,

$$\hat{\beta} = \frac{\widehat{\mathrm{cov}}(Y, X_1)}{\widehat{\mathrm{var}}(X_1)} \to \frac{\sigma_Z^2}{\sigma_Z^2 + \sigma_{X_2}^2}$$

and

$$\hat{\alpha} \to 0.$$

Decomposition of R^2 (enhancement example)

Therefore for large n,

$$
\begin{aligned}
n^{-1}\|Y-\hat{Y}\|^2 &\approx n^{-1}\|Z+\epsilon-\sigma_Z^2 X_1/(\sigma_Z^2+\sigma_{X_2}^2)\|^2 \\
&= n^{-1}\|\sigma_{X_2}^2 Z/(\sigma_Z^2+\sigma_{X_2}^2)+\epsilon-\sigma_Z^2 X_2/(\sigma_Z^2+\sigma_{X_2}^2)\|^2 \\
&= \sigma_{X_2}^4\sigma_Z^2/(\sigma_Z^2+\sigma_{X_2}^2)^2+\sigma^2+\sigma_Z^4\sigma_{X_2}^2/(\sigma_Z^2+\sigma_{X_2}^2)^2 \\
&= \sigma_{X_2}^2\sigma_Z^2/(\sigma_Z^2+\sigma_{X_2}^2)+\sigma^2.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
R_1^2 &= 1-\frac{n^{-1}\|Y-\hat{Y}\|^2}{n^{-1}\|Y-\bar{Y}\|^2} \\
&\approx 1-\frac{\sigma_{X_2}^2\sigma_Z^2/(\sigma_Z^2+\sigma_{X_2}^2)+\sigma^2}{\sigma_Z^2+\sigma^2} \\
&= \frac{\sigma_Z^2}{(\sigma_Z^2+\sigma^2)(1+\sigma_{X_2}^2/\sigma_Z^2)}
\end{aligned}
$$

.

Decomposition of R^2 (enhancement example)

Thus

$$R_1^2/R^2 \approx 1/(1 + \sigma_{X_2}^2/\sigma_Z^2),$$

which is strictly less than one if $\sigma_{X_2}^2 > 0$.

Since $R_2^2 = 0$, it follows that $R^2 > R_1^2 + R_2^2$.

The reason for this is that while X_2 contains no directly useful information about Y (hence $R_2^2 = 0$), it can remove the "measurement error" in X_1, making X_1 a better predictor of Z.

Decomposition of R^2 (enhancement example)

We can also calculate the limiting partial R^2 for adding X_2 to a model that already contains X_1:

$$\frac{\sigma^2_{X_2}}{\sigma^2_{X_2} + \sigma^2(1 + \sigma^2_{X_2}/\sigma^2_Z)}.$$

Partial R^2 example 2

Suppose the design matrix satisfies

$$X'X/n = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & r \\ 0 & r & 1 \end{pmatrix}$$

and the data generating model is

$$Y = X_1 + X_2 + \epsilon$$

with $\mathrm{var}\,\epsilon = \sigma^2$.

Partial R^2 example 2

We will calculate the partial R^2 for X_1, using the fact that the partial R^2 is the regular R^2 for regressing

$$(I - P_{-1})Y$$

on

$$(I - P_{-1})X_1$$

where P_{-1} is the projection onto $\mathrm{span}\left(\{1, X_2\}\right)$.

Since this is a simple linear regression, the partial R^2 can be expressed

$$\widehat{\mathrm{cor}}((I - P_{-1})Y, (I - P_{-1})X_1)^2.$$

Partial R^2 example 2

The numerator of the partial R^2 is the square of

$$
\begin{aligned}
\widehat{\mathrm{cov}}((I-P_{-1})Y,(I-P_{-1})X_1) &= Y'(I-P_{-1})X_1/n \\
&= (X_1+X_2+\epsilon)'(X_1-rX_2)/n \\
&\to 1-r^2.
\end{aligned}
$$

The denominator contains two factors. The first is

$$
\begin{aligned}
\|(I-P_{-1})X_1\|^2/n &= X_1'(I-P_{-1})X_1/n \\
&= X_1'(X_1-rX_2)/n \\
&\to 1-r^2.
\end{aligned}
$$

Partial R^2 example 2

The other factor in the denominator is $Y'(I - P_{-1})Y/n$:

$$
\begin{aligned}
Y'(I - P_{-1})Y/n &= (X_1 + X_2)'(I - P_{-1})(X_1 + X_2)/n + \epsilon'(I - P_{-1})\epsilon/n + \\
&\quad 2\epsilon'(I - P_{-1})(X_1 + X_2)/n \\
&\approx (X_1 + X_2)'(X_1 - rX_2)/n + \sigma^2 \\
&\to 1 - r^2 + \sigma^2.
\end{aligned}
$$

Thus we get that the partial R^2 is approximately equal to

$$
\frac{1 - r^2}{1 - r^2 + \sigma^2}.
$$

If $r = 1$ then the result is zero (X_1 has no unique explanatory power), and if $r = 0$, the result is $1/(1 + \sigma^2)$, indicating that after controlling for X_2, around $1/(1 + \sigma^2)$ fraction of the remaining variance is explained by X_1 (the rest is due to ϵ).

Summary

Each of the three R^2 values can be expressed either in terms of variance ratios, or as a squared correlation coefficient:

	Multiple R^2	Semi-partial R^2	Partial R^2
VR	$\|\hat{Y} - \bar{Y}\|^2 / \|Y - \bar{Y}\|^2$	$R^2 - R^2_{-k}$	$(R^2 - R^2_{-k})/(1 - R^2_{-k})$
Correlation	$\widehat{\mathrm{cor}}(\hat{Y}, Y)^2$	$\widehat{\mathrm{cor}}(Y, X_k^{\perp})^2$	$\widehat{\mathrm{cor}}((I - P_{-k})Y, X_k^{\perp})^2$