October 13, 2019

Algebra of Least Squares

Geometry of least squares

Recall that out data is like a table $[\boldsymbol{Y}\ \boldsymbol{X}]$ where $\boldsymbol{Y}$ collects n observations on the dependent variable and $\boldsymbol{X}$ collects n observations on the k-dimensional independent variable:

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix} = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,k} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n,1} & X_{n,k} & \cdots & X_{n,k} \end{bmatrix}_{n\times k} \quad \text{and } \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} .$$

We can think of $\boldsymbol{Y}$ and the columns of $\boldsymbol{X}$ as members of the n-dimensional Euclidean space $\mathbb{R}^n$. One can define a subspace of $\mathbb{R}^n$ called the *column space* of a $n \times k$ matrix $\boldsymbol{X}$, that is a collection of all vectors in $\mathbb{R}^n$ that can be written as linear combinations of the columns of $\boldsymbol{X}$:

$$\mathcal{S}(\boldsymbol{X}) = \left\{ \boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{z} = \boldsymbol{X}\boldsymbol{b},\, \boldsymbol{b} = (b_1, b_2, \ldots, b_k)' \in \mathbb{R}^k \right\}.$$

For two vectors $\boldsymbol{a}, \boldsymbol{b}$ in $\mathbb{R}^n$, the distance between $\boldsymbol{a}$ and $\boldsymbol{b}$ is given by the Euclidean norm[1] of their difference $\|\boldsymbol{a} - \boldsymbol{b}\| = \sqrt{(\boldsymbol{a} - \boldsymbol{b})' (\boldsymbol{a} - \boldsymbol{b})}$. Thus, the least squares problem, minimization of the sum-of-squared errors $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})' (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})$, is to find, out of all elements of $\mathcal{S}(\boldsymbol{X})$, the one closest to $\boldsymbol{Y}$:

$$\min_{\boldsymbol{y} \in \mathcal{S}(\boldsymbol{X})} \|\boldsymbol{Y} - \boldsymbol{y}\|^2 .$$

The closest point is found by "dropping a perpendicular". That is, a solution to the least squares problem, $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ must be chosen so that the residual vector $\widehat{\boldsymbol{e}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ is orthogonal (perpendicular) to each column of $\boldsymbol{X}$:

$$\widehat{\boldsymbol{e}}' \boldsymbol{X} = \boldsymbol{0}.$$

As a result, $\widehat{\boldsymbol{e}}$ is orthogonal to every element of $\mathcal{S}(\boldsymbol{X})$. Indeed, if $\boldsymbol{z} \in \mathcal{S}(\boldsymbol{X})$, then there exists $\boldsymbol{b} \in \mathbb{R}^k$ such that $\boldsymbol{z} = \boldsymbol{X}\boldsymbol{b}$, and

$$\begin{aligned} \widehat{\boldsymbol{e}}' \boldsymbol{z} &= \widehat{\boldsymbol{e}}' \boldsymbol{X}\boldsymbol{b} \\ &= 0. \end{aligned}$$

The collection of the elements of $\mathbb{R}^n$ orthogonal to $\mathcal{S}(\boldsymbol{X})$ is called the *orthogonal complement* of $\mathcal{S}(\boldsymbol{X})$:

$$\mathcal{S}^{\perp}(\boldsymbol{X}) = \left\{ \boldsymbol{z} \in \mathbb{R}^n : \boldsymbol{z}' \boldsymbol{X} = \boldsymbol{0} \right\}.$$

Every element of $\mathcal{S}^{\perp}(\boldsymbol{X})$ is orthogonal to every element in $\mathcal{S}(\boldsymbol{X})$.

[1] For a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$, its Euclidean norm is defined as $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}'\boldsymbol{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}$.

The solution to the least squares problem is given by

$$\begin{aligned}\widehat{\boldsymbol{Y}} &= \boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} \\ &= \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y},\end{aligned}$$

where

$$\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$$

is called the *orthogonal projection matrix.* For any vector $\boldsymbol{z} \in \mathbb{R}^n$,

$$\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{z} \in \mathcal{S}(\boldsymbol{X}).$$

Furthermore, the residual vector will be in $\mathcal{S}^{\perp}(\boldsymbol{X})$:

$$\boldsymbol{z} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{z} \in \mathcal{S}^{\perp}(\boldsymbol{X}). \tag{1}$$

To show (1), first note, that, since the columns of $\boldsymbol{X}$ are in $\mathcal{S}(\boldsymbol{X})$,

$$\begin{aligned}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X} &= \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X} \\ &= \boldsymbol{X},\end{aligned}$$

and, since $\boldsymbol{P}_{\boldsymbol{X}}$ is a symmetric matrix,

$$\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}'.$$

Now,

$$\begin{aligned}\boldsymbol{X}'\left(\boldsymbol{z} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{z}\right) &= \boldsymbol{X}'\boldsymbol{z} - \boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{z} \\ &= \boldsymbol{X}'\boldsymbol{z} - \boldsymbol{X}'\boldsymbol{z} \\ &= \boldsymbol{0}.\end{aligned}$$

Thus, by the definition, the residuals $\boldsymbol{z} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{z}$ belongs to $\mathcal{S}^{\perp}(\boldsymbol{X})$. The residuals can be written as

$$\begin{aligned}\widehat{\boldsymbol{e}} &= \boldsymbol{Y} - \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} \\ &= \left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}}\right)\boldsymbol{Y} \\ &= \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{Y},\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{M}_{\boldsymbol{X}} &= \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}} \\ &= \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}',\end{aligned}$$

is a projection matrix onto $\mathcal{S}^{\perp}(\boldsymbol{X})$.

The projection matrices $\boldsymbol{P}_{\boldsymbol{X}}$ and $\boldsymbol{M}_{\boldsymbol{X}}$ have the following properties:

- $\boldsymbol{P_X} + \boldsymbol{M_X} = \boldsymbol{I}_n$. This implies, that for any $\boldsymbol{z} \in \mathbb{R}^n$,

$$\boldsymbol{z} = \boldsymbol{P_X z} + \boldsymbol{M_X z}.$$

- Symmetric:

$$\begin{aligned} \boldsymbol{P}'_{\boldsymbol{X}} &= \boldsymbol{P_X}, \\ \boldsymbol{M}'_{\boldsymbol{X}} &= \boldsymbol{M_X}. \end{aligned}$$

- Idempotent: $\boldsymbol{P_X P_X} = \boldsymbol{P_X}$, and $\boldsymbol{M_X M_X} = \boldsymbol{M_X}$.

$$\begin{aligned} \boldsymbol{P_X P_X} &= \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\ &= \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\ &= \boldsymbol{P_X} \\ \boldsymbol{M_X M_X} &= \left(\boldsymbol{I}_n - \boldsymbol{P_X}\right)\left(\boldsymbol{I}_n - \boldsymbol{P_X}\right) \\ &= \boldsymbol{I}_n - 2\boldsymbol{P_X} + \boldsymbol{P_X P_X} \\ &= \boldsymbol{I}_n - \boldsymbol{P_X} \\ &= \boldsymbol{M_X}. \end{aligned}$$

- Orthogonal:

$$\begin{aligned} \boldsymbol{P_X M_X} &= \boldsymbol{P_X}\left(\boldsymbol{I}_n - \boldsymbol{P_X}\right) \\ &= \boldsymbol{P_X} - \boldsymbol{P_X P_X} \\ &= \boldsymbol{P_X} - \boldsymbol{P_X} \\ &= \boldsymbol{0}. \end{aligned}$$

  This property implies that $\boldsymbol{M_X X} = 0$:

$$\begin{aligned} \boldsymbol{M_X X} &= \left(\boldsymbol{I}_n - \boldsymbol{P_X}\right)\boldsymbol{X} \\ &= \boldsymbol{X} - \boldsymbol{P_X X} \\ &= \boldsymbol{X} - \boldsymbol{X} \\ &= \boldsymbol{0}. \end{aligned}$$

Note that, in the above discussion, none of "statistical assumptions" (such as $\mathbb{E}\left(e_i|\boldsymbol{X}_i\right) = 0$) have been used. Given data, $\boldsymbol{Y}$ and $\boldsymbol{X}$, one can always perform least squares, regardless of what data generating process stands behind the data. However, one needs a model to discuss the statistical properties of an estimator (such as unbiasedness and etc).

Partitioned regression

We can partition the matrix of regressors $\boldsymbol{X}$ as follows:

$$\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2],$$

and write the model as

$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e},$$

where $\boldsymbol{X}_1$ is a $n \times k_1$ matrix, $\boldsymbol{X}_2$ is $n \times k_2$, $k_1 + k_2 = k$, and

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix},$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are k_1 and k_2-vectors respectively. Such a decomposition allows one to focus on a group of variables and their corresponding parameters, say $\boldsymbol{X}_1$ and $\boldsymbol{\beta}_1$. If

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix},$$

then one can write the following version of the normal equations (first-order conditions of the least square):

$$\left(\boldsymbol{X}'\boldsymbol{X}\right)\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{Y}$$

as

$$\begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{X}_1 & \boldsymbol{X}_1'\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_2 \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1'\boldsymbol{Y} \\ \boldsymbol{X}_2'\boldsymbol{Y} \end{pmatrix}.$$

One can obtain the expressions for $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ by inverting the partitioned matrix on the left-hand side of the equation above.

Alternatively, let's define $\boldsymbol{M}_2$ to be the projection matrix on the space orthogonal to the space $\mathcal{S}\left(\boldsymbol{X}_2\right)$:

$$\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2\left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'.$$

Then,

$$\widehat{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{Y}. \tag{2}$$

In order to show that, first write

$$\boldsymbol{Y} = \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2 + \widehat{\boldsymbol{e}}. \tag{3}$$

Note that by the construction,

$$\begin{aligned} \boldsymbol{M}_2\widehat{\boldsymbol{e}} &= \widehat{\boldsymbol{e}} \text{ } (\widehat{\boldsymbol{e}} \text{ is orthogonal to } \boldsymbol{X}_2), \\ \boldsymbol{M}_2\boldsymbol{X}_2 &= \boldsymbol{0}, \\ \boldsymbol{X}_1'\widehat{\boldsymbol{e}} &= \boldsymbol{0}, \\ \boldsymbol{X}_2'\widehat{\boldsymbol{e}} &= \boldsymbol{0}. \end{aligned}$$

Substitute equation (3) into the right-hand side of equation (2):

$$\begin{aligned}
&\left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{Y}\\
&=\left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\left(\boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1+\boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2+\widehat{\boldsymbol{e}}\right)\\
&=\left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1\\
&+\left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\widehat{\boldsymbol{e}}\qquad(\boldsymbol{M}_2\boldsymbol{X}_2=\boldsymbol{0}\text{ and }\boldsymbol{M}_2\widehat{\boldsymbol{e}}=\widehat{\boldsymbol{e}})\\
&=\widehat{\boldsymbol{\beta}}_1.
\end{aligned}$$

Since $\boldsymbol{M}_2$ is symmetric and idempotent, one can write

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_1&=\left((\boldsymbol{M}_2\boldsymbol{X}_1)'(\boldsymbol{M}_2\boldsymbol{X}_1)\right)^{-1}(\boldsymbol{M}_2\boldsymbol{X}_1)'(\boldsymbol{M}_2\boldsymbol{Y})\\
&=\left(\widetilde{\boldsymbol{X}}_1'\widetilde{\boldsymbol{X}}_1\right)^{-1}\widetilde{\boldsymbol{X}}_1'\widetilde{\boldsymbol{Y}},
\end{aligned}$$

where

$$\begin{aligned}
\widetilde{\boldsymbol{X}}_1&=\boldsymbol{M}_2\boldsymbol{X}_1\\
&=\boldsymbol{X}_1-\boldsymbol{X}_2\left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1\text{ residuals from the regression of columns of }\boldsymbol{X}_1\text{ on }\boldsymbol{X}_2,\\
\widetilde{\boldsymbol{Y}}&=\boldsymbol{M}_2\boldsymbol{Y}\\
&=\boldsymbol{Y}-\boldsymbol{X}_2\left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\boldsymbol{Y}\text{ residuals from the regression of }\boldsymbol{Y}\text{ on }\boldsymbol{X}_2.
\end{aligned}$$

Thus, to obtain coefficients for the first k_1 regressors, instead of running the full regression with k_1+k_2 regressors, one can regress $\boldsymbol{Y}$ on $\boldsymbol{X}_2$ to obtain the residuals $\widetilde{\boldsymbol{Y}}$, regress $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$ to obtain the residuals $\widetilde{\boldsymbol{X}}_1$, and then regress $\widetilde{\boldsymbol{Y}}$ on $\widetilde{\boldsymbol{X}}_1$ to obtain $\widehat{\boldsymbol{\beta}}_1$. In other words, $\widehat{\boldsymbol{\beta}}_1$ shows the effect of $\boldsymbol{X}_1$ *after controlling* for $\boldsymbol{X}_2$.

Similarly to $\widehat{\boldsymbol{\beta}}_1$, one can write:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_2&=\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{Y},\text{ where}\\
\boldsymbol{M}_1&=\boldsymbol{I}_n-\boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'.
\end{aligned}$$

For example, consider a simple regression

$$Y_i=\beta_1+\beta_2X_i+e_i,$$

for $i=1,\ldots,n$.

Let's define a n-vector of ones:

$$\boldsymbol{1}=\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}_{n\times1}.$$

In this case, the matrix of regressors is given by

$$\begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} \mathbf{1} & \boldsymbol{X} \end{pmatrix}.$$

Consider

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}',$$

and

$$\widehat{\beta}_2 = \frac{\boldsymbol{X}'\boldsymbol{M}_1\boldsymbol{Y}}{\boldsymbol{X}'\boldsymbol{M}_1\boldsymbol{X}}.$$

Now, $\mathbf{1}'\mathbf{1} = n$. Therefore,

$$\begin{aligned} \boldsymbol{M}_1 &= \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}', \text{ and} \\ \boldsymbol{M}_1\boldsymbol{X} &= \boldsymbol{X} - \mathbf{1}\frac{\mathbf{1}'\boldsymbol{X}}{n} \\ &= \boldsymbol{X} - \overline{X}\mathbf{1} \\ &= \begin{pmatrix} X_1 - \overline{X} \\ X_2 - \overline{X} \\ \vdots \\ X_n - \overline{X} \end{pmatrix}, \end{aligned}$$

where $\overline{X}$ is the sample average:

$$\begin{aligned} \overline{X} &= \frac{\mathbf{1}'\boldsymbol{X}}{n} \\ &= n^{-1}\sum_{i=1}^{n} X_i. \end{aligned}$$

Thus, the matrix $\boldsymbol{M}_1$ transforms the vector $\boldsymbol{X}$ into the vector of deviations from the average. We can write

$$\begin{aligned} \widehat{\beta}_2 &= \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)Y_i}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \\ &= \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}. \end{aligned}$$

Goodness of fit

Write

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{Y} + \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{Y} \\ &= \widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{e}}, \end{aligned}$$

where, by the construction,

$$\begin{aligned}\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{e}} &= (\boldsymbol{P_X Y})'(\boldsymbol{M_X Y}) \\ &= \boldsymbol{Y}'\boldsymbol{P_X M_X Y} \\ &= 0.\end{aligned}$$

Suppose that the model contains an intercept, i.e. the first column of $\boldsymbol{X}$ is the vector of ones $\mathbf{1}$. The *total variation* in $\boldsymbol{Y}$ is

$$\begin{aligned}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 &= \boldsymbol{Y}'\boldsymbol{M}_1\boldsymbol{Y} \\ &= \left(\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{e}}\right)'\boldsymbol{M}_1\left(\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{e}}\right) \\ &= \widehat{\boldsymbol{Y}}'\boldsymbol{M}_1\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{Y}}'\boldsymbol{M}_1\widehat{\boldsymbol{e}} + 2\widehat{\boldsymbol{Y}}'\boldsymbol{M}_1\widehat{\boldsymbol{e}}.\end{aligned}$$

Since the model contains an intercept,

$$\begin{aligned}\mathbf{1}'\widehat{\boldsymbol{e}} &= 0, \text{ and} \\ \boldsymbol{M}_1\widehat{\boldsymbol{e}} &= \widehat{\boldsymbol{e}}.\end{aligned}$$

However, $\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{e}} = 0$, and, therefore,

$$\begin{aligned}\boldsymbol{Y}'\boldsymbol{M}_1\boldsymbol{Y} &= \widehat{\boldsymbol{Y}}'\boldsymbol{M}_1\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}, \text{ or} \\ \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 &= \sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{\widehat{Y}}\right)^2 + \sum_{i=1}^{n}\widehat{e}_i^2.\end{aligned}$$

Note that

$$\begin{aligned}\overline{Y} &= \frac{\mathbf{1}'\boldsymbol{Y}}{n} \\ &= \frac{\mathbf{1}'\widehat{\boldsymbol{Y}}}{n} + \frac{\mathbf{1}'\widehat{\boldsymbol{e}}}{n} \\ &= \frac{\mathbf{1}'\widehat{\boldsymbol{Y}}}{n} \\ &= \overline{\widehat{Y}}.\end{aligned}$$

Hence, the averages of $\boldsymbol{Y}$ and its predicted values $\widehat{\boldsymbol{Y}}$ are equal, and we can write:

$$\begin{aligned}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 &= \sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n}\widehat{e}_i^2, \text{ or} \\ TSS &= ESS + RSS,\end{aligned} \tag{4}$$

where

$$TSS = \sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 \text{ total sum-of-squares,}$$

$$
\begin{aligned}
ESS &= \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2 \text{ explained sum-of-squares,} \\
RSS &= \sum_{i=1}^{n} \widehat{e}_i^2 \text{ residual sum-of-squares.}
\end{aligned}
$$

The ratio of the ESS to the TSS is called the *coefficient of determination* or R^2:

$$
\begin{aligned}
R^2 &= \frac{\sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} \\
&= 1 - \frac{\sum_{i=1}^{n} \widehat{e}_i^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} \\
&= 1 - \frac{\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}}{\boldsymbol{Y}'\boldsymbol{M}_1\boldsymbol{Y}}.
\end{aligned}
$$

Properties of R^2:

- Bounded between 0 and 1 as implied by decomposition (4). This property does not hold if the model does not have an intercept, and one should not use the above definition of R^2 in this case. If $R^2 = 1$ then $\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}} = 0$, which can happen only if $\boldsymbol{Y} \in \mathcal{S}(\boldsymbol{X})$, i.e. $\boldsymbol{Y}$ is *exactly* a linear combination of the columns of $\boldsymbol{X}$.

- R^2 increases by adding more regressors. Suppose we have n observations on regressors $(Z_1, ..., Z_k)$ and $(W_1, ..., W_m)$ and dependent variable Y. Consider two regressions: the "long" regression with all regressors and the "short" regression with only $(Z_1, ..., Z_k)$. It can be shown that the R^2 of the long regression must be smaller or equal to the R^2 of the short regression.

- R^2 shows how much of the *sample* variation in $\boldsymbol{Y}$ was explained by $\boldsymbol{X}$. However, our objective is to estimate *population* relationships and not to explain the *sample* variation. High R^2 is not necessary an indicator of the good regression model, and a low R^2 is not an evidence against it.

- Since R^2 increases with inclusion of additional regressors, instead researchers often report the *adjusted coefficient of determination* $\overline{R}^2$:

$$
\begin{aligned}
\overline{R}^2 &= 1 - \frac{n-1}{n-k}\left(1 - R^2\right) \\
&= 1 - \frac{\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}/\left(n-k\right)}{\boldsymbol{Y}'\boldsymbol{M}_1\boldsymbol{Y}/\left(n-1\right)}.
\end{aligned}
$$

  The adjusted coefficient of determination discounts the fit when the number of the regressors k is large relative to the number of observations n. $\overline{R}^2$ may decrease with k. However, there is no strong argument for using such an adjustment.