The general linear model: estimation

Political Science 307

January 25, 2022

Suppose we posit a theoretical process that generates Y_i as a linear function of K variables $X_{1i}, X_{2i}, \ldots, X_{Ki}$ and a stochastic error ϵ_i:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_K X_{Ki} + \epsilon_i.$$

Writing the variables Y_i, X_{ki}, and ϵ_i as vectors, the process can be summarized compactly as

$$\mathbf{y} = \beta_1 \mathbf{1} + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 \cdots + \beta_K \mathbf{x}_K + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix} = \beta_1 \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_2 \begin{bmatrix} x_{21} \\ x_{22} \\ x_{23} \\ \vdots \\ x_{2N} \end{bmatrix} + \beta_3 \begin{bmatrix} x_{31} \\ x_{32} \\ x_{33} \\ \vdots \\ x_{3N} \end{bmatrix} + \cdots + \beta_K \begin{bmatrix} x_{K1} \\ x_{K2} \\ x_{K3} \\ \vdots \\ x_{KN} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_{21} & x_{31} & \cdots & x_{K1} \\ 1 & x_{22} & x_{32} & \cdots & x_{K2} \\ 1 & x_{23} & x_{33} & \cdots & x_{K3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{2N} & x_{3N} & \cdots & x_{KN} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Sampling from the population, we use the N observations to estimate the empirical model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}.$$

The estimation criterion, of course, is to

choose the vector $\mathbf{b}$ to minimize $\mathbf{e}'\mathbf{e}$.

The solution, and all related estimators, will be a function of a matrix of product moments of the independent variables $\mathbf{X}$,

$$\begin{aligned}\mathbf{X}'\mathbf{X} &= \begin{bmatrix} \mathbf{x}_1'\mathbf{x}_1 & \mathbf{x}_1'\mathbf{x}_2 & \mathbf{x}_1'\mathbf{x}_3 & \cdots & \mathbf{x}_1'\mathbf{x}_K \\ \mathbf{x}_1'\mathbf{x}_2 & \mathbf{x}_2'\mathbf{x}_2 & \mathbf{x}_2'\mathbf{x}_3 & \cdots & \mathbf{x}_2'\mathbf{x}_K \\ \mathbf{x}_1'\mathbf{x}_3 & \mathbf{x}_2'\mathbf{x}_3 & \mathbf{x}_3'\mathbf{x}_3 & \cdots & \mathbf{x}_3'\mathbf{x}_K \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_1'\mathbf{x}_K & \mathbf{x}_2'\mathbf{x}_K & \mathbf{x}_3'\mathbf{x}_K & \cdots & \mathbf{x}_K'\mathbf{x}_K \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^N x_{1i}^2 & \sum_{i=1}^N x_{1i}x_{2i} & \sum_{i=1}^N x_{1i}x_{3i} & \cdots & \sum_{i=1}^N x_{1i}x_{Ki} \\ \sum_{i=1}^N x_{1i}x_{2i} & \sum_{i=1}^N x_{2i}^2 & \sum_{i=1}^N x_{2i}x_{3i} & \cdots & \sum_{i=1}^N x_{2i}x_{Ki} \\ \sum_{i=1}^N x_{1i}x_{3i} & \sum_{i=1}^N x_{2i}x_{3i} & \sum_{i=1}^N x_{3i}^2 & \cdots & \sum_{i=1}^N x_{3i}x_{Ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^N x_{1i}x_{Ki} & \sum_{i=1}^N x_{2i}x_{Ki} & \sum_{i=1}^N x_{3i}x_{Ki} & \cdots & \sum_{i=1}^N x_{Ki}^2 \end{bmatrix},\end{aligned}$$

a matrix with sums of squares on the diagonal and sums of cross products off the diagonal, and a vector of the product moments of the independent variables $\mathbf{X}$ and the dependent variable $\mathbf{y}$,

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \mathbf{x}_1'\mathbf{y} \\ \mathbf{x}_2'\mathbf{y} \\ \mathbf{x}_3'\mathbf{y} \\ \vdots \\ \mathbf{x}_K'\mathbf{y} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^N x_{1i}y_i \\ \sum_{i=1}^N x_{2i}y_i \\ \sum_{i=1}^N x_{3i}y_i \\ \vdots \\ \sum_{i=1}^N x_{Ki}y_i, \end{bmatrix},$$

a vector of sums of cross products.

1 The Gauss-Markov assumptions

To derive the estimator, establish that it is unbiased, and prove that it is efficient, we require four assumptions, again named for Gauss and Markov, each generalizations of the assumptions of the bivariate linear model.

1. $\text{rank}(\mathbf{X}) = K$.
   The regressor matrix $\mathbf{X}$ has dimensions $N \times K$, where N is the number of observations and K is the number of regressors (including the constant). Thus, the assumption requires that $\mathbf{X}$ has *full column rank*, that there are no linear dependencies among the columns, that there is no perfect *collinearity* among the variables. One important consequence of the assumption is that $(\mathbf{X}'\mathbf{X})^{-1}$ exists, as we shall require. Furthermore,

$\text{rank}(\mathbf{X}) = K \Rightarrow \mathbf{X}'\mathbf{X}$ is positive definite, as is easily proved: If $\mathbf{X}$ has full column rank, $\mathbf{Xa} = \mathbf{0}$ if and only if $\mathbf{a} = \mathbf{0}$. Let $\mathbf{Xa} = \mathbf{t}$. Because $\mathbf{X}$ has full column rank, $\mathbf{t} \neq \mathbf{0}$, so

$$\mathbf{a}'\mathbf{X}'\mathbf{X}\mathbf{a} = \mathbf{t}'\mathbf{t} = \sum_i t_i^2 > 0$$

and $\mathbf{X}'\mathbf{X}$ is positive definite.

2. $E(\boldsymbol{\epsilon}) = \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_N) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}.$

   That is, each residual has expectation 0.

3. $\mathbf{X}$ is nonstochastic.
   This assumption implies that

   - $\mathbf{X}$ is fixed in repeated samples.
   - $\mathbf{X}$ is measured without error.
   - $\mathbf{X}$ is exogenous.

   Further, together with Assumption 2, $E(\boldsymbol{\epsilon}) = \mathbf{0}$, Assumption 3 implies

$$\begin{aligned} E(\mathbf{X}'\boldsymbol{\epsilon}) &= \mathbf{X}'E(\boldsymbol{\epsilon}) \\ &= \mathbf{0}. \end{aligned}$$

4. $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \begin{bmatrix} E(\epsilon_1^2) & E(\epsilon_1\epsilon_2) & \cdots & E(\epsilon_1\epsilon_N) \\ E(\epsilon_1\epsilon_2) & E(\epsilon_2^2) & \cdots & E(\epsilon_2\epsilon_N) \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_1\epsilon_N) & E(\epsilon_2\epsilon_N) & \cdots & E(\epsilon_N^2) \end{bmatrix} = \begin{bmatrix} \sigma_\epsilon^2 & 0 & \cdots & 0 \\ 0 & \sigma_\epsilon^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_\epsilon^2 \end{bmatrix} = \sigma_\epsilon^2\mathbf{I}.$

   This assumption encompasses both of the i.i.d. assumptions of the bivariate case:

   - The residuals are homoscedastic: $E(\epsilon_i^2) = \sigma_\epsilon^2$.
   - The residuals are independent across observations, not serially autocorrelated or spatially autocorrelated: $E(\epsilon_i\epsilon_j) = 0, \ \forall\ i \neq j$.

   When this assumption is met, the disturbances are said to be *spherical*.

2 Derivation of the least squares estimator

To derive the ordinary least squares estimator for $\boldsymbol{\beta}$, we minimize $\mathbf{e}'\mathbf{e}$ with respect to $\mathbf{b}$. To do so, we will, as before, take the derivative of the function $\mathbf{e}'\mathbf{e}$ with respect to $\mathbf{b}$, set it equal to $\mathbf{0}$, and solve for $\mathbf{b}$. Because $\mathbf{e}'\mathbf{e}$ is a multivariable function in the coefficients $\mathbf{b}$, however, we first need a brief review of multivariable differential calculus.

2.1 Vectors of derivatives: A digression

$\partial f(\mathbf{a})/\partial \mathbf{a}$ is a vector of first partial derivatives of a multivariable function. For each of the k elements in $\mathbf{a}$, we take the partial derivative of $f(\mathbf{a})$ with respect to that element, which becomes the kth element in the vector of partial derivatives:

$$\frac{\partial}{\partial \mathbf{a}} f(\mathbf{a}) = \begin{bmatrix} \frac{\partial f(\mathbf{a})}{\partial a_1} \\ \frac{\partial f(\mathbf{a})}{\partial a_2} \\ \vdots \\ \frac{\partial f(\mathbf{a})}{\partial a_K} \end{bmatrix}.$$

$\partial f(\mathbf{a})/\partial a_k$ is the first *partial derivative* of $f(\mathbf{a})$ with respect to a_k. Because $f(\mathbf{a})$ is a function of $a_1, a_2, \ldots, a_K$, each derivative is "partial" in that it depends upon the values of the other elements of $\mathbf{a}$. In loose terms, it is the derivative of $f(\mathbf{a})$ with respect to, say, a_1 holding $a_2, a_3, \ldots, a_K$ constant. The partial derivative is calculated just like any other derivative: in taking the derivative with respect to a_k, we treat the other $N-1$ variables $a_{(k)}$ as constants. In terms of interpretation, however, the multivariable domain of $f(\mathbf{a})$ makes a substantial difference. A multivariable function defines a surface (e.g., a paraboloid), and the set of partial derivatives defines a plane (or hyperplane) tangent to it. Accordingly, the vector of first partial derivatives is called the *gradient* of $f(\mathbf{a})$, designated $\nabla f(\mathbf{a})$.

Example. Let $f(a_1, a_2) = 2a_1^2 a_2 + 8a_2^2$. Then

$$\frac{\partial}{\partial \mathbf{a}} f(a_1, a_2) = \begin{bmatrix} \partial f(a_1, a_2)/\partial a_1 \\ \partial f(a_1, a_2)/\partial a_2 \end{bmatrix} = \begin{bmatrix} \partial/\partial a_1 (2a_1^2 a_2 + 8a_2^2) \\ \partial/\partial a_2 (2a_1^2 a_2 + 8a_2^2) \end{bmatrix} = \begin{bmatrix} 4a_1 a_2 \\ 2a_1^2 + 16a_2 \end{bmatrix}.$$

2.1.1 Derivatives of vector products

The derivatives of vector products are particularly simple, owing to the linearity of vector products. The product

$$\mathbf{z}'\mathbf{a} = z_1 a_1 + z_2 a_2 + \cdots + z_K a_K,$$

for instance, is a linear function in K variables a_k. If we evaluate its rate of change with respect to the elements of $\mathbf{a}$ we find

$$\begin{aligned}
\frac{\partial}{\partial a_1}\mathbf{z}'\mathbf{a} = \frac{\partial}{\partial a_1}(z_1a_1 + z_2a_2 + \cdots + z_Ka_K) &= z_1 \\
\frac{\partial}{\partial a_2}\mathbf{z}'\mathbf{a} = \frac{\partial}{\partial a_2}(z_1a_1 + z_2a_2 + \cdots + z_Ka_K) &= z_2 \\
\frac{\partial}{\partial a_3}\mathbf{z}'\mathbf{a} = \frac{\partial}{\partial a_3}(z_1a_1 + z_2a_2 + \cdots + z_Ka_K) &= z_3,
\end{aligned}$$

and so on, or more succinctly

$$\frac{\partial \mathbf{z}'\mathbf{a}}{\partial \mathbf{a}} = \begin{bmatrix} \partial \mathbf{z}'\mathbf{a}/\partial a_1 \\ \partial \mathbf{z}'\mathbf{a}/\partial a_2 \\ \vdots \\ \partial \mathbf{z}'\mathbf{a}/\partial a_K \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_K \end{bmatrix} = \mathbf{z}.$$

Correspondingly,

$$\frac{\partial \mathbf{z}'\mathbf{a}}{\partial \mathbf{a}'} = [\partial \mathbf{z}'\mathbf{a}/\partial a_1 \;\; \partial \mathbf{z}'\mathbf{a}/\partial a_2 \;\; \cdots \;\; \partial \mathbf{z}'\mathbf{a}/\partial a_K] = [z_1 \;\; z_2 \;\; \cdots \;\; z_K] = \mathbf{z}'.$$

The slope of a linear multivariable function, that is, is constant in all dimensions. A linear multivariable function defines a plane (in $\mathbb{R}^3$) or a hyperplane (in higher-order spaces).

Extending the idea, the partial derivative of a matrix $\mathbf{Z}$ postmultiplied by a vector $\mathbf{a}$ with respect to its first element a_1 is

$$\frac{\partial \mathbf{Z}\mathbf{a}}{\partial a_1} = \begin{bmatrix} \frac{\partial}{\partial a_1}(z_{11}a_1 + z_{12}a_2 + \cdots + z_{1K}a_K) \\ \frac{\partial}{\partial a_1}(z_{21}a_1 + z_{22}a_2 + \cdots + z_{2K}a_K) \\ \vdots \\ \frac{\partial}{\partial a_1}(z_{K1}a_1 + z_{K2}a_2 + \cdots + z_{KK}a_K) \end{bmatrix} = \begin{bmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{K1} \end{bmatrix},$$

the first column of $\mathbf{Z}$. With respect to its second element a_2 it is

$$\frac{\partial \mathbf{Z}\mathbf{a}}{\partial a_2} = \begin{bmatrix} \frac{\partial}{\partial a_2}(z_{11}a_1 + z_{12}a_2 + \cdots + z_{1K}a_K) \\ \frac{\partial}{\partial a_2}(z_{21}a_1 + z_{22}a_2 + \cdots + z_{2K}a_K) \\ \vdots \\ \frac{\partial}{\partial a_2}(z_{K1}a_1 + z_{K2}a_2 + \cdots + z_{KK}a_K) \end{bmatrix} = \begin{bmatrix} z_{12} \\ z_{22} \\ \vdots \\ z_{K2} \end{bmatrix},$$

the second column of $\mathbf{Z}$, and so forth. Thus,

$$\frac{\partial \mathbf{Z}\mathbf{a}}{\partial \mathbf{a}} = \mathbf{Z}.$$

Finally, take $\mathbf{a}'\mathbf{Z}\mathbf{a}$, which is a quadratic in $\mathbf{a}$. The partial derivative of $\mathbf{a}'\mathbf{Z}\mathbf{a}$ with respect to the first element a_1 is

$$\begin{aligned}\frac{\partial \mathbf{a}'\mathbf{Z}\mathbf{a}}{\partial a_1} &= \frac{\partial}{\partial a_1}(z_{11}a_1^2 + z_{22}a_2^2 + \cdots + z_{KK}a_K^2 + (z_{12}+z_{21})a_1a_2 + (z_{13}+z_{31})a_1a_3 + \cdots) \\ &= 2z_{11}a_1 + (z_{12}+z_{21})a_2 + (z_{13}+z_{31})a_3 + \cdots + (z_{1K}+z_{K1})a_K \\ &= (\mathbf{z}_{r1} + \mathbf{z}_{c1})'\mathbf{a},\end{aligned}$$

where $\mathbf{z}_{r1}$ is the first row and $\mathbf{z}_{c1}$ is the first column of $\mathbf{Z}$. Likewise, the partial derivative of $\mathbf{a}'\mathbf{Z}\mathbf{a}$ with respect to a_2 is

$$\begin{aligned}\frac{\partial \mathbf{a}'\mathbf{Z}\mathbf{a}}{\partial a_2} &= \frac{\partial}{\partial a_1}(z_{11}a_1^2 + z_{22}a_2^2 + \cdots + z_{KK}a_K^2 + (z_{21}+z_{12})a_1a_2 + \cdots + (z_{23}+z_{32})a_2a_3 + \cdots) \\ &= 2z_{22}a_2 + (z_{21}+z_{12})a_1 + (z_{23}+z_{32})a_3 + \cdots + (z_{2K}+z_{K2})a_K \\ &= (\mathbf{z}_{r2} + \mathbf{z}_{c2})'\mathbf{a},\end{aligned}$$

where $\mathbf{z}_{r2}$ and $\mathbf{z}_{c2}$ are the second row and second column of $\mathbf{Z}$, and so on. Gathered together, then,

$$\frac{\partial \mathbf{a}'\mathbf{Z}\mathbf{a}}{\partial \mathbf{a}} = \begin{bmatrix} 2z_{11} & z_{12}+z_{21} & z_{13}+z_{31} & \cdots & z_{1K}+z_{K1} \\ z_{21}+z_{12} & 2z_{22} & z_{23}+z_{32} & \cdots & z_{2K}+z_{K2} \\ z_{31}+z_{13} & z_{32}+z_{23} & 2z_{33} & \cdots & z_{3K}+z_{K3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{K1}+z_{1K} & z_{K2}+z_{2K} & z_{K3}+z_{3K} & \cdots & 2z_{KK} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_K \end{bmatrix} = (\mathbf{Z} + \mathbf{Z}')\mathbf{a}.$$

If $\mathbf{Z}$ is symmetric, as matrices in quadratic forms are usually taken to be, then the derivative reduces to

$$\frac{\partial \mathbf{a}'\mathbf{Z}\mathbf{a}}{\partial \mathbf{a}} = 2\mathbf{Z}\mathbf{a}.$$

2.1.2 Second partial derivatives and the Hessian matrix

The matrix of second partial derivatives of a multivariable function is called the *Hessian matrix*. It is defined as

$$\begin{aligned}\frac{\partial^2}{\partial \mathbf{a}\partial \mathbf{a}'} f(\mathbf{a}) &= \frac{\partial}{\partial \mathbf{a}'}\left(\frac{\partial}{\partial \mathbf{a}} f(\mathbf{a})\right) \\ &= \begin{bmatrix} \partial/\partial a_1 \begin{bmatrix} \partial f(\mathbf{a})/\partial a_1 \\ \partial f(\mathbf{a})/\partial a_2 \\ \vdots \\ \partial f(\mathbf{a})/\partial a_K \end{bmatrix} & \partial/\partial a_2 \begin{bmatrix} \partial f(\mathbf{a})/\partial a_1 \\ \partial f(\mathbf{a})/\partial a_2 \\ \vdots \\ \partial f(\mathbf{a})/\partial a_K \end{bmatrix} & \cdots & \partial/\partial a_K \begin{bmatrix} \partial f(\mathbf{a})/\partial a_1 \\ \partial f(\mathbf{a})/\partial a_2 \\ \vdots \\ \partial f(\mathbf{a})/\partial a_K \end{bmatrix} \end{bmatrix}\end{aligned}$$

$$= \begin{bmatrix} \partial^2 f(\mathbf{a})/\partial a_1^2 & \partial^2 f(\mathbf{a})/\partial a_1 \partial a_2 & \cdots & \partial^2 f(\mathbf{a})/\partial a_1 \partial a_K \\ \partial^2 f(\mathbf{a})/\partial a_2 \partial a_1 & \partial^2 f(\mathbf{a})/\partial a_2^2 & \cdots & \partial^2 f(\mathbf{a})/\partial a_2 \partial a_K \\ \vdots & \vdots & \ddots & \vdots \\ \partial^2 f(\mathbf{a})/\partial a_K \partial a_1 & \partial^2 f(\mathbf{a})/\partial a_K \partial a_2 & \cdots & \partial^2 f(\mathbf{a})/\partial a_K^2 \end{bmatrix}.$$

Because $\partial^2 f(\mathbf{a})/\partial a_i \partial a_j = \partial^2 f(\mathbf{a})/\partial a_j \partial a_i$, the Hessian is symmetric.

Example. Again, let $f(a_1, a_2) = 2a_1^2 a_2 + 8a_2^2$. Then

$$\begin{aligned} \frac{\partial^2}{\partial \mathbf{a} \partial \mathbf{a}'} f(a_1, a_2) &= \begin{bmatrix} \partial^2 f(a_1, a_2)/\partial a_1^2 & \partial^2 f(a_1, a_2)/\partial a_1 \partial a_2 \\ \partial^2 f(a_1, a_2)/\partial a_2 \partial a_1 & \partial^2 f(a_1, a_2)/\partial a_2^2 \end{bmatrix} \\ &= \begin{bmatrix} \partial/\partial a_1(4a_1 a_2) & \partial/\partial a_2(4a_1 a_2) \\ \partial/\partial a_1(2a_1^2 + 16a_2) & \partial/\partial a_2(2a_1^2 + 16a_2) \end{bmatrix} \\ &= \begin{bmatrix} 4a_2 & 4a_1 \\ 4a_1 & 16 \end{bmatrix}. \end{aligned}$$

The Hessian indicates the nature of stationary points in the optimization of a multivariable function: a Hessian that is positive definite indicates a local minimum; a Hessian that is negative definite indicates a local maximum.

2.2 Derivation

With these results in hand, our task is to find

$$\min_{\mathbf{b}} \mathbf{e}'\mathbf{e}.$$

First we substitute for $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b}$,

$$\begin{aligned} \mathbf{e}'\mathbf{e} &= (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{y}'\mathbf{y} - \mathbf{b}'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}. \end{aligned}$$

Differentiating,

$$\frac{\partial}{\partial \mathbf{b}} \mathbf{e}'\mathbf{e} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}.$$

For a minimum,

$$\begin{aligned} \partial \mathbf{e}'\mathbf{e}/\partial \mathbf{b} &= \mathbf{0} \\ -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} &= \mathbf{0} \\ \mathbf{X}'\mathbf{X}\mathbf{b} &= \mathbf{X}'\mathbf{y} \\ \mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \end{aligned}$$

The second order condition for a minimum requires that the Hessian, the matrix of second partial derivatives, $\partial^2 \mathbf{e}'\mathbf{e}/\partial\mathbf{b}\partial\mathbf{b}'$ is positive definite. Here,

$$\frac{\partial^2}{\partial\mathbf{b}\partial\mathbf{b}'}\mathbf{e}'\mathbf{e} = \frac{\partial}{\partial\mathbf{b}'}(-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}) = 2(\mathbf{X}'\mathbf{X})' = 2\mathbf{X}'\mathbf{X},$$

which, as we know, is a positive definite matrix multiplied by a positive scalar and therefore is positive definite. Thus, the estimator $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ minimizes $\mathbf{e}'\mathbf{e}$.

In deriving the least squares estimator $\mathbf{b}$, we used only the first two Gauss-Markov assumptions:

1. $\text{rank}(\mathbf{X}) = K$
2. $E(\boldsymbol{\epsilon}) = \mathbf{0}$.

As before, their use builds them into the estimator, with three mathematical consequences.

- The observed residuals $\mathbf{e}$ are uncorrelated with the regressors $\mathbf{X}$.

$$\begin{aligned}\mathbf{X}'\mathbf{e} &= \mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\mathbf{b} \\ &= \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{0}.\end{aligned}$$

  The source of this mathematical fact of the model is the first order condition for the minimum:

$$\begin{aligned}\partial\mathbf{e}'\mathbf{e}/\partial\mathbf{b} &= \mathbf{0} \\ -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} &= \mathbf{0} \\ -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) &= \mathbf{0} \\ -2\mathbf{X}'\mathbf{e} &= \mathbf{0} \\ \mathbf{X}'\mathbf{e} &= \mathbf{0}.\end{aligned}$$

  In the general model as in the bivariate model, the most important assumption of the least squares model cannot be tested from the data.

- The mean of the residuals equals 0. This has already been shown. Partition $\mathbf{X} = [\mathbf{1} \;\; \mathbf{X}_x]$. Then

$$\begin{aligned}\mathbf{X}'\mathbf{e} &= \mathbf{0} \\ \begin{bmatrix}\mathbf{1}' \\ \mathbf{X}_x'\end{bmatrix}\mathbf{e} &= \begin{bmatrix}0 \\ \mathbf{0}\end{bmatrix} \\ \begin{bmatrix}\mathbf{1}'\mathbf{e} \\ \mathbf{X}_x'\mathbf{e}\end{bmatrix} &= \begin{bmatrix}0 \\ \mathbf{0}\end{bmatrix},\end{aligned}$$

which implies $E(\mathbf{e}) = (1/N)\mathbf{1}'\mathbf{e} = 0$.

- The vector of variable means $[\bar{y} \;\; \bar{\mathbf{x}}']$ is on the regression plane. To show this, find $\hat{y}$ for the mean vector $\bar{\mathbf{x}}$:

$$\begin{aligned}\hat{y}_{\bar{x}} &= \bar{\mathbf{x}}'\mathbf{b} \\ &= (1/N)\mathbf{1}'\mathbf{X}\mathbf{b} \\ &= (1/N)\mathbf{1}'(\mathbf{y} - \mathbf{e}) \\ &= (1/N)\mathbf{1}'\mathbf{y} - (1/N)\mathbf{1}'\mathbf{e} \\ &= \bar{y}.\end{aligned}$$

Note as well the following relationships. With $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$,

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}\mathbf{b} = \hat{\mathbf{y}} \tag{1}$$

and

$$(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{e}. \tag{2}$$

In (1), $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is called the *projection* of $\mathbf{y}$ into the column space of $\mathbf{X}$. The projection matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is $N \times N$, symmetric, and idempotent:

$$\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

As (2) shows, the projection matrix subtracted from $\mathbf{I}_N$ and the expression postmultiplied by $\mathbf{y}$ gives the ordinary least squares residuals $\mathbf{e}$. As we will see, the matrix $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, which we will often designate by $\mathbf{M}$, is ubiquitous in least squares. Like the projection matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, the matrix $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is $N \times N$, symmetric, and idempotent:

$$\begin{aligned}\mathbf{M}\mathbf{M} &= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= \mathbf{I} - 2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{I} - 2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{M}.\end{aligned}$$

Its rank, as we will show shortly, is $N - K$. If we premultiply $\mathbf{M}$ by $\mathbf{y}'$, we get

$$\mathbf{y}'\mathbf{M} = \mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \mathbf{y}' - \mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{y}' - \hat{\mathbf{y}}' = \mathbf{e}'.$$

Accordingly,

$$\begin{aligned}\mathbf{y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} &= \mathbf{y}'\mathbf{M}'\mathbf{M}\mathbf{y} \\ &= \mathbf{y}'\mathbf{M}\mathbf{M}\mathbf{y} \\ &= \mathbf{y}'\mathbf{M}\mathbf{y} \\ &= \mathbf{e}'\mathbf{e}\end{aligned}$$

is the sum of squared least squares residuals. You will also see this expression with $\mathbf{y}'$ and $\mathbf{y}$ distributed through the product,

$$\begin{aligned}\mathbf{y}'\mathbf{M}\mathbf{y} &= \mathbf{y}'(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} \\ &= \mathbf{y}'\mathbf{y}-\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{y}'\mathbf{y}-\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{y}'\mathbf{y}-\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}=\mathbf{e}'\mathbf{e}. \qquad (3)\end{aligned}$$

If $\mathbf{y}$ and $\mathbf{X}$ are mean deviated, that is, we have $\dot{\mathbf{y}}=\mathbf{C}\mathbf{y}$ and $\dot{\mathbf{X}}=\mathbf{C}\mathbf{X}$, where

$$\mathbf{C}=(\mathbf{I}-(1/N)\mathbf{1}\mathbf{1}')$$

is the *centering matrix*, then (3) is the decomposition of the sum of squares in the analysis of variance, as we will show in a moment.

3 Interpretation

The generalization of the least squares regression model changes nothing in interpretation except that the coefficients represent the marginal effect of x_k on y, that is, the effect of x_k on y holding the other regressors statistically constant.

Geometrically, the regression model $\mathbf{y}=\mathbf{X}\mathbf{b}+\mathbf{e}$ defines a *plane* in $\mathbb{R}^K$. Least squares minimizes the sum of the squared vertical distances between the data points and the plane; in doing so, it makes the estimated residuals orthogonal to the regression plane: $\mathbf{e}'\hat{\mathbf{y}}=\mathbf{e}'\mathbf{X}\mathbf{b}=0$.

3.1 The bivariate least squares estimator as a special case

The multivariate least squares estimator "looks" a little like the bivariate estimator, but it may be instructive to show that the bivariate estimator is a special case of the multivariate estimator. To begin, partition $\mathbf{X}$ into its columns: $\mathbf{X}=[\mathbf{1}\ \ \mathbf{x}]$. Then,

- $\mathbf{X}'\mathbf{X}=\begin{bmatrix}\mathbf{1}'\\ \mathbf{x}'\end{bmatrix}[\mathbf{1}\ \ \mathbf{x}]=\begin{bmatrix}\mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{x}\\ \mathbf{x}'\mathbf{1} & \mathbf{x}'\mathbf{x}\end{bmatrix}=\begin{bmatrix}N & \sum_i x_i\\ \sum_i x_i & \sum_i x_i^2\end{bmatrix}$;
- $|\mathbf{X}'\mathbf{X}|=N\sum_i x_i^2-(\sum_i x_i)^2$;
- $(\mathbf{X}'\mathbf{X})^{-1}=\frac{1}{N\sum_i x_i^2-(\sum_i x_i)^2}\begin{bmatrix}\sum_i x_i^2 & -\sum_i x_i\\ -\sum_i x_i & N\end{bmatrix}$;
- $\mathbf{X}'\mathbf{y}=\begin{bmatrix}\mathbf{1}'\\ \mathbf{x}'\end{bmatrix}\mathbf{y}=\begin{bmatrix}\mathbf{1}'\mathbf{y}\\ \mathbf{x}'\mathbf{y}\end{bmatrix}=\begin{bmatrix}\sum_i y_i\\ \sum_i x_i y_i\end{bmatrix}$.

The bivariate estimators, then, are:

$$\begin{aligned}
\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \frac{1}{N\sum_i x_i^2 - (\sum_i x_i)^2}\begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & N \end{bmatrix}\begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix} \\
&= \begin{bmatrix} \frac{\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i}{N\sum_i x_i^2 - (\sum_i x_i)^2} \\ \frac{N\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{N\sum_i x_i^2 - (\sum_i x_i)^2} \end{bmatrix} \\
&= \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}.
\end{aligned}$$

Clearly,

$$b_1 = \frac{N\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{N\sum_i x_i^2 - (\sum_i x_i)^2} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2};$$

recall the alternative forms of the variance and covariance.

That

$$b_0 = \frac{N\sum_i x_i^2 - \sum_i x_i \sum_i x_i y_i}{N\sum_i x_i^2 - (\sum_i x_i)^2} = \bar{y} - b_1\bar{x}$$

is less evident, but:

$$\begin{aligned}
\frac{\sum_i x_i^2 \sum_i y_i - \sum_i x_i \sum_i x_i y_i}{N\sum_i x_i^2 - (\sum_i x_i)^2} &= \frac{\bar{y}\sum_i x_i^2 - \bar{x}\sum_i x_i y_i}{\sum_i x_i^2 - (\sum_i x_i)^2/N} \\
&= \frac{\bar{y}(\sum_i x_i^2 - (\sum_i x_i)^2/N) - \bar{x}\sum_i x_i y_i + \bar{y}(\sum_i x_i)^2/N}{\sum_i x_i^2 - (\sum_i x_i)^2/N} \\
&= \bar{y} - \bar{x}\frac{\sum_i x_i y_i - \sum_i x_i \sum_i y_i/N}{\sum_i x_i^2 - (\sum_i x_i)^2/N} \\
&= \bar{y} - b_1\bar{x}.
\end{aligned}$$

3.2 Inside the general linear estimator

In general form, the ordinary least squares estimator is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

which has the advantage of brevity and the disadvantage of obscurity. To get a better sense of the meaning in the calculations, consider the regression of y onto three independent variables, x_1, x_2 and x_3, all variables mean-deviated. The model is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e},$$

with $\mathbf{X} = [\mathbf{x}_1 \;\; \mathbf{x}_2 \;\; \mathbf{x}_3]$. The OLS coefficient vector is

$$\begin{aligned}\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} &= \begin{bmatrix} \mathbf{x}_1'\mathbf{x}_1 & \mathbf{x}_1'\mathbf{x}_2 & \mathbf{x}_1'\mathbf{x}_3 \\ \mathbf{x}_1'\mathbf{x}_2 & \mathbf{x}_2'\mathbf{x}_2 & \mathbf{x}_2'\mathbf{x}_3 \\ \mathbf{x}_1'\mathbf{x}_3 & \mathbf{x}_2'\mathbf{x}_3 & \mathbf{x}_3'\mathbf{x}_3 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x}_1'\mathbf{y} \\ \mathbf{x}_2'\mathbf{y} \\ \mathbf{x}_3'\mathbf{y} \end{bmatrix} \\ &= \begin{bmatrix} (\mathbf{X}'\mathbf{X})^{-1}_{11} & (\mathbf{X}'\mathbf{X})^{-1}_{12} & (\mathbf{X}'\mathbf{X})^{-1}_{13} \\ (\mathbf{X}'\mathbf{X})^{-1}_{21} & (\mathbf{X}'\mathbf{X})^{-1}_{22} & (\mathbf{X}'\mathbf{X})^{-1}_{23} \\ (\mathbf{X}'\mathbf{X})^{-1}_{31} & (\mathbf{X}'\mathbf{X})^{-1}_{32} & (\mathbf{X}'\mathbf{X})^{-1}_{33} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1'\mathbf{y} \\ \mathbf{x}_2'\mathbf{y} \\ \mathbf{x}_3'\mathbf{y} \end{bmatrix}.\end{aligned}$$

The first coefficient b_1 is the product of the first row of $(\mathbf{X}'\mathbf{X})^{-1}$ and the column vector $\mathbf{X}'\mathbf{y}$. According to the formula for the inverse of a partitioned matrix, in this case $\mathbf{X}'\mathbf{X}$ partitioned at the first row and the first column, the $(1,1)$ element in the inverse is

$$(\mathbf{X}'\mathbf{X})^{-1}_{11} = (\mathbf{x}_1'\mathbf{x}_1 - \mathbf{x}_1'\mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}\mathbf{X}_{23}'\mathbf{x}_1)^{-1},$$

with

$$\mathbf{X}_{23} = [\mathbf{x}_2 \;\; \mathbf{x}_3] \;\; \text{and} \;\; \mathbf{X}_{23}'\mathbf{X}_{23} = \begin{bmatrix} \mathbf{x}_2'\mathbf{x}_2 & \mathbf{x}_2'\mathbf{x}_3 \\ \mathbf{x}_2'\mathbf{x}_3 & \mathbf{x}_3'\mathbf{x}_3 \end{bmatrix},$$

and the second and third elements in the first row of the inverse compose the row vector

$$\begin{aligned}(\mathbf{X}'\mathbf{X})^{-1}_{12,13} &= -(\mathbf{x}_1'\mathbf{x}_1 - \mathbf{x}_1'\mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}\mathbf{X}_{23}'\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1} \\ &= -(\mathbf{X}'\mathbf{X})^{-1}_{11}\mathbf{x}_1'\mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}.\end{aligned}$$

Multiplying out the components, the elements in the first row of $(\mathbf{X}'\mathbf{X})^{-1}$ are

$$\begin{aligned}(\mathbf{X}'\mathbf{X})^{-1}_{11} &= \left(\mathbf{x}_1'\mathbf{x}_1 - [\mathbf{x}_1'\mathbf{x}_2 \;\; \mathbf{x}_1'\mathbf{x}_3]\frac{1}{\mathbf{x}_2'\mathbf{x}_2\mathbf{x}_3'\mathbf{x}_3 - (\mathbf{x}_2'\mathbf{x}_3)^2}\begin{bmatrix} \mathbf{x}_3'\mathbf{x}_3 & -\mathbf{x}_2'\mathbf{x}_3 \\ -\mathbf{x}_2'\mathbf{x}_3 & \mathbf{x}_2'\mathbf{x}_2 \end{bmatrix}\begin{bmatrix} \mathbf{x}_1'\mathbf{x}_2 \\ \mathbf{x}_1'\mathbf{x}_3 \end{bmatrix}\right)^{-1} \\ &= \left(\mathbf{x}_1'\mathbf{x}_1 - \frac{\mathbf{x}_3'\mathbf{x}_3(\mathbf{x}_1'\mathbf{x}_2)^2 - 2\mathbf{x}_1'\mathbf{x}_2\mathbf{x}_1'\mathbf{x}_3\mathbf{x}_2'\mathbf{x}_3 + \mathbf{x}_2'\mathbf{x}_2(\mathbf{x}_1'\mathbf{x}_3)^2}{\mathbf{x}_2'\mathbf{x}_2\mathbf{x}_3'\mathbf{x}_3 - (\mathbf{x}_2'\mathbf{x}_3)^2}\right)^{-1} \\ &= \left(\frac{\mathbf{x}_1'\mathbf{x}_1\mathbf{x}_2'\mathbf{x}_2\mathbf{x}_3'\mathbf{x}_3 - \mathbf{x}_1'\mathbf{x}_1(\mathbf{x}_2'\mathbf{x}_3)^2 - \mathbf{x}_2'\mathbf{x}_2(\mathbf{x}_1'\mathbf{x}_3)^2 - \mathbf{x}_3'\mathbf{x}_3(\mathbf{x}_1'\mathbf{x}_2)^2 + 2\mathbf{x}_1'\mathbf{x}_2\mathbf{x}_1'\mathbf{x}_3\mathbf{x}_2'\mathbf{x}_3}{\mathbf{x}_2'\mathbf{x}_2\mathbf{x}_3'\mathbf{x}_3 - (\mathbf{x}_2'\mathbf{x}_3)^2}\right)^{-1} \\ &= \frac{\mathbf{x}_2'\mathbf{x}_2\mathbf{x}_3'\mathbf{x}_3 - (\mathbf{x}_2'\mathbf{x}_3)^2}{\mathbf{x}_1'\mathbf{x}_1\mathbf{x}_2'\mathbf{x}_2\mathbf{x}_3'\mathbf{x}_3 - \mathbf{x}_1'\mathbf{x}_1(\mathbf{x}_2'\mathbf{x}_3)^2 - \mathbf{x}_2'\mathbf{x}_2(\mathbf{x}_1'\mathbf{x}_3)^2 - \mathbf{x}_3'\mathbf{x}_3(\mathbf{x}_1'\mathbf{x}_2)^2 + 2\mathbf{x}_1'\mathbf{x}_2\mathbf{x}_1'\mathbf{x}_3\mathbf{x}_2'\mathbf{x}_3},\end{aligned}$$

$$\begin{aligned}(\mathbf{X}'\mathbf{X})^{-1}_{12} &= -\frac{\mathbf{x}_3'\mathbf{x}_3\mathbf{x}_1'\mathbf{x}_2 - \mathbf{x}_1'\mathbf{x}_3\mathbf{x}_2'\mathbf{x}_3}{\mathbf{x}_1'\mathbf{x}_1\mathbf{x}_2'\mathbf{x}_2\mathbf{x}_3'\mathbf{x}_3 - \mathbf{x}_1'\mathbf{x}_1(\mathbf{x}_2'\mathbf{x}_3)^2 - \mathbf{x}_2'\mathbf{x}_2(\mathbf{x}_1'\mathbf{x}_3)^2 - \mathbf{x}_3'\mathbf{x}_3(\mathbf{x}_1'\mathbf{x}_2)^2 + 2\mathbf{x}_1'\mathbf{x}_2\mathbf{x}_1'\mathbf{x}_3\mathbf{x}_2'\mathbf{x}_3}, \\ (\mathbf{X}'\mathbf{X})^{-1}_{13} &= -\frac{\mathbf{x}_2'\mathbf{x}_2\mathbf{x}_1'\mathbf{x}_3 - \mathbf{x}_1'\mathbf{x}_2\mathbf{x}_2'\mathbf{x}_3}{\mathbf{x}_1'\mathbf{x}_1\mathbf{x}_2'\mathbf{x}_2\mathbf{x}_3'\mathbf{x}_3 - \mathbf{x}_1'\mathbf{x}_1(\mathbf{x}_2'\mathbf{x}_3)^2 - \mathbf{x}_2'\mathbf{x}_2(\mathbf{x}_1'\mathbf{x}_3)^2 - \mathbf{x}_3'\mathbf{x}_3(\mathbf{x}_1'\mathbf{x}_2)^2 + 2\mathbf{x}_1'\mathbf{x}_2\mathbf{x}_1'\mathbf{x}_3\mathbf{x}_2'\mathbf{x}_3}.\end{aligned}$$

Multiplying this first row of $(\mathbf{X}'\mathbf{X})^{-1}$ by $\mathbf{X}'\mathbf{y}$ we get

$$b_1 = \frac{(\mathbf{x}_2'\mathbf{x}_2\mathbf{x}_3'\mathbf{x}_3 - (\mathbf{x}_2'\mathbf{x}_3)^2)\mathbf{x}_1'\mathbf{y} - (\mathbf{x}_3'\mathbf{x}_3\mathbf{x}_1'\mathbf{x}_2 - \mathbf{x}_1'\mathbf{x}_3\mathbf{x}_2'\mathbf{x}_3)\mathbf{x}_2'\mathbf{y} - (\mathbf{x}_2'\mathbf{x}_2\mathbf{x}_1'\mathbf{x}_3 - \mathbf{x}_1'\mathbf{x}_2\mathbf{x}_2'\mathbf{x}_3)\mathbf{x}_3'\mathbf{y}}{\mathbf{x}_1'\mathbf{x}_1\mathbf{x}_2'\mathbf{x}_2\mathbf{x}_3'\mathbf{x}_3 - \mathbf{x}_1'\mathbf{x}_1(\mathbf{x}_2'\mathbf{x}_3)^2 - \mathbf{x}_2'\mathbf{x}_2(\mathbf{x}_1'\mathbf{x}_3)^2 - \mathbf{x}_3'\mathbf{x}_3(\mathbf{x}_1'\mathbf{x}_2)^2 + 2\mathbf{x}_1'\mathbf{x}_2\mathbf{x}_1'\mathbf{x}_3\mathbf{x}_2'\mathbf{x}_3}.$$

There are several ways to understand the decomposition of the multiple regression estimator. First, we can understand it as a combination of variances and covariances. Because the variables are mean deviated,

$$\mathrm{Var}(x_i) = \frac{\mathbf{x}_i'\mathbf{x}_i}{N} \quad \text{and} \quad \mathrm{Cov}(x_i, x_j) = \frac{\mathbf{x}_i'\mathbf{x}_j}{N}.$$

Letting $V_i = \mathrm{Var}(x_i)$ and $C_{ij} = \mathrm{Cov}(x_i, x_j)$, we can rewrite the expression as

$$b_1 = \frac{(V_2V_3 - C_{23}^2)\mathrm{Cov}(x_1, y) - (V_3C_{12} - C_{13}C_{23})\mathrm{Cov}(x_2, y) - (V_2C_{13} - C_{12}C_{23})\mathrm{Cov}(x_3, y)}{V_1V_2V_3 - V_1C_{23}^2 - V_2C_{12}^2 - V_3C_{13}^2 + 2C_{12}C_{13}C_{23}}.$$

Second, we can understand the coefficient estimator b_1 as a combination of regressions. Look closely at

$$(\mathbf{X}'\mathbf{X})_{11}^{-1} = (\mathbf{x}_1'\mathbf{x}_1 - \mathbf{x}_1'\mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}\mathbf{X}_{23}'\mathbf{x}_1)^{-1}.$$

The vector

$$\mathbf{d}_{1\cdot 23} = (\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}\mathbf{X}_{23}'\mathbf{x}_1,$$

is the coefficients from a regression of x_1 onto x_2 and x_3 and

$$\mathbf{x}_1 - \mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}\mathbf{X}_{23}'\mathbf{x}_1 = (\mathbf{I} - \mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}\mathbf{X}_{23}')\mathbf{x}_1 = \mathbf{M}_{23}\mathbf{x}_1 = \mathbf{e}_{1\cdot 23}$$

is the vector of residuals from the regression of x_1 onto x_2 and x_3. $\mathbf{M}_{23} = \mathbf{I} - \mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}\mathbf{X}_{23}'$ is a symmetric, idempotent matrix. Accordingly, factoring $\mathbf{x}_1'$ from the left and $\mathbf{x}_1$ from the right, we see that

$$(\mathbf{X}'\mathbf{X})_{11}^{-1} = \left(\mathbf{x}_1'(\mathbf{I} - \mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}\mathbf{X}_{23}')\mathbf{x}_1\right)^{-1} = (\mathbf{x}_1'\mathbf{M}_{23}\mathbf{x}_1)^{-1} = (\mathbf{x}_1'\mathbf{M}_{23}'\mathbf{M}_{23}\mathbf{x}_1)^{-1} = (\mathbf{e}_{1\cdot 23}'\mathbf{e}_{1\cdot 23})^{-1}$$

is the reciprocal of the sum of squared residuals from a regression of x_1 on x_2 and x_3.

Examining the rest of the first row of $(\mathbf{X}'\mathbf{X})^{-1}$,

$$(\mathbf{X}'\mathbf{X})_{12,13}^{-1} \;=\; -(\mathbf{X}'\mathbf{X})_{11}^{-1}\mathbf{x}_1'\mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}.$$

in the same way, we see that

$$(\mathbf{X}'\mathbf{X})_{12,13}^{-1} = -\frac{1}{\mathbf{e}_{1\cdot 23}'\mathbf{e}_{1\cdot 23}}\mathbf{d}_{1\cdot 23}'.$$

Multiplying the first row of $(\mathbf{X}'\mathbf{X})^{-1}$ by $\mathbf{X}'\mathbf{y}$, then,

$$b_1 = \frac{\mathbf{x}_1'\mathbf{y}}{\mathbf{e}_{1\cdot 23}'\mathbf{e}_{1\cdot 23}} - d_{2:x_1}\frac{\mathbf{x}_2'\mathbf{y}}{\mathbf{e}_{1\cdot 23}'\mathbf{e}_{1\cdot 23}} - d_{3:x_1}\frac{\mathbf{x}_3'\mathbf{y}}{\mathbf{e}_{1\cdot 23}'\mathbf{e}_{1\cdot 23}} = \frac{\mathbf{x}_1'\mathbf{y} - d_{2:x_1}\mathbf{x}_2'\mathbf{y} - d_{3:x_1}\mathbf{x}_3'\mathbf{y}}{\mathbf{e}_{1\cdot 23}'\mathbf{e}_{1\cdot 23}}.$$

The denominator is the variance in x_1 that is not "explained" by x_2 and x_3, that is, the unique variance of x_1. The numerator is the covariance of x_1 and y minus the covariance of x_2 and y weighted by (conceptually) the partial correlation of x_1 and x_2 and minus the covariance of x_3 and y weighted by the partial correlation of x_1 and x_3. In other words, the numerator is the unique covariance of x_1 and y.

Finally, we can understand the coefficient estimator b_1 as a ratio of residuals. In

$$b_1 = \frac{\mathbf{x}_1'\mathbf{y} - d_{2:x_1}\mathbf{x}_2'\mathbf{y} - d_{3:x_1}\mathbf{x}_3'\mathbf{y}}{\mathbf{e}_{1\cdot 23}'\mathbf{e}_{1\cdot 23}},$$

the denominator is $\mathbf{e}_{1\cdot 23}'\mathbf{e}_{1\cdot 23} = \mathbf{x}_1'\mathbf{M}_{23}\mathbf{x}_1$, the sum of the squared residuals from a regression of x_1 onto x_2 and x_3. The numerator may be rewritten as

$$\begin{aligned}
\mathbf{x}_1'\mathbf{y} - d_{2:x_1}\mathbf{x}_2'\mathbf{y} - d_{3:x_1}\mathbf{x}_3'\mathbf{y} &= \mathbf{x}_1'\mathbf{y} - \mathbf{d}_{1\cdot 23}'\mathbf{X}_{23}'\mathbf{y} \\
&= \mathbf{x}_1'\mathbf{y} - \mathbf{x}_1'\mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}\mathbf{X}_{23}'\mathbf{y} \\
&= \mathbf{x}_1'(\mathbf{I} - \mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}\mathbf{X}_{23}')\mathbf{y} \\
&= \mathbf{x}_1'\mathbf{M}_{23}\mathbf{y}.
\end{aligned}$$

Thus,

$$b_1 = \frac{\mathbf{x}_1'\mathbf{M}_{23}\mathbf{y}}{\mathbf{x}_1'\mathbf{M}_{23}\mathbf{x}_1}.$$

As we have already seen,

$$\mathbf{M}_{23}\mathbf{x}_1 = \mathbf{e}_{1\cdot 23}$$

is the vector of residuals from the regression of x_1 onto x_2 and x_3. Likewise,

$$\mathbf{M}_{23}\mathbf{y} = \mathbf{e}_{y\cdot 23}$$

is the vector of residuals from the regression of y onto x_2 and x_3. If we regress the residuals $\mathbf{e}_{y\cdot 23}$ onto the residuals $\mathbf{e}_{1\cdot 23}$, we get

$$(\mathbf{e}_{1\cdot 23}'\mathbf{e}_{1\cdot 23})^{-1}\mathbf{e}_{1\cdot 23}'\mathbf{e}_{y\cdot 23} = (\mathbf{x}_1'\mathbf{M}_{23}\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{M}_{23}\mathbf{y} = b_1.$$

The multiple regression coefficient b_1 is the same as the coefficient obained by regressing the residuals from a regression of y to x_2 and x_3 onto the residuals from a regression of x_1 on x_2 and x_3. We will make use of this fact later.

4 Unbiasedness of the least squares estimator b

The estimator

$$\begin{aligned}\mathbf{b} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\ &= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}.\end{aligned}$$

$\mathbf{b}$ is an unbiased estimator of $\boldsymbol{\beta}$:

$$\begin{aligned}E(\mathbf{b}) &= E(\boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}) \\ &= E(\boldsymbol{\beta})+E(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\ &= \boldsymbol{\beta}+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta}.\end{aligned}$$

Thus, $\mathbf{b}$ is unbiased. ■

Note the use of three of the Gauss-Markov assumptions in the proof of unbiasedness:

1. $\text{rank}(\mathbf{X}) = K$.
2. $E(\boldsymbol{\epsilon}) = \mathbf{0}$.
3. $\mathbf{X}$ is nonstochastic.

5 The variance of b

The variance of $\mathbf{b}$ is a $K \times K$ matrix of coefficient variances and covariances:

$$\text{Var}(\mathbf{b}) = \begin{bmatrix} \text{Var}(b_1) & \text{Cov}(b_1,b_2) & \text{Cov}(b_1,b_3) & \cdots & \text{Cov}(b_1,b_K) \\ \text{Cov}(b_1,b_2) & \text{Var}(b_2) & \text{Cov}(b_2,b_3) & \cdots & \text{Cov}(b_2,b_K) \\ \text{Cov}(b_1,b_3) & \text{Cov}(b_2,b_3) & \text{Var}(b_3) & \cdots & \text{Cov}(b_3,b_K) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(b_1,b_K) & \text{Cov}(b_2,b_K) & \text{Cov}(b_3,b_K) & \cdots & \text{Var}(b_K) \end{bmatrix},$$

which is symmetric and positive definite (as long as $\mathbf{X}$ has full column rank). It equals

$$\begin{aligned}
\text{Var}(\mathbf{b}) &= E\left(\mathbf{b} - E(\mathbf{b})\right)\left(\mathbf{b} - E(\mathbf{b})\right)' \\
&= E(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})' \\
&= E\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\right)\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\right)' \\
&= E\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma_\epsilon^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}$$

The standard errors of the coefficients are the square roots of the diagonal elements of the matrix $\text{Var}(\mathbf{b})$.

In deriving $\text{Var}(\mathbf{b})$, we required Assumption 4: $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma_\epsilon^2\mathbf{I}$. Thus, if Assumption 4 is violated, $\mathbf{b}$ is still unbiased, but it is no longer efficient (and the estimated standard errors are incorrect).

5.0.1 Inside the least squares coefficient variance estimator

Earlier we considered the regression of a mean-deviated variable y onto three mean-deviated regressors, x_1, x_2, and x_3. Using the formula for the inverse of a partitioned matrix, we found that the $(1,1)$ element of $(\mathbf{X}'\mathbf{X})^{-1}$ is

$$(\mathbf{X}'\mathbf{X})^{-1}_{11} = \left(\mathbf{x}_1'(\mathbf{I} - \mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}\mathbf{X}_{23}')\mathbf{x}_1\right)^{-1} = \left(\mathbf{x}_1'\mathbf{M}_{23}\mathbf{x}_1\right)^{-1} = \left(\mathbf{e}_{1\cdot 23}'\mathbf{e}_{1\cdot 23}\right)^{-1},$$

and therefore

$$\begin{aligned}
\text{Var}(b_1) &= \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}_{11} \\
&= \sigma_\epsilon^2\left(\mathbf{x}_1'(\mathbf{I} - \mathbf{X}_{23}(\mathbf{X}_{23}'\mathbf{X}_{23})^{-1}\mathbf{X}_{23}')\mathbf{x}_1\right)^{-1} \\
&= \sigma_\epsilon^2\left(\mathbf{x}_1'\mathbf{M}_{23}\mathbf{x}_1\right)^{-1} \\
&= \sigma_\epsilon^2\left(\mathbf{e}_{1\cdot 23}'\mathbf{e}_{1\cdot 23}\right)^{-1},
\end{aligned}$$

where $\mathbf{e}_{1\cdot 23}$ is the vector of residuals from the regression of $\mathbf{x}_1$ onto $[\mathbf{x}_2 \;\; \mathbf{x}_3]$. The expression illustrates that

- $\text{SE}(b_k)$ is directly proportional to σ_ϵ^2.
- $\text{SE}(b_k)$ is inversely proportional to $\text{Var}(x_k) \propto \mathbf{x}_k'\mathbf{x}_k$, ceteris paribus.
- $\text{SE}(b_k)$ is inversely proportional to the correlation between x_k and the other independent variables. The better the fit of the regression of x_k onto the other independent variables, the smaller the variance of the residuals from the regression of x_k onto the other regressors. Multicollinearity, that is, increases the coefficient standard errors.

6 The best linear unbiasedness of b

The proof of the efficiency of $\mathbf{b}$ is the Gauss-Markov Theorem. The proof proceeds by creating an arbitrary other unbiased linear estimator and showing that its variance exceeds $\text{Var}(\mathbf{b})$.

The Gauss-Markov Theorem. Let $\mathbf{b}_* = \mathbf{C}\mathbf{y}$, where $\mathbf{C}$ is $K \times N$, be any other linear unbiased estimator of $\boldsymbol{\beta}$. The unbiasedness of $\mathbf{b}_*$ requires

$$\begin{aligned} E(\mathbf{b}_*) &= E(\mathbf{C}\mathbf{y}) \\ &= E(\mathbf{C}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})) \\ &= E\mathbf{C}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}E\boldsymbol{\epsilon} \\ &= \boldsymbol{\beta}, \end{aligned}$$

which implies $\mathbf{C}\mathbf{X} = \mathbf{I}$.

The variance of $\mathbf{b}_*$ is

$$\begin{aligned} \text{Var}(\mathbf{b}_*) &= E(\mathbf{b}_* - \boldsymbol{\beta})(\mathbf{b}_* - \boldsymbol{\beta})' \\ &= E\mathbf{C}\boldsymbol{\epsilon}(\mathbf{C}\boldsymbol{\epsilon})' \\ &= \mathbf{C}E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')\mathbf{C}' \\ &= \sigma_\epsilon^2 \mathbf{C}\mathbf{C}'. \end{aligned}$$

Now let $\mathbf{D} = \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, so that $\mathbf{D}\mathbf{y} = \mathbf{b}_* - \mathbf{b}$. If we postmultiply $\mathbf{D}$ by $\mathbf{X}$, then, we get

$$\begin{aligned} \mathbf{D}\mathbf{X} &= \mathbf{C}\mathbf{X} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} \\ &= \mathbf{I} - \mathbf{I} \\ &= \mathbf{0}. \end{aligned}$$

Using $\mathbf{D} = \mathbf{C} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, we can rewrite

$$\begin{aligned} \text{Var}(\mathbf{b}_*) &= \sigma_\epsilon^2 \mathbf{C}\mathbf{C}' \\ &= \sigma_\epsilon^2 (\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{D} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \\ &= \sigma_\epsilon^2 \left(\mathbf{D}\mathbf{D}' + \mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}' + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) \\ &= \sigma_\epsilon^2 \mathbf{D}\mathbf{D}' + \sigma_\epsilon^2 (\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma_\epsilon^2 \mathbf{D}\mathbf{D}' + \text{Var}(\mathbf{b}). \end{aligned}$$

$\mathbf{D}\mathbf{D}'$ is positive semi-definite: Let $\mathbf{q} = \mathbf{D}'\mathbf{a}$; then

$$\mathbf{a}'\mathbf{D}\mathbf{D}'\mathbf{a} = \mathbf{q}'\mathbf{q} \geq 0.$$

Therefore,

$$\text{Var}(\mathbf{b}) \leq \text{Var}(\mathbf{b}_*)$$

for any arbitrary unbiased estimator $\mathbf{b}_*$, and $\mathbf{b}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$. ■

7 Estimation of σ_ϵ^2

At this point, we have everything we need for estimation except an estimate of σ_ϵ^2. We will want to prove that

$$s_e^2 = \mathbf{e}'\mathbf{e}/(N-K)$$

is an unbiased estimator of σ_ϵ^2.

The proof begins with

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \mathbf{X}\mathbf{b} \\ &= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} \\ &= \mathbf{M}\mathbf{y},\end{aligned}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. We will have frequent occasion to use $\mathbf{M}$, which you have already seen. Recall that

- $\mathbf{M}$ is symmetric: $\mathbf{M}' = \mathbf{M}$.
- $\mathbf{M}$ is idempotent: $\mathbf{M}\mathbf{M} = \mathbf{M}$.
- $\mathbf{M}\mathbf{X} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$.

Thus,

$$\begin{aligned}\mathbf{e} &= \mathbf{M}\mathbf{y} \\ &= \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \mathbf{M}\boldsymbol{\epsilon}.\end{aligned}$$

Because $\mathbf{M}$ is symmetric and idempotent,

$$\mathbf{e}'\mathbf{e} = \boldsymbol{\epsilon}'\mathbf{M}'\mathbf{M}\boldsymbol{\epsilon} = \boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}.$$

Taking expectations,

$$\begin{aligned}E(\mathbf{e}'\mathbf{e}) &= E(\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}) \\ &= E(\mathrm{tr}(\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon})) \\ &= E(\mathrm{tr}(\mathbf{M}\boldsymbol{\epsilon}\boldsymbol{\epsilon}')) \\ &= \mathrm{tr}(\mathbf{M})E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') \\ &= \sigma_\epsilon^2\mathrm{tr}(\mathbf{M}),\end{aligned}$$

using two properties of the matrix trace,

1. $\text{tr}(c) = c$.

2. $\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{BCA}) = \text{tr}(\mathbf{CAB})$.

Substituting for $\mathbf{M}$, we can use these properties to find $\text{tr}(\mathbf{M})$:

$$\begin{aligned}
\text{tr}(\mathbf{M}) &= \text{tr}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\
&= \text{tr}(\mathbf{I}) - \text{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\
&= \text{tr}(\mathbf{I}) - \text{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) \\
&= \text{tr}(\mathbf{I}_{N\times N}) - \text{tr}(\mathbf{I}_{K\times K}) \\
&= N - K.
\end{aligned}$$

For idempotent matrices, the rank of the matrix equals its trace, and so $\text{rank}(\mathbf{M}) = \text{tr}(\mathbf{M}) = N - K$.

Thus, we have the residual variance estimator

$$s_e^2 = \frac{\mathbf{e}'\mathbf{e}}{N-K},$$

and it follows immediately that $E(s_e^2) = \sigma_\epsilon^2$, that is, s_e^2 is an unbiased estimator of σ_ϵ^2. Finally, for idempotent matrices in quadratic forms—sums of squares—the rank of the matrix equals the degrees of freedom in the sum of squares. The number of degrees of freedom in the estimation of σ_ϵ^2 with $s_e^2 = E(\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon})/(N-K)$ is $N-K$.

The square root $s_e = \sqrt{s_e^2}$ is known as the *standard error of the regression* (SER) or the *standard error of the estimate* (SEE). In estimation, then,

$$\text{Var}(\mathbf{b}) = s_e^2(\mathbf{X}'\mathbf{X})^{-1}$$

gives the estimated variance of the coefficients $\mathbf{b}$. The standard errors of the coefficients are the square roots of the diagonal elements of $\text{Var}(\mathbf{b})$.

8 The decomposition of the sum of squares

As you know, the *centering matrix*,

$$\mathbf{C} = \mathbf{I} - (1/N)\mathbf{J} = \mathbf{I} - (1/N)\mathbf{1}\mathbf{1}',$$

mean deviates a vector $\mathbf{x}$. $\mathbf{C}$ is symmetric and idempotent.

If we partition $\mathbf{X} = [\mathbf{1} \;\; \mathbf{X}_x]$ and $\mathbf{b}' = [b_1 \;\; \mathbf{b}_x']$, then

$$\begin{aligned}
\mathbf{y} &= \mathbf{X}\mathbf{b} + \mathbf{e} \\
&= \mathbf{1}b_1 + \mathbf{X}_x\mathbf{b}_x + \mathbf{e}.
\end{aligned}$$

Premultiplying by the centering matrix $\mathbf{C}$,

$$\begin{aligned}\mathbf{Cy} &= \mathbf{C1}b_1 + \mathbf{CX}_x\mathbf{b}_x + \mathbf{Ce} \\ &= \mathbf{CX}_x\mathbf{b}_x + \mathbf{e},\end{aligned}$$

because

1. $\mathbf{C1} = (\mathbf{I} - (1/N)\mathbf{11}')\mathbf{1} = \mathbf{1} - (1/N)\mathbf{1}N = \mathbf{0}$, and
2. $\mathbf{Ce} = (\mathbf{I} - (1/N)\mathbf{11}')\mathbf{e} = \mathbf{e} - (1/N)\mathbf{1}(0) = \mathbf{e}$.

Thus, the decomposition of the sum of squares is

$$\begin{aligned}\mathbf{y}'\mathbf{Cy} &= \mathbf{b}'\mathbf{X}'\mathbf{CXb} + 2\mathbf{b}'\mathbf{X}'\mathbf{Ce} + \mathbf{e}'\mathbf{e} \\ &= \mathbf{b}'\mathbf{X}'\mathbf{CXb} + \mathbf{e}'\mathbf{e} \\ &= \mathbf{b}_x'\mathbf{X}_x'\mathbf{CX}_x\mathbf{b}_x + \mathbf{e}'\mathbf{e} \\ \text{TSS} &= \text{ESS} + \text{RSS}.\end{aligned}$$

8.1 The coefficient of determination: R^2

Then, as before, we define the multiple coefficient of determination as

$$\begin{aligned}R^2 &= \mathbf{b}'\mathbf{X}'\mathbf{CXb}/\mathbf{y}'\mathbf{Cy} &&= \text{ESS/TSS} \\ &= 1 - (\mathbf{e}'\mathbf{e}/\mathbf{y}'\mathbf{Cy}) &&= 1 - (\text{RSS/TSS}),\end{aligned}$$

and the adjusted R^2 as

$$\bar{R}^2 = 1 - \frac{\mathbf{e}'\mathbf{e}/(N-K)}{\mathbf{y}'\mathbf{Cy}/(N-1)} = 1 - \frac{N-1}{N-K}(1-R^2).$$

The substantive interpretation of R^2, however, has not changed. It gives the percentage of the variance in Y_i statistically explained by the model $\mathbf{Xb}$. Its square root, R, is the *multiple correlation coefficient*

$$R = r_{Y\hat{Y}}.$$