Math Camp for Economists:

OLS with Matrices

Justin C. Wiltshire

Department of Economics
University of Victoria
Summer 2023

OLS with Matrices

In this lesson:

- We assume you have already fully grasped the earlier reviews of univariate OLS, matrix algebra, and optimization
- To review OLS in matrix form, we also need to review a bit of matrix calculus
- We'll then review multivariate OLS using matrix algebra

Brief review of matrix calculus

Suppose we have a function $\mathbf{y} = T(\mathbf{x})$ where $\mathbf{y}$ is a $m \times 1$ vector, $\mathbf{x}$ is a $n \times 1$ vector. Denote the $m \times n$ matrix of first derivatives as

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}$$

$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$ if $\mathbf{y} = \mathbf{Ax}$

Let $\mathbf{y} = \mathbf{Ax}$, where $\mathbf{y}$ is a $m \times 1$ vector, $\mathbf{x}$ is a $n \times 1$ vector, and $\mathbf{A}$ is a $m \times n$ matrix that is independent of $\mathbf{x}$. Then $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}$

How do we know?

- We know the ith element of $\mathbf{y}$ is $\sum_{j=1}^{n} a_{ij} x_j$
- Then it follows that $\frac{\partial y_i}{\partial x_j} = a_{ij}$ for all i, j, which is the (i, j)th element of $\mathbf{A}$
- Thus $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}$

Exercise:

(1) Create a 3×3 matrix, $\mathbf{A}$, with at least two non-zero elements in each row and some elements > 1

(2) Use $\mathbf{A}$ and the column vector $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}'$ to find $\mathbf{y} = \mathbf{Ax}$

(3) Differentiate y_i wrt x_j for each $i, j = 1, 2, 3$ to populate the $(i, j)th$ element of a new 3×3 matrix, $\mathbf{B}$. How does $\mathbf{B}$ relate to $\mathbf{A}$?

$\frac{\partial b}{\partial \mathbf{x}}$ and $\frac{\partial b}{\partial \mathbf{y}}$ if $b = \mathbf{y}'\mathbf{A}\mathbf{x}$

Let $\mathbf{y}$ be a $m \times 1$ vector, $\mathbf{x}$ be a $n \times 1$ vector, and $\mathbf{A}$ be a $m \times n$ matrix that is independent of $\mathbf{x}$ and $\mathbf{y}$. Define $b = \mathbf{y}'\mathbf{A}\mathbf{x}$ is a scalar (how can you know?). Then $\frac{\partial b}{\partial \mathbf{x}} = \mathbf{y}'\mathbf{A}$ and $\frac{\partial b}{\partial \mathbf{y}} = \mathbf{x}'\mathbf{A}'$

How do we know?

- Define $\mathbf{w}' = \mathbf{y}'\mathbf{A} \Rightarrow b = \mathbf{w}'\mathbf{x}$
- Then it follows that $\frac{b}{\partial \mathbf{x}} = \mathbf{w}' = \mathbf{y}'\mathbf{A}$
- Since b is a scalar, we can write $b = \mathbf{y}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{A}'\mathbf{y} = b'$
- Then $\frac{\partial b}{\partial \mathbf{y}} = \frac{\partial b'}{\partial \mathbf{y}} = \mathbf{x}'\mathbf{A}'$

$\frac{\partial b}{\partial \mathbf{x}}$ if $b = \mathbf{x}'\mathbf{A}\mathbf{x}$

Let $\mathbf{x}$ be a $n \times 1$ vector, and $\mathbf{A}$ be a $n \times n$ matrix that is independent of $\mathbf{x}$. Define $b = \mathbf{x}'\mathbf{A}\mathbf{x}$ is a scalar. Then $\frac{\partial b}{\partial \mathbf{x}} = \mathbf{x}'(\mathbf{A} + \mathbf{A}')$ or $\frac{\partial b}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}')\mathbf{x}$. Pick the one that satisfies the needed dimensions, as b is a scalar

How do we know?

- We know that by definition $b = \sum_{j=1}^{n} \sum_{i=1}^{n} a_{ij} x_i x_j$
- Differentiate b wrt the kth element of $\mathbf{x}$: $\frac{\partial b}{\partial x_k} = \sum_{j=1}^{n} a_{kj} x_j + \sum_{i=1}^{n} a_{ik} x_i$ for all k
- Then $\frac{\partial b}{\partial \mathbf{x}} = \mathbf{x}'\mathbf{A} + \mathbf{x}'\mathbf{A}' = \mathbf{x}'(\mathbf{A} + \mathbf{A}')$ $\left(\text{or } \frac{\partial b}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}')\mathbf{x}\right)$

Exercise with $\frac{\partial b}{\partial \mathbf{x}}$ if $b = \mathbf{x}'\mathbf{A}\mathbf{x}$

Exercise: Let $\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix}$

(1) Find $b = \mathbf{x}'\mathbf{A}\mathbf{x}$

(2) Differentiate b wrt x_k, $k = 1, 2$ to populate the $(1, k)$th element of a 1×2 vector $\mathbf{c}$

(3) Find $\mathbf{A} + \mathbf{A}'$

(4) Noting $b = b'$, find the 1×2 vector $\mathbf{d} = \frac{\partial b}{\partial \mathbf{x}}$

(5) How does $\mathbf{c}$ relate to $\mathbf{d}$?

$\frac{\partial b}{\partial \mathbf{x}}$ if $b = \mathbf{x}'\mathbf{A}\mathbf{x}$ and $\mathbf{A}$ is symmetric

Let $\mathbf{x}$ be a $n \times 1$ vector, and $\mathbf{A}$ be a $n \times n$ symmetric matrix that is independent of $\mathbf{x}$. Define $b = \mathbf{x}'\mathbf{A}\mathbf{x}$ is a scalar. Then $\dfrac{\partial b}{\partial \mathbf{x}} = 2\mathbf{x}'\mathbf{A}$ (alternatively, $\dfrac{\partial b}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$)

How do we know? It follows directly from the more general previous result because for a symmetric matrix $\mathbf{A}$ we have $\mathbf{A}' = \mathbf{A} \Rightarrow (\mathbf{A} + \mathbf{A}') = (\mathbf{A} + \mathbf{A}) = 2\mathbf{A}$

$\frac{\partial \mathbf{A}^{-1}}{\partial c}$ if $\mathbf{A}$ is a nonsingular matrix with elements functions of a scalar parameter c

Let $\mathbf{A}$ be a $m \times m$ nonsingular matrix whose elements are functions of a scalar parameter c. Then

$$\frac{\partial \mathbf{A}}{\partial \mathbf{c}} = \begin{pmatrix} \frac{\partial a_{11}}{\partial c} & \cdots & \frac{\partial a_{1m}}{\partial c} \\ \vdots & \ddots & \vdots \\ \frac{\partial a_{m1}}{\partial c} & \cdots & \frac{\partial a_{mm}}{\partial c} \end{pmatrix}$$

and $\frac{\partial \mathbf{A}^{-1}}{\partial c} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial c}\mathbf{A}^{-1}$

How do we know?

- We know that by definition $\mathbf{A}^{-1}\mathbf{A} = I$
- Then $\frac{\partial(\mathbf{A}^{-1}\mathbf{A})}{\partial c} = \mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial c} + \frac{\partial \mathbf{A}^{-1}}{\partial c}\mathbf{A} = \mathbf{0}$
- Rearranging and recalling $\mathbf{A}^{-1}\mathbf{A} = I$ yields $\frac{\partial \mathbf{A}^{-1}}{\partial c} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial c}\mathbf{A}^{-1}$

Exercise with $\frac{\partial \mathbf{A}^{-1}}{\partial c}$ if $\mathbf{A}$ is a nonsingular matrix with elements functions of a scalar parameter c

Exercise: Let $\mathbf{A} = \begin{pmatrix} c & 2c \\ 2c & 2c \end{pmatrix}$

(1) Differentiate $\mathbf{A}$ wrt c

(2) Find $\mathbf{A}^{-1}$

(3) Find $\mathbf{A}^{-1}\mathbf{A}$

(4) Differentiate $\mathbf{A}^{-1}$ wrt c. Call this $\frac{\partial \mathbf{A}^{-1}}{\partial c} = \mathbf{W}$

(5) Find $-\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial c}\mathbf{A}^{-1} = \mathbf{Z}$

(6) How does $\mathbf{Z}$ relate to $\mathbf{W}$?

OLS: the true model

Assume a specific reduced form model of the DGP for y holds in the population such that for each individual $i = 1, ..., n$

$$y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_{k-1} X_{k-1,i} + \epsilon_i$$

- y_i is i's observed value of the outcome y
- $X_{j,i}$ is i's observed value of variable X_j which is independent of X_m for all $j \neq m$, $j, m = 1, ..., k-1$
- ϵ_i is the error (or "disturbance") for i
- β_j, $j = 0, ..., k-1$ are (unknown) population parameters

The true model in matrix form

We can write this in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{1,1} & X_{2,1} & \dots & X_{k-1,1} \\ 1 & X_{1,2} & X_{2,2} & \dots & X_{k-1,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1,n} & X_{2,n} & \dots & X_{k-1,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

or, more compactly, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with rows $y_i = \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i$, $i = 1, ..., n$

- This is linear in the parameters
- $\mathbf{X}$ is a $n \times k$ matrix of n observations of $k-1$ (assumed) independent "regressor" variables (left-augmented with a column of ones for the constant β_0)
  - → The independence assumption means $\mathbf{X}$ has full column rank
  - → This will ensure that the inverse of $\mathbf{X}'\mathbf{X}$ exists

The sum of squared residuals given some estimator for β, $\hat{\beta}$

Let $\hat{\beta}$ be some estimator for β

- This implies the fitted values are $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$
- We denote the residuals (these are the prediction errors. You may prefer to denote them $\hat{\epsilon}$):

$$\begin{aligned} \mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}\hat{\beta} \end{aligned}$$

- Then the sum of squared residuals (SSR) is

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

  where this last equality arises because $\mathbf{y}'\mathbf{X}\hat{\beta} = \hat{\beta}'\mathbf{X}'\mathbf{y}$

Exercise:

(1) Why is this last statement true? Recall our exercises with the laws of matrix algebra

(2) Show this using a 3×1 vector $\mathbf{y}$, a 3×2 matrix $\mathbf{X}$ with full column rank, and a 2×1 vector $\hat{\beta}$

Estimating β with OLS

We estimate $\hat{\boldsymbol{\beta}}_{OLS}$ by choosing $\hat{\boldsymbol{\beta}}$ to minimize $\mathbf{e}'\mathbf{e}$. That is:

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\hat{\boldsymbol{\beta}}} \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

Exercise: Use the matrix differentiation we just reviewed

(1) What are the FOCs for this minimization problem? Ensure the matrix exists as you've written it

(2) Assuming X has full column rank, solve for $\hat{\boldsymbol{\beta}}_{OLS}$

(3) What is the necessary SOC for $\hat{\boldsymbol{\beta}}_{OLS}$ to be the global minimizer of $\mathbf{e}'\mathbf{e}$?

(4) Let $\mathbf{y} = \begin{bmatrix} 2 \\ 3 \\ 1 \\ 2 \end{bmatrix}$ and $\mathbf{X} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$

a) How many (assumed) independent variables are there? How many parameters to estimate?
b) Is $\mathbf{X}'\mathbf{X}$ invertible? How do you know?
c) Find $\hat{\boldsymbol{\beta}}_{OLS}$
d) Does $\hat{\boldsymbol{\beta}}_{OLS}$ minimize $\mathbf{e}'\mathbf{e}$? How do you know?

Properties of $\hat{\beta}_{OLS}$

We just found that the FOCs were

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \hat{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\beta}$$
$$= \mathbf{0}$$
$$\Rightarrow \mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y}$$

By the definition of the residuals, we have $\mathbf{y} = \mathbf{X}\hat{\beta} + \mathbf{e}$. Plug this into the equality above:

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'(\mathbf{X}\hat{\beta} + \mathbf{e})$$
$$= \mathbf{X}'\mathbf{X}\hat{\beta} + \mathbf{X}'\mathbf{e}$$
$$\Rightarrow \mathbf{X}'\mathbf{e} = \mathbf{0}$$

That is, $\mathbf{X}$ is "orthogonal" to $\mathbf{e}$. This result can also be written as $\sum_{i=1}^{n} x_{j,i}e_i = 0$ for all $j = 0, ..., k-1$ and says the sample covariance of each of the $k-1$ regressors with the residuals is zero

→ Recall that $\mathbf{x}_0$ is a vector of ones which cannot vary with $\mathbf{e}$

Important properties of the OLS estimators given $\mathbf{X}'\mathbf{e} = \mathbf{0}$

Provided we include a constant in the model, we learn several important properties about the OLS estimators from $\mathbf{X}'\mathbf{e} = \mathbf{0}$

- The observed values of $\mathbf{X}$ are uncorrelated with the residuals
- The residuals always sum to zero: $\sum_{i=1}^{n} e_i = 0$
- The mean of the residuals is zero: $\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$
- The regression hyperplane passes through the means of the observed values. That is, $(\bar{x}, \bar{y})$ is always on the OLS regression line: $\bar{y} = \mathbf{x}\hat{\beta}$
- The predicted values $\hat{\mathbf{y}}$ are uncorrelated with the residuals
  - $\rightarrow$ Given $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ we have $\hat{\mathbf{y}}'\mathbf{e} = (\mathbf{X}\hat{\beta})'\mathbf{e} = \hat{\beta}'\mathbf{X}'\mathbf{e} = \hat{\beta}'\mathbf{0} = \mathbf{0}$
- $\overline{\hat{\mathbf{y}}} = \overline{\mathbf{y}}$

Note that these properties hold by construction and involve the residuals. They say **nothing** about the unobserved errors, ϵ

Exercises with the residuals

Exercise: Using $\mathbf{y}$, $\mathbf{X}$, and your $\hat{\beta}_{OLS}$ from the previous exercise

(1) Find $\mathbf{e}$

(2) Find the sum of the residuals, $\sum_{i=1}^{n} e_i$

(3) Find the mean of the residuals, $\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$

(4) Find $\mathbf{X}'\mathbf{e}$

(5) Find $\hat{\mathbf{y}}'\mathbf{e}$

Gauss-Markov Theorem

Can we say anything yet about whether $\hat{\beta}_{OLS}$ is a "good" estimator for β? No! Again, we need some assumptions on the true population model. Actually, we already put some on it, but we need more!

Under a set of assumptions/conditions on the true population model, no *other* linear and unbiased estimator will have a smaller sampling variance than the OLS estimator ($\hat{\beta}_{OLS}$)

- This is the Gauss-Markov theorem
- Obviously (pretty much a restatement of the definition), it suggests that $\hat{\beta}_{OLS}$ is linear, unbiased (for β), and has the smallest sampling variance of this class of estimators
- You'll recall this being boiled down to "OLS is BLUE (**B**est **L**inear **U**nbiased **E**stimator)"
  - → "Best" means it has the smallest sampling variance among this class of estimators
  - → "Linear" means it is linear in the parameters
  - → "Unbiased" means (narrowing what we said earlier) that $Bias(\hat{\beta}_{OLS}, \beta) = E\left[\hat{\beta}_{OLS}\right] - \beta = 0$

Gauss-Markov assumptions

Several assumptions about the true population model are needed for OLS to be BLUE:

(1) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$

$\rightarrow$ There's a linear relationship between $\mathbf{X}$ and $\mathbf{y}$ (already assumed)

(2) $\mathbf{X}$ is a $n \times k$ matrix with full column rank: $rank(\mathbf{X}) = k$ (already assumed)

$\rightarrow$ No perfect multicollinearity

(3) Zero conditional mean of the errors (Note: $E[\epsilon] = 0$ is trivial with a constant term included in $\mathbf{y}$)

$$E[\boldsymbol{\epsilon}|\mathbf{X}] = E\begin{bmatrix} \epsilon_1|\mathbf{X} \\ \vdots \\ \epsilon_n|\mathbf{X} \end{bmatrix} = \begin{bmatrix} E[\epsilon_1] \\ \vdots \\ E[\epsilon_n] \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}$$

$\rightarrow$ $\mathbf{X}$ tells us nothing about the expected value of the errors

$\Rightarrow$ $E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$

Gauss-Markov assumptions cont'd.

(4) The errors are homoskedastic and uncorrelated

$$E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}] = E\begin{bmatrix} \epsilon_1|\mathbf{X} \\ \vdots \\ \epsilon_n|\mathbf{X} \end{bmatrix}\begin{bmatrix} \epsilon_1|\mathbf{X} & \dots & \epsilon_n|\mathbf{X} \end{bmatrix}$$

$$= E\begin{bmatrix} \epsilon_1^2|\mathbf{X} & \dots & \epsilon_1\epsilon_n|\mathbf{X} \\ \vdots & \ddots & \vdots \\ \epsilon_n\epsilon_1|\mathbf{X} & \dots & \epsilon_n^2|\mathbf{X} \end{bmatrix}$$

$$= \begin{bmatrix} E[\epsilon_1^2|\mathbf{X}] & \dots & E[\epsilon_1\epsilon_n]|\mathbf{X} \\ \vdots & \ddots & \vdots \\ E[\epsilon_n\epsilon_1|\mathbf{X}] & \dots & E[\epsilon_n^2]|\mathbf{X} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix} = \sigma^2 I$$

$$\Rightarrow \Omega = \sigma^2 I$$

Implications of the Gauss-Markov assumptions

Given these assumptions, and given that $\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ together yield $\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}$

$\rightarrow$ Note: Let $\mathbf{B} = \mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Then for

(1) $\hat{\boldsymbol{\beta}}_{OLS}$ is unbiased for $\boldsymbol{\beta}$:

$$\begin{aligned}
E[\hat{\boldsymbol{\beta}}_{OLS}|\mathbf{X}] &= E[\boldsymbol{\beta}|\mathbf{X}] + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}|\mathbf{X}] \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\boldsymbol{\epsilon}|\mathbf{X}] \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{0} \\
&= \boldsymbol{\beta}
\end{aligned}$$

Implications of the Gauss-Markov assumptions cont'd.

(2) $Var(\hat{\boldsymbol{\beta}}_{OLS}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$:

$$\begin{aligned}
Var(\hat{\boldsymbol{\beta}}_{OLS}|\mathbf{X}) &= E[(\hat{\boldsymbol{\beta}}_{OLS} - E[\hat{\boldsymbol{\beta}}_{OLS}])(\hat{\boldsymbol{\beta}}_{OLS} - E[\hat{\boldsymbol{\beta}}_{OLS}])'|\mathbf{X}] \\
&= E[(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} \\
&= (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'(\sigma^2 I)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}
\end{aligned}$$

Implications of the Gauss-Markov assumptions cont'd.

(3) $\hat{\boldsymbol{\beta}}_{OLS}$ has the smallest sampling variance among all linear unbiased estimators. That is, for any $k \times 1$ vector $\mathbf{c} \neq \mathbf{0}$, we have $Var(\mathbf{c}'\hat{\boldsymbol{\beta}}_{OLS}) \leq Var(\mathbf{c}'\tilde{\boldsymbol{\beta}})$

- Note that all linear estimators take the form $\tilde{\boldsymbol{\beta}} = \mathbf{Ay}$
- Define $\mathbf{B} = \mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ such that

$$\begin{aligned}\tilde{\boldsymbol{\beta}} = \mathbf{Ay} &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{B}]\mathbf{y} \\ &= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{B}](\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} + \mathbf{BX}\boldsymbol{\beta} + [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{B}]\boldsymbol{\epsilon}\end{aligned}$$

Thus

$$E[\mathbf{Ay}|\mathbf{X}] = \boldsymbol{\beta} + \mathbf{BX}\boldsymbol{\beta}$$

As we are considering only unbiased linear estimators, we must choose $\mathbf{A}$ such that $\mathbf{BX} = \mathbf{0}$

Implications of the Gauss-Markov assumptions cont'd

Then

$$\begin{aligned}
Var(\tilde{\boldsymbol{\beta}}|\mathbf{X}) &= E[(\tilde{\boldsymbol{\beta}} - E[\tilde{\boldsymbol{\beta}}])(\tilde{\boldsymbol{\beta}} - E[\tilde{\boldsymbol{\beta}}])'|\mathbf{X}] \\
&= E[[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{B}]\boldsymbol{\epsilon}\boldsymbol{\epsilon}'[\mathbf{B}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]|\mathbf{X}] \\
&= [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X} + \mathbf{B}]'E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}][\mathbf{B}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\
&= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{B}\mathbf{B}'] \\
&= Var(\hat{\boldsymbol{\beta}}_{OLS}) + \sigma^2\mathbf{B}\mathbf{B}'
\end{aligned}$$

Thus for any $k \times 1$ vector $\mathbf{c} \neq \mathbf{0}$

$$Var(\mathbf{c}'\tilde{\boldsymbol{\beta}}) = Var(\mathbf{c}'\hat{\boldsymbol{\beta}}_{OLS}) + \sigma^2\mathbf{c}'\mathbf{B}\mathbf{B}'\mathbf{c} = Var(\mathbf{c}'\hat{\boldsymbol{\beta}}_{OLS}) + \sigma^2(\mathbf{B}'\mathbf{c})'\mathbf{B}\mathbf{c} \geq Var(\mathbf{c}'\hat{\boldsymbol{\beta}}_{OLS})$$

Thus under the Gauss-Markov assumptions $\hat{\boldsymbol{\beta}}_{OLS}$ is the best linear unbiased estimator

References

This drew notes from:

- Dennis L. Hartmannn's notes on Matrix Differentiation
- Michael J. Rosenfeld's notes on OLS in Matrix Form (Caution: there are some typos in these notes)
- Anthony Tay's notes on OLS using Matrix Algebra