Lecture 4: Multiple Linear Regression II

Gauss Markov Model and Statistical Properties of MLR

Ailin Zhang

2023-05-17

Recap: LS Estimates - Method 1

- To find $(\hat{\beta}_0, \hat{\beta}_1)$ that minimize:

$$L(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- One can set the derivatives of L with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ to 0

$$\frac{\partial L}{\partial \hat{\beta}_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial L}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

Recap: LS Estimates - Method 1

- This results in the 2 equations with 2 unknowns:

$$n\hat{\beta}_0 + \hat{\beta}_1 \underbrace{\sum_{i=1}^{n} x_i}_{n\bar{x}} = \underbrace{\sum_{i=1}^{n} y_i}_{n\bar{y}} \xrightarrow{\text{divide by n}} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 \underbrace{\sum_{i=1}^{n} x_i}_{n\bar{x}} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \xrightarrow{\text{replace } \hat{\beta}_0} (\bar{y} - \hat{\beta}_1 \bar{x}) n\bar{x} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\iff \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) = \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

$$\iff \hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Recap: LS Estimates - Method 1

- Normal equations: a system of equations whose solution is the Ordinary Least Squares (OLS) estimator of the regression coefficients

$$
\begin{aligned}
\hat{\beta}_0 \cdot n + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} \quad & + \cdots \quad + \hat{\beta}_p \sum_{i=1}^{n} x_{ip} = \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} x_{i1} \quad & + \cdots \quad + \hat{\beta}_p \sum_{i=1}^{n} x_{i1} x_{ip} = \sum_{i=1}^{n} x_{i1} y_i \\
& \vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} \quad & + \cdots \quad + \hat{\beta}_p \sum_{i=1}^{n} x_{ik} x_{ip} = \sum_{i=1}^{n} x_{ik} y_i \\
& \vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ip} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ip} x_{i1} \quad & + \cdots \quad + \hat{\beta}_p \sum_{i=1}^{n} x_{ip} x_{ip} = \sum_{i=1}^{n} x_{ip} y_i
\end{aligned}
$$

Recap: OLS Estimates - Method 1

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ip} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}x_{i1} & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{ik}x_{i1} & \sum_{i=1}^n x_{ik}x_{i2} & \cdots & \sum_{i=1}^n x_{ik}x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ip} & \sum_{i=1}^n x_{ip}x_{i1} & \sum_{i=1}^n x_{ip}x_{i2} & \cdots & \sum_{i=1}^n x_{ip}x_{ip} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \\ \vdots \\ \sum_{i=1}^n x_{ip}y_i \end{bmatrix}$$

- In matrix notation, the normal equation is $(X^TX)\hat{\beta} = X^TY$
- And the least squares estimate is $\hat{\beta} = (X^TX)^{-1}X^TY$

Recap: OLS Estimates - Method 2

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n1} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = (\mathbf{1}_n, X_1, X_2, \dots, X_p)$$

$RSS(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = (Y - X\beta)^T (Y - X\beta)$, $\nabla_\beta RSS(\beta) = 0$

$$\begin{aligned} \nabla_\beta RSS(\beta) &= \nabla_\beta \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 = \sum_{i=1}^{n} \nabla_\beta (y_i - x_i^T \beta)^2 \\ &= \sum_{i=1}^{n} 2 \cdot (y_i - x_i^T \beta) \underbrace{\nabla_\beta (y_i - x_i^T \beta)}_{= -x_i \text{ dims:}(p+1)\times 1} = -2 \sum_{i=1}^{n} (y_i - x_i^T \beta) x_i = 0 \end{aligned}$$

We can also acquire normal equations: $\sum_{i=1}^{n} (y_i - x_i^T \hat{\beta}) x_i = 0$

Recap: OLS Estimates - Method 2

Using the matrix notation, we have

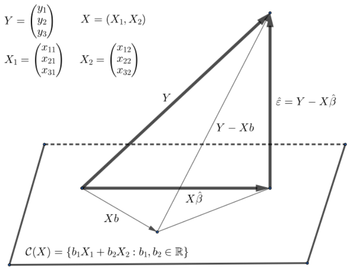
$$LHS = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})x_i = \begin{bmatrix} x_1, x_2, \dots, x_n \end{bmatrix} \begin{bmatrix} y_1 - x_1^T\hat{\beta} \\ y_2 - x_2^T\hat{\beta} \\ \vdots \\ y_n - x_n^T\hat{\beta} \end{bmatrix}$$

$$= X^T(Y - X\hat{\beta}) = 0$$

$$\Rightarrow X^TY - X^TX\hat{\beta} = 0$$

$$(X^TX)\hat{\beta} = X^TY$$

Recap: OLS Estimates - Method 3



- The OLS problem is to find the best linear combination of the column vectors of X to approximate the response vector Y
- By projection, the residual vector $\hat{\epsilon} = Y - X\hat{\beta}$ must be orthogonal to $C(X)$, or, equivalently, the residual vector is orthogonal to $X_1, \ldots, X_p$
- This geometric intuition in turn implies that $X_1^T(Y - X\hat{\beta}) = 0, X_2^T(Y - X\hat{\beta}) = 0, \ldots, X_p^T(Y - X\hat{\beta}) = 0$

  which is essentially the normal equation

Agenda

- Gauss-Markov Model (model assumptions)
- Properties of the OLS Estimator ($\hat{\beta}$)
    - Mean
    - Covariance
- Properties of Variance ($\hat{\sigma}^2$)
- Gauss-Markov Theorem

Gauss-Markov Model

- Without any stochastic assumptions, the OLS in the last lecture is purely algebraic.
- From now on, we want to discuss the statistical properties of $\hat{\beta}$ and associated quantities.
- A simple starting point is the following Gauss–Markov model with a fixed design matrix X and unknown parameters β, ϵ:

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

  - $X^T X$ is non-degenerate, $E(\epsilon) = 0$, $\text{cov}(\epsilon) = \sigma^2 I_n$
  - Y has mean $X\beta$ and covariance matrix $\sigma^2 I_n$
  - At the individual level, $y_i = x_i^T \beta + \epsilon_i$, where the error terms are uncorrelated with mean 0 and variance σ^2.

Properties of OLS estimator ($\hat{\beta}$)

$\hat{\beta} = (X^T X)^{-1} X^T Y$

Theorem

Under the Gauss-Markov Model,

$$E(\hat{\beta}) = \beta$$

$$cov(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

Proof. Because $E(Y) = X\beta$, we have

$E(\hat{\beta}) = E\{(X^T X)^{-1} X^T Y\} = (X^T X)^{-1} X^T E(Y) = (X^T X)^{-1} X^T X\beta = \beta$

Because $\text{cov}(Y) = \sigma^2 I_n$, we have

$$\begin{aligned} \text{cov}(\hat{\beta}) &= \text{cov}\{(X^T X)^{-1} X^T Y\} = (X^T X)^{-1} X^T \text{cov}(Y) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} \end{aligned}$$

Mean and Covariance matrix of $\hat{Y}$ and $\hat{\epsilon}$

- Since $Y = \hat{Y} + \hat{\epsilon}$, where $\hat{Y} = X\hat{\beta} = HY$ and the residual vector is $\hat{\epsilon} = Y - \hat{Y} = (I_n - H)Y$, $H = X(X^T X)^{-1}X^T$
- Both H and $I_n - H$ are projection matrices. You can prove that $H(I_n - H) = (I_n - H)H = 0$

Theorem

Under the Gauss-Markov model,

$$E\begin{pmatrix}\hat{Y}\\ \hat{\epsilon}\end{pmatrix} = \begin{pmatrix}X\beta\\ 0\end{pmatrix}, \qquad cov\begin{pmatrix}\hat{Y}\\ \hat{\epsilon}\end{pmatrix} = \sigma^2\begin{pmatrix}H & 0_{n\times n}\\ 0_{n\times n} & I_n - H\end{pmatrix}$$

- So $\hat{Y}$ and $\hat{\epsilon}$ are uncorrelated.

Mean and Covariance matrix of $\hat{Y}$ and $\hat{\epsilon}$

Proof.

$$\begin{pmatrix} \hat{Y} \\ \hat{\epsilon} \end{pmatrix} = \begin{pmatrix} HY \\ (I_n - H)Y \end{pmatrix} = \begin{pmatrix} H \\ I_n - H \end{pmatrix} Y$$

$$E \begin{pmatrix} \hat{Y} \\ \hat{\epsilon} \end{pmatrix} = \begin{pmatrix} H \\ I_n - H \end{pmatrix} E(Y) = \begin{pmatrix} H \\ I_n - H \end{pmatrix} X\beta = \begin{pmatrix} X\beta \\ 0 \end{pmatrix}$$

$$\mathrm{cov} \begin{pmatrix} \hat{Y} \\ \hat{\epsilon} \end{pmatrix} = \begin{pmatrix} H \\ I_n - H \end{pmatrix} \mathrm{cov}(Y) \begin{pmatrix} H^T & (I_n - H)^T \end{pmatrix}$$
$$= \sigma^2 \begin{pmatrix} H \\ I_n - H \end{pmatrix} \begin{pmatrix} H & I_n - H \end{pmatrix} = \sigma^2 \begin{pmatrix} H & 0 \\ 0 & I_n - H \end{pmatrix}$$

Remark. $\mathrm{cov}(\hat{\epsilon}_i, \hat{\epsilon}_j) = -H_{ij}$ for $i \neq j$

Variance Estimation (σ^2)

- σ^2 is the variance of each ϵ_i, but the ϵ_i's are not observable
- We estimate ϵ by the residuals $\hat{\epsilon}_i = y_i - x_i^T \beta$
- $RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2$
  - We have shown that $\hat{\epsilon}_i$ has mean zero and variance $\sigma^2(1 - H_{ii})$.
  - So the mean of RSS (Residual Sum of Squares):

$$\sum_{i=1}^{n} \sigma^2(1 - H_{ii}) = \sigma^2[n - \text{trace}(H)] = \sigma^2[n - (p + 1)]$$

Theorem

Define

$$\hat{\sigma}^2 = \frac{RSS}{n - (p + 1)} = \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{n - (p + 1)}$$

Then $E(\hat{\sigma}^2) = \sigma^2$ is an unbiased estimator under the Gauss Markov model.

Substitute σ^2 by $\hat{\sigma}^2$

- Now we can modify $\text{cov}(\hat{\beta})$

  We denote $\widehat{\text{cov}}(\hat{\beta}) = \hat{\sigma}^2(X^T X)^{-1}$

- Standard error (s.e.) of $\hat{\beta}_j$ for $j = 0, 1, \ldots, p$:

$$\sqrt{\text{j th diagonal entry of } \hat{\sigma}^2(X^T X)^{-1}}$$

Gauss–Markov Theorem (BLUE)

- Why should we focus on it and are there any better estimators?

  Under the Gauss–Markov model, the answer is definite: we focus on the OLS estimator because it is optimal in the sense of best linear unbiased estimator (BLUE).

  - It has the smallest covariance matrix among all linear unbiased estimators.

Theorem

Under the Gauss–Markov model, the OLS estimator $\hat{\beta}$ for β is the best linear unbiased estimator (BLUE) in the sense that $cov(\hat{\beta}) \preceq cov(\tilde{\beta})$ for any estimator $\tilde{\beta}$ satisfying

1. $\tilde{\beta} = AY$ for some $A \in \mathbb{R}^{(p+1)\times n}$ not depending on Y;

2. $\tilde{\beta}$ is unbiased for β.

Gauss–Markov Theorem

Proof. Let OLS estimator $\hat{\beta} = \hat{A}Y$, $\tilde{\beta} = AY$ $E(\tilde{\beta}) = E(AY) = AX\beta = \beta$,
For unbiased estimator $AX = I_{p+1}$ must hold

$$\begin{aligned}
\text{cov}(\tilde{\beta}) &= \text{cov}(AY) \\
&= \text{cov}(AY - \hat{A}Y + \hat{A}Y) \\
&= \text{cov}[(A - \hat{A})Y] + \text{cov}(\hat{A}Y) \\
&+ \text{cov}\{(A - \hat{A})Y, \hat{A}Y\} + \text{cov}\{\hat{A}Y, (A - \hat{A})Y\}
\end{aligned}$$

$$\begin{aligned}
\text{cov}\{\hat{A}Y, (A - \hat{A})Y\} &= \hat{A}\text{cov}(Y)(A - \hat{A})^T \\
&= \sigma^2 \hat{A}(A - \hat{A})^T \\
&= \sigma^2(\hat{A}A^T - \hat{A}\hat{A}^T) \\
&= \sigma^2[(X^TX)^{-1}X^TA^T - (X^TX)^{-1}X^TX(X^TX)^{-1}] = 0
\end{aligned}$$

Similarly, you can prove $\text{cov}\{(A - \hat{A})Y, \hat{A}Y\} = 0$

Gauss–Markov Theorem

Proof. (Continued)

The above covariance decomposition simplifies to

$$\text{cov}(\tilde{\beta}) = \text{cov}[(A - \hat{A})Y] + \text{cov}(\hat{\beta})$$

Therefore, we have

$$\text{cov}(\tilde{\beta}) - \text{cov}(\hat{\beta}) = \text{cov}[(A - \hat{A})Y] \succeq 0$$

Which implies that $\text{cov}(\hat{\beta}) \preceq \text{cov}(\tilde{\beta})$

Discussion

How to find other linear unbiased estimators?

Review: Multivariate Statistics

- Expectation of a random matrix

  Consider a random matrix $X \in \mathbb{R}^{m \times n}$

  We define
  $$\mathrm{E}(X) = (\mathrm{E}(x_{ij})),$$
  i.e., taking expectations element-wise.

  - Let A be a constant matrix of appropriate dimension, then $\mathrm{E}(AX) = A\mathrm{E}(X)$.
  - Let B be another constant matrix of appropriate dimension, then $\mathrm{E}(XB) = \mathrm{E}(X)B$.

Review: Multivariate Statistics

- Variance of a random vector

  Let $X \in \mathbb{R}^n$ be a random vector.

  We define
  $$\mathrm{Var}(X) = \mathrm{E}[(X - \mu_X)(X - \mu_X)^\top].$$
  Then the (i, j)-th element of $\mathrm{Var}(X)$ is $\mathrm{cov}(x_i, x_j)$. The diagonal elements are $\mathrm{Var}(x_i)$.

  - Let A be a constant matrix of appropriate dimension, then $\mathrm{Var}(AX) = A\mathrm{Var}(X)A^\top$.

    Proof.
    $$\begin{aligned}
    \mathrm{Var}(AX) &= \mathrm{E}[(AX - \mathrm{E}(AX))(AX - \mathrm{E}(AX))^\top] \\
    &= \mathrm{E}[(AX - A\mu_X)(AX - A\mu_X)^\top] \\
    &= \mathrm{E}[A(X - \mu_X)(X - \mu_X)^\top A^\top] \\
    &= A\mathrm{E}[(X - \mu_X)(X - \mu_X)^\top]A^\top = A\mathrm{Var}(X)A^\top
    \end{aligned}$$

Review: Multivariate Statistics

- Covariance between two random vectors

We can also define

$$\mathrm{cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)^\top],$$

then

$$\begin{aligned}
\mathrm{cov}(AX, BY) &= \mathrm{E}[(AX - A\mu_X)(BY - B\mu_Y)^\top] \\
&= \mathrm{E}[A(X - \mu_X)(Y - \mu_Y)^\top B^\top] \\
&= A\mathrm{E}[(X - \mu_X)(Y - \mu_Y)^\top]B^\top \\
&= A\mathrm{cov}(X, Y)B^\top
\end{aligned}$$