JOHNS HOPKINS
BLOOMBERG SCHOOL of PUBLIC HEALTH

HW1 grading
Quiz 1 → solution
PS2 is posted
Datasets → NMES

Lecture 7

Vector representation of MLR continued, assessing the impact of Gaussian residuals assumption

MLR model expressed in vector notation

We have for each subject i, $i = 1, \dots, n$

$$\underbrace{Y_i}_{} = \underbrace{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi}}_{\mu_i} + \varepsilon_i$$

error residual

$$\begin{array}{c|c}
i & \\
1 & Y_1 = 1\,\beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \dots \beta_p X_{p1} + \varepsilon_1 \\
2 & Y_2 = 1\,\beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \dots \beta_p X_{p2} + \varepsilon_2 \\
\vdots & \vdots \\
n & Y_n = 1 \cdot \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \dots \beta_p X_{pn} + \varepsilon_p
\end{array}$$

$$\underset{\sim}{Y}_{n\times 1} = \underbrace{\underset{\sim}{1}\,\beta_0 + \beta_1 \underset{\sim}{X_1}{}_{n\times 1} + \beta_2 \underset{\sim}{X_2}{}_{n\times 1} + \dots \beta_p \underset{\sim}{X_n}{}_{n\times 1}} + \underset{\sim}{\varepsilon}_{n\times 1}$$

$$\underset{\sim}{Y}_{n\times 1} = (\underset{\sim}{1}, \underset{\sim}{X_1}, \underset{\sim}{X_2}, \dots, \underset{\sim}{X_p}) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{p+1\times 1} + \underset{\sim}{\varepsilon}_{n\times 1}$$

$X_{n\times(p+1)}$ design matrix

MLR model expressed in vector notation

$$\Rightarrow \underset{\sim}{Y}_{n\times 1} = X_{n\times(p+1)} \underset{\sim}{\beta}_{(p+1)\times 1} + \underset{\sim}{\varepsilon}_{n\times 1}$$

Say $n=3,\ p=2 \Rightarrow Y_i = \underbrace{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}} + \varepsilon_i$

$$\underset{\sim}{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ 1 & X_{13} & X_{23} \end{bmatrix} \quad \underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad \underset{\sim}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix}$$

$$X_{3\times 3}\, \underset{\sim}{\beta}_{3\times 1} = \begin{pmatrix} 1\cdot\beta_0 + \beta_1 X_{11} + \beta_2 X_{21} \\ 1\cdot\beta_0 + \beta_1 X_{12} + \beta_2 X_{22} \\ 1\cdot\beta_0 + \beta_1 X_{13} + \beta_2 X_{23} \end{pmatrix}_{3\times 1} \begin{matrix} -\ \mu_1 \\ -\ \mu_2 \\ -\ \mu_3 \end{matrix} = \underset{\sim}{\mu}_{3\times 1}$$

Defines the mean of Y_i for each i

MLR model expressed in vector notation

What about distribution of $\underset{\sim}{\varepsilon}$? and $\underset{\sim}{Y}$?

In general, we can define the multivariate normal distribution as: $\underset{\sim}{Y}_{n\times 1} \sim MVN(\underset{\sim}{\mu}_{n\times 1}, V_{n\times n})$

where $\underset{\sim}{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$ $\underset{\sim}{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$

$V = \begin{bmatrix} v_{11} & & v_{ij} \\ & v_{22} & \\ v_{ji} & & \ddots \; v_{nn} \end{bmatrix}$

$\boxed{v_{ii} = Var(Y_i)}$

$v_{ij} = v_{ji} = Cov(Y_i, Y_j)$

If $\varepsilon_i \sim N(0, \sigma^2)$, independent

$\underset{\sim}{\varepsilon} \sim MVN(\underset{\sim}{0}, \sigma^2 I_{n\times n})$

$I = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ 0 & & \ddots \; 1 \end{bmatrix}$ identity matrix

$\underset{\sim}{Y} \sim MVN(\underbrace{X\underset{\sim}{\beta}}_{\underset{\sim}{\mu}}, \sigma^2 I)$

$V = \begin{bmatrix} \sigma^2 & & 0 \\ 0 & \sigma^2 & \\ & & \ddots \; \sigma^2 \end{bmatrix}$

MLE or LS solution expressed in vector notation

MLR model: $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$, $\underset{\sim}{\varepsilon} \sim MVN(\underset{\sim}{0}, \sigma^2 I)$

MLE or least squares: Going to drop the "$\sim$"

Choose $\hat{\beta}$ and $\hat{\sigma}^2$ to minimize $\sum_{i=1}^{n} (y_i - X_i\beta)^2$

X_i = ith row of X

$$\sum_{i=1}^{n} (y_i - X_i\beta)^2 = \underset{1\times n}{(Y - X\beta)'}\,\underset{n\times 1}{(Y - X\beta)}$$

$(Y_{n\times1} - X\beta_{n\times1})_{n\times1}$

$(Y_1 - \mu_1,\ Y_2 - \mu_2)\begin{pmatrix} Y_1 - \mu_1 \\ Y_2 - \mu_2 \end{pmatrix}$

$(Y_1 - \mu_1)^2 + (Y_2 - \mu_2)^2$

$$U_\beta(\beta) = \frac{d}{d\beta}(Y - X\beta)'(Y - X\beta)$$

$$= X'(Y - X\beta)$$

Set to 0, solve for $\hat{\beta} = (X'X)^{-1}X'Y$

SLR

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum (X_i - \bar{X})^2\right.}$$

Predicted values and residuals in vector notation

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

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij}\, y_j$$

$$\hat{Y} = X\hat{\beta} = \underbrace{X(X'X)^{-1}X'}_{n\times(p+1)}\,Y = \underset{n\times n}{H}\,\underset{n\times 1}{Y}$$

Hat matrix

hat

$$\underline{\hat{R}} = Y - \hat{Y} = Y - HY = (I-H)Y$$

Distribution of $\hat{\beta}$

Note that if $\underset{\sim}{Y} \sim MVN(\underset{\sim}{\mu}, V)$, then

$$A\underset{\sim}{Y} \sim MVN(A\underset{\sim}{\mu}, \underline{AVA'})$$

$$A' = [(X'X)^{-1}X']' = X(X'X)^{-1}$$

$$\hat{\beta} = \underbrace{(X'X)^{-1}X'}_{A}Y$$

$$E(\hat{\beta}) = E(AY) = E\left((X'X)^{-1}X'Y\right) = (X'X)^{-1}X'E(Y) = (X'X)^{-1}\underbrace{X'X}\beta = \beta$$

$$Var(\hat{\beta}) = Var(AY) = (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1}$$

$$\sigma^2(X'X)^{-1}\underbrace{X'X}_{SLR}\underbrace{(X'X)^{-1}}_{\hat{\beta}_1 = \sigma^2/SSX} = \sigma^2(X'X)^{-1}$$

Distribution of $\hat{Y}$

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

$$\hat{Y} \sim MVN(X\beta, \sigma^2 H)$$

$$E(\hat{Y}) = H\,E(Y) = HX\beta = X(X'X)^{-1}X'X\beta = X\beta$$

$$Var(\hat{Y}) = H\,Var(Y)\,H' = X(X'X)^{-1}X'\sigma^2 I\,X(X'X)^{-1}X'$$

$$= \sigma^2 X(X'X)^{-1}\underbrace{X'X(X'X)^{-1}}X'$$

$$= \sigma^2 X(X'X)^{-1}X' = \sigma^2 H$$

Properties of the Hat matrix

1) hat matrix is symmetric

$$H' = [X(X'X)^{-1}X']' = X(X'X)^{-1}X'$$

2) hat matrix is idempotent: $H \cdot H = H$

Distribution of $\hat{R}$

$$\hat{R} = Y - \hat{Y} = Y - HY = \underbrace{(I-H)}_{A} Y$$

$$\hat{R} \sim MVN($$

$$E(\hat{R}) = E(Y - \hat{Y}) = X\beta - X\beta = 0$$

$$Var(\hat{R}) = \sigma^2 (I-H)$$

Relationship between $\hat{Y}$ and $\hat{R}$

$$\operatorname{Cov}(\underbrace{\hat{Y}}_{HY}, \underbrace{\hat{R}}_{(I-H)Y}) = E\left[\underbrace{HY}\,\underbrace{\{(I-H)Y\}'}\right]$$

$$= E\left[HY\,Y'(I-H)\right]$$

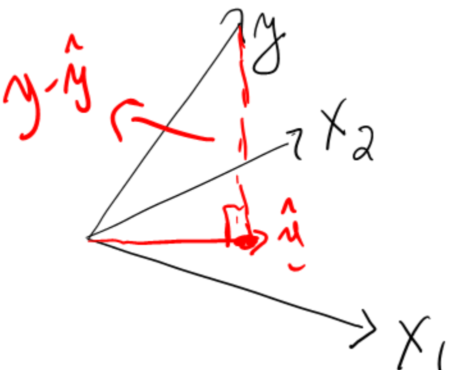
$$= H\,E(YY')(I-H)$$

$$= H\,\sigma^2 I\,(I-H)$$

$$= \sigma^2\,\underbrace{(H)(I-H)} = 0$$

Geometry of least squares

Consider $\underset{\sim}{Y}_{n\times 1}$, $\underset{\sim}{X}_{1\; n\times 1}$, $\underset{\sim}{X}_{2\; n\times 1}$



$\underline{H}$ projects y onto the plane spanned by X_1, X_2

$\Rightarrow \hat{y} = Hy = X\hat{\beta}$

1) minimize the distance between y and $\hat{y} = X\hat{\beta}$

2) Shortest distance is the one that has a right angle between the predicted value and residual

3) residual is orthogonal to the plane spanned by X

4) Score equations:

$$X'(Y - X\beta) = 0$$

Simulation study

- We derived the distribution of the estimated regression coefficients assuming the residuals were Gaussian.
- Does approximate normality of the estimated regression coefficients hold even when the residuals are non-Gaussian?

$$\hat{\beta} \sim MVN\left(\beta, \sigma^2 (X'X)^{-1}\right)$$

Next time....

- Deriving the distribution of linear combinations of regression coefficients
- Deriving the distribution of non-linear combinations of regression coefficients using the Delta method
- LAB: You will generate the distribution of combinations of regression coefficients using bootstrap!