Lecture6 Handout

Elizabeth Colantuoni

2/10/2021

I. Objectives

Upon completion of this session, you will be able to do the following:

- Derive the distribution for the maximum likelihood estimates for MLR based on the properties of functions of Gaussian random variables
- Use vector notation to specify the multiple linear regression model
- Derive the least squares estimators using vector notation

II. Properties of maximum likelihood estimates in MLR

In Lecture 5, we derived the maximum likelihood estimates (MLEs) for $\underset{\sim}{\beta}$ and σ^2 under the classical linear regression model assumptions.

Recall that the MLEs for β_0 and β_1 in the classical **simple linear regression model** can be expressed as:

$$\hat{\beta}_0 = \overline{y} - \beta_1 \overline{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(y_i - \overline{y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Now we will use properties of sums of independent Gaussian random variables to derive the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$.

1. Review of properties of sums of independent Gaussian random variables

Suppose $Y_1, ..., Y_n$ are independent with distribution $N(\mu_i, \sigma_u^2)$ for $i = 1, ..., n$.

Define $d = \sum_{i=1}^{n} a_i Y_i$, a linear combination of Ys with weights a_i.

Then $d \sim N(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2)$.

2. Application to simple linear regression

Now, we will derive the distribution for $\hat{\beta}_1$ and then $\hat{\beta}_0$.

A. Distribution for $\hat{\beta}_1$

We can define $\hat{\beta}_1 = \sum_{i=1}^{n} a_i y_i$, where $a_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{(X_i - \overline{X})}{SSX}$.

Therefore, $\hat{\beta}_1 \sim N(\sum_{i=1}^{n} a_i(\beta_0 + \beta_1 X_i), \sigma^2 \sum_{i=1}^{n} a_i^2)$.

The mean and variance are:

$$
\begin{aligned}
E(\hat{\beta}_1) &= \sum_{i=1}^{n} a_i(\beta_0 + \beta_1 X_i) \\
&= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(\beta_0 + \beta_1 X_i)}{SSX} \\
&= \beta_0 \frac{\sum_{i=1}^{n}(X_i - \overline{X})}{SSX} + \beta_1 \frac{\sum_{i=1}^{n}(X_i - \overline{X})X_i}{SSX} \\
&\quad \text{Note: } \sum_{i=1}^{n}(X_i - \overline{X}) = 0 \\
&\quad \text{Note: } \sum_{i=1}^{n}(X_i - \overline{X})X_i = \sum_{i=1}^{n}(X_i - \overline{X})(X_i - \overline{X}) = SSX \\
&= \beta_1 \text{ i.e. } \hat{\beta}_1 \text{ is an unbiased estimator for } \beta_1
\end{aligned}
$$

and $Var(\hat{\beta}_1) = \frac{\sigma^2}{SSX}$.

B. Distribution for $\hat{\beta}_0$

Given $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{SSX})$, we can derive the distribution for $\hat{\beta}_0$. First, $\hat{\beta}_0$ will be Gaussian given that $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$, a linear function of Gaussian random variables.

$$
\begin{aligned}
E(\hat{\beta}_0) &= E(\overline{Y} - \hat{\beta}_1\overline{X} \\
&= E(\tfrac{1}{n}\sum_{i=1}^{n} Y_i - \hat{\beta}_1\overline{X}) \\
&= E(\tfrac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i) - \hat{\beta}_1\overline{X}) \\
&= E(\beta_0 + \beta_1\overline{X} - \hat{\beta}_1\overline{X}) \\
&= \beta_0 + \beta_1\overline{X} - \beta_1\overline{X} \\
&= \beta_0 \text{ i.e. } \hat{\beta}_0 \text{ is an unbiased estimator for } \beta_0
\end{aligned}
$$

$$
\begin{aligned}
Var(\hat{\beta}_0) &= Var(\overline{Y} - \hat{\beta}_1\overline{X}) \\
&= \tfrac{\sigma^2}{n} - \overline{X}^2 Var(\hat{\beta}_1) \\
&= \tfrac{\sigma^2}{n} - \tfrac{\sigma^2\overline{X}^2}{SSX} \\
&\quad \text{After some algebra....} \\
&= \frac{\sigma^2 \sum_{i=1}^{n} X_i^2}{nSSX}
\end{aligned}
$$

3. Implications for data analysis

Here are some take aways from the calculations above.

1. The estimators for β based on MLE are equal to the least squares solution under the assumption of independent Gaussian residuals.
2. For $j = 1, ..., p$, $\hat{\beta}_j$ is a linear combination of $Y_1, .., Y_n$, so $\hat{\beta}_j$ is also Gaussian if Ys are Gaussian. Further, $\hat{\beta}_j$ will be approximately Gaussian when Ys are not Gaussian with n sufficiently large by the Central Limit Theorem.
3. $\hat{\beta}_j$ is not robust; i.e. one “bad” or “influential” observation can distort results.

III. MLR in vector notation

In this section, we will walk back through the derivations of $\underset{\sim}{\hat{\beta}}$ but expressing the regression models using vector and matrix notation.

Consider the following structure for our regression problem for $i = 1, ..., n$:

$$Y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + ... + X_{pi}\beta_p + \epsilon_i$$

where ϵ_i are independently distributed as $N(0, \sigma^2)$

We can then stack each individuals data into a table structure:

$$\begin{array}{ccccccccccccc}
\underline{i} & & & & & & & & & & & & \\
1 & Y_1 & = & 1 \times \beta_0 & + & X_{11}\beta_1 & + & X_{21}\beta_2 & + & ... & + & X_{p1}\beta_p & = & \epsilon_1 \\
2 & Y_2 & = & 1 \times \beta_0 & + & X_{12}\beta_1 & + & X_{22}\beta_2 & + & ... & + & X_{p2}\beta_p & = & \epsilon_2 \\
. & & & & & & & & & & & & \\
. & & & & & & & & & & & & \\
. & & & & & & & & & & & & \\
\text{n} & Y_n & = & 1 \times \beta_0 & + & X_{1n}\beta_1 & + & X_{2n}\beta_2 & + & ... & + & X_{pn}\beta_p & = & \epsilon_n
\end{array}$$

We can then think about creating vectors that contain the same type of information for each element of our model:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_p \end{pmatrix} \begin{matrix} = \\ = \\ . \\ . \\ . \\ = \end{matrix} \begin{pmatrix} 1 \times \beta_0 \\ 1 \times \beta_0 \\ . \\ . \\ . \\ 1 \times \beta_0 \end{pmatrix} \begin{matrix} + \\ + \\ . \\ . \\ . \\ + \end{matrix} \begin{pmatrix} X_{11}\beta_1 \\ X_{12}\beta_1 \\ . \\ . \\ . \\ X_{1n}\beta_1 \end{pmatrix} \begin{matrix} + \\ + \\ . \\ . \\ . \\ + \end{matrix} \begin{pmatrix} X_{21}\beta_2 \\ X_{22}\beta_2 \\ . \\ . \\ . \\ X_{2n}\beta_2 \end{pmatrix} \begin{matrix} + & ... & + \\ + & ... & + \\ . & . & . \\ . & . & . \\ . & . & . \\ + & ... & + \end{matrix} \begin{pmatrix} X_{p1}\beta_p \\ X_{p2}\beta_p \\ . \\ . \\ . \\ X_{pn}\beta_p \end{pmatrix} \begin{matrix} = \\ = \\ . \\ . \\ . \\ = \end{matrix} \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

We can then express the system of equations in vector notation:

$$\begin{array}{ccccccccccccc}
\underset{\sim}{Y} & = & \underset{\sim}{1}\beta_0 & + & \underset{\sim}{X}_1\beta_1 & + & \underset{\sim}{X}_2\beta_2 & + & ... & + & \underset{\sim}{X}_p\beta_p & = & \underset{\sim}{\epsilon} \\
n \times 1 & & n \times 1 & & n \times 1 & & n \times 1 & & & & n \times 1 & & n \times 1
\end{array}$$

Further, we can append the column vectors represented by $\underset{\sim}{1}$, $\underset{\sim}{X}_1$, $\ldots$, $\underset{\sim}{X}_p$ into a matrix and multiply this matrix with the vector of regression coefficients:

$$\underset{\sim}{Y} = (\underset{\sim}{1}, \underset{\sim}{X}_1, \underset{\sim}{X}_2, ..., \underset{\sim}{X}_p) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{pmatrix} + \underset{\sim}{\epsilon}$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & ... & X_{p1} \\ 1 & X_{12} & X_{22} & ... & X_{p2} \\ . & . & . & ... & . \\ . & . & . & ... & . \\ . & . & . & ... & . \\ 1 & X_{1n} & X_{2n} & ... & X_{pn} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

This organization of the model leaves us with the following matrix representation of the MLR:

$$\underset{n\times 1}{\underset{\sim}{Y}} = \underset{n\times(p+1)}{X} \; \underset{(p+1)\times 1}{\underset{\sim}{\beta}} + \underset{n\times 1}{\underset{\sim}{\epsilon}}$$

How do we express the distribution of $\underset{\sim}{\epsilon}$?

A. Multivariate Gaussian distribution

The multivariate Gaussian disribution describes the marginal and joint distribution of 2 or more Gaussian random variables. In matrix notation, we can define the multivariate Gaussian distribution as:

$$\underset{\sim}{Y} \sim MVN(\underset{\sim}{\mu}, V)$$

where,

$$\underset{\sim}{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{pmatrix}, \underset{\sim}{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_n \end{pmatrix}, \text{ and } V = \begin{bmatrix} v_{11} & v_{12} & . & . & . & v_{1n} \\ v_{21} & v_{22} & . & . & . & v_{2n} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ v_{n1} & v_{n2} & . & . & . & v_{nn} \end{bmatrix}$$

with $v_{ii} = Var(Y_i)$ and $v_{ij} = Cov(Y_i, Y_j)$.

Given our MLR under the assumption that $\epsilon_i iid N(0, \sigma^2)$ then we have:

- $E(Y_i) = \mu_i = X_i \underset{\sim}{\beta}$ where $X_i = (1, X_{1i}, X_{2i}, ..., X_{pi})$ (the ith row of X)
- $Var(Y_i) = \sigma^2$ and $Cov(Y_i, Y_j) = 0$

And we can express the multivariate normal distribution for $\underset{\sim}{\epsilon}$ as $\underset{\sim}{\epsilon} \sim N(\underset{\sim}{0}, \sigma^2 I)$ where I is the identity matrix with $1s$ on the diagonal elements and $0s$ on the off-diagonal elements.

B. Maximum likelihood estimation using vector notation

Using vector notation, our MLR is:

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

In the remainder of this section, I will drop the $\sim$ so you should assume Y and β are $n \times 1$ and $(p+1) \times 1$ vectors, respectively, and X is the $n \times (p+1)$ design matrix.

Our goal is to select estimates of β and σ^2 to minimize $\sum_{i=1}^{n} r_i(\hat{\beta}) = \sum_{i=1}^{n} (y_i - X_i\beta)^2$.

Using the vector notation, we can express the sums of squared residuals as:

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

where Y' is the transpose of Y: if Y is a $n \times 1$ vector, then Y' is a $1 \times n$ vector. If X is a $n \times (p+1)$ matrix, then X' is a $(p+1) \times n$ matrix.

The score equations for β can be written as:

$$U_\beta(\beta) = \frac{\partial}{\partial\beta}(Y - X\beta)'(Y - X\beta) = X'(Y - X\beta) = 0$$

Solving the score equations for β, we have:

$$\begin{aligned} X'Y - X'X\beta &= 0 \\ X'Y &= (X'X)\beta \\ (X'X)^{-1}X'Y &= \hat{\beta} \end{aligned}$$