# Lecture5 Handout

Elizabeth Colantuoni

2/8/2021

## I. Objectives

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

- State the multiple linear regression model and the least squares solution
- Describe the least squares solution for simple linear regression
- Derive the maximum likelihood estimates of the regression coefficients and residual variance assuming the residuals are normally distributed
- Demonstrate for simple linear regression, the maximum likelihood solution is the least square solution

## II. Multiple linear regression model

Consider a population of interest where we define an outcome of interest ($Y$) and a set of explanatory variables ($X_1, X_2, ..., X_p$). $Y$ is a random variable and we assume $X_1, X_2, ..., X_p$ are fixed (i.e. not random variables, no measurement error)

For a sample of size $n$, we observe $(y_i, X_{1i}, X_{2i}, ..., X_{pi})$ for each $i = 1, 2, ..., n$.

The classical multiple linear regression (MLR) model assumes:

$$Y_i = \mu_i(\underset{\sim}{\beta}, X_i) + \epsilon_i$$

where

- $X$ represents the design matrix, which includes a column of $1s$ and the vectors $X_1, X_2, ..., X_p$ where each vector contains the data defining $X_j$ for observation $i = 1, ..., n$, $j = 1, ..., p$
- $X_i$ represents the row of the design matrix corresponding to the values of the explanatory variables for subject $i$, i.e. $(1, X_{1i}, X_{2i}, ..., X_{pi})$.
- $\underset{\sim}{\beta}$ represens the vector of regression coefficients $\beta_0, \beta_1, ..., \beta_p$.
- $\mu_i(\underset{\sim}{\beta}, X_i)$ is the systematic component and $\epsilon_i$ is the random component
- $\epsilon_i \sim N(0, \sigma^2), Cov(\epsilon_i, \epsilon_j) = 0$
- $\mu_i(\underset{\sim}{\beta}, X_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ...\beta_p X_{pi}$

### A. Least Squares Solution

The goal of MLR is to estimate $\underset{\sim}{\beta}$ and $\sigma^2$ using the observed data $(y_i, X_{1i}, X_{2i}, ..., X_{pi})$ for $i = 1, ..., n$.

The (ordinary) least squares solution finds the values of $\underset{\sim}{\beta}$ that minimizes:

$$\sum_{i=1}^{n}(y_i - \mu_i(\underset{\sim}{\beta}, X_i))^2$$

i.e. the least squares solution finds $\underset{\sim}{\beta}$ that minimizes the squared residuals (or errors)

### B. Least squares solution for SLR

For a simple linear regression, i.e. p = 1:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

the least squares estimates for $\beta_0$ and $\beta_1$ are:

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

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{X}$$

where $\overline{y}$ and $\overline{X}$ are the sample average values of $y$ and $X$ respectively.

Note: The formula for $\hat{\beta}_1$ is the ratio of the covariance between $y$ and $X$ and the variance of $X$.

Note: If we standardize $y$ and $X$ such that $\overline{y} = \overline{(}X) = 0$ and $SD(y) = SD(X) = 1$ then $\hat{\beta}_1 = r$, the Pearson correlation coefficient.

## III. Maximum Likelihood Inference in MLR

In this section, we will derive the maximum likelihood estimates for $\underset{\sim}{\beta}$ and $\sigma^2$ and functions thereof (next time) to address scientific questions of interest.

The maximum likelihood procedure utilizes the assumptions of our MLR model plus the data we observe:

$$\text{Model: } Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2) \text{ and independent}$$

$$\text{Data: } (y_i, X_i), i = 1, ..., n$$

The principle of maximum likelihood estimation is to choose values of $\underset{\sim}{\beta}$ and $\sigma^2$ that make the observed data maximally likely to have occurred.

More notation:

- $Y_i$ is a random variable, $y_i$ is the observation of the random variable for subject $i$

- $\underset{\sim}{Y}$ and $\underset{\sim}{y}$ are vectors (lists or matrices each with single column) of independent random variables or observations
- $X_i$ is the row-vector of explanatory variables for subject $i$ and are fixed/known (i.e. not random variables)
- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_i X_{1i} + ... + \hat{\beta}_p X_{pi}$ is the predicted value or the estimated mean of the distribution of $Y$ at the value of $X_i$
- $\hat{R}_i = Y_i - \hat{Y}_i$ is the residual

## A. Likelihood function definition

Under the assumption that $e_i \sim N(0, \sigma^2)$, $Y_i$ are random and $X_i$ are fixed, then we have that:

$$Y_i \sim N(\mu_i(\underset{\sim}{\beta}, X_i), \sigma^2)$$

Then we can think about two quantities:

- $f(\underset{\sim}{y}|\mu(\underset{\sim}{\beta}, X), \sigma^2) = \prod_{i=1}^{n} f(y_i|\mu_i(\underset{\sim}{\beta}, X_i), \sigma^2)$ by independence, which is the probability density function for $\underset{\sim}{y}$, which is viewed as a function of $\underset{\sim}{y}$ with $\mu(\underset{\sim}{\beta}, X)$ and $\sigma^2$ fixed. From the probability density function, we would answer questions about the probability of observing $\underset{\sim}{y}$ in certain ranges given $\mu(\underset{\sim}{\beta}, X)$ and $\sigma^2$.
- $L(\mu(\underset{\sim}{\beta}, X), \sigma^2|\underset{\sim}{y}) = \prod_{i=1}^{n} L(\mu_i(\underset{\sim}{\beta}, X_i), \sigma^2|y_i)$ by independence, which is the likelihood function. The likelihood function is viewed as a function of $\mu(\underset{\sim}{\beta}, X)$ and $\sigma^2$ for fixed $\underset{\sim}{y}$.
- In maximum likelihood estimation, we identify the values of $\underset{\sim}{\beta}$ and $\sigma^2$ that maximize the likelihood function given $\underset{\sim}{y}$.

## B. Maximum likelihood estimation under gaussian residuals

Given the normality assumption, we can write the likelihood function as:

$$\begin{aligned}
L(\underset{\sim}{\beta}, \sigma^2|\underset{\sim}{y}) &= \prod_{i=1}^{n} L(\mu_i(\underset{\sim}{\beta}, X_i), \sigma^2|y_i) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{1}{2\sigma^2}(y_i - \mu_i(\underset{\sim}{\beta}, X_i))^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - ... - \beta_p X_{pi})^2)
\end{aligned}$$

Similarly, we can express the log likelihood function as:

$$\begin{aligned}
l(\underset{\sim}{\beta}, \sigma^2|\underset{\sim}{y}) &= log L(\underset{\sim}{\beta}, \sigma^2|\underset{\sim}{y}) \\
&= \sum_{i=1}^{n} \left(-\frac{1}{2} log(2\pi) - log(\sigma) - \frac{1}{2\sigma^2}(y_i - \mu_i(\underset{\sim}{\beta}, X_i))^2\right)
\end{aligned}$$

Now find $\hat{\underset{\sim}{\beta}}$ and $\hat{\sigma}^2$ that maximize $l(\underset{\sim}{\beta}, \sigma^2|\underset{\sim}{y})$ by differentiating with respect to $\underset{\sim}{\beta}$ and $\sigma^2$, setting equal to 0 and solving for $\hat{\underset{\sim}{\beta}}$ and $\hat{\sigma}^2$.

### 1. Solution for $\beta_j$

Notice that $l(\underset{\sim}{\beta}, \sigma^2|\underset{\sim}{y})$ as a function of $\underset{\sim}{\beta}$ is proportional to $\displaystyle\sum_{i=1}^{n}(y_i - \mu_i(\underset{\sim}{\beta}, X_i))^2$.

Define the score equation for $\beta_j$ as:

$$
\begin{aligned}
U_{\beta_j}(\underset{\sim}{\beta}|\sigma^2) &= \tfrac{\partial}{\partial \beta_j} l(\underset{\sim}{\beta}, \sigma^2|\underset{\sim}{y}) \\
&= \tfrac{\partial}{\partial \beta_j} \sum_{i=1}^{n} \left( -\frac{1}{2\sigma^2}(y_i - \mu_i(\underset{\sim}{\beta}, X_i))^2 \right) \\
&= -\tfrac{2}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i(\underset{\sim}{\beta}, X_i))(-X_{ij})
\end{aligned}
$$

We then take the score equations, set them equal to 0 and solve for $\hat{\underset{\sim}{\beta}}$. Therefore, we solve the following system of equations (also known as score equations) with $p+1$ equations and $p+1$ unknown parameters:

$$
U_{\underset{\sim}{\beta}} = \sum_{i=1}^{n} (y_i - \mu_i(\underset{\sim}{\beta}, X_i)) \begin{bmatrix} 1 \\ X_{1i} \\ X_{2i} \\ . \\ . \\ . \\ X_{pi} \end{bmatrix} = 0
$$

Notice that when solving for $\underset{\sim}{\beta}$, we do not need to know $\sigma^2$.

NOTE: By looking at the score equations, you will see that the maximum likelihood estimate for $\underset{\sim}{\beta}$ will be achieved when minimizing the sums of squared residuals, i.e. the least squares solution.

NOTE: To confirm that our estimates in fact maximize the likelihood function, you would need to take the second derivative of the score equations and show they are negative when evaluated at the maximum likelihood estimate.

### 2. Solution for $\sigma^2$

Given $\hat{\underset{\sim}{\beta}}$, the score equation for $\sigma^2$ is:

$$
\begin{aligned}
U_{\sigma^2}(\hat{\underset{\sim}{\beta}}) &= \tfrac{\partial}{\partial \sigma^2} \sum_{i=1}^{n} \left( -log(\sigma) - \frac{1}{2\sigma^2}(y_i - \mu_i(\hat{\underset{\sim}{\beta}}, X_i))^2 \right) \\
&= \sum_{i=1}^{n} \left( -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y_i - \mu_i(\hat{\underset{\sim}{\beta}}, X_i))^2 \right)
\end{aligned}
$$

Setting the score equation equal to zero and solving for $\sigma^2$, the maximum likelihood estimate is:

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_i(\underset{\sim}{\hat{\beta}}, X_i))^2$$

You can show that $E(\hat{\sigma}^2_{MLE}) = \frac{n-(p+1)}{n}\sigma^2$.

Therefore, in practice, we use the following unbiased estimate for $\sigma^2$:

$$\tilde{\sigma}^2 = \frac{1}{n-(p+1)}\sum_{i=1}^{n}(y_i - \mu_i(\underset{\sim}{\hat{\beta}}, X_i))^2$$

### 3. Simple linear regression, p = 1

For a simple linear regression model, we are required to solve the following equations:

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i) &= 0 \\ \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i)X_i &= 0 \end{aligned}$$

Solving the first equation, we have:

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i) &= 0 \\ \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1\sum_{i=1}^{n} X_i &= 0 \\ \sum_{i=1}^{n} y_i - \beta_1\sum_{i=1}^{n} X_i &= n\beta_0 \\ \tfrac{1}{n}\sum_{i=1}^{n} y_i - \beta_1\frac{1}{n}\sum_{i=1}^{n} X_i &= \beta_0 \\ \overline{y} - \beta_1\overline{X} &= \hat{\beta}_0 \end{aligned}$$

Solving the second equation, we have:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i-\beta_0-\beta_1X_i)X_i &= 0 \\
\sum_{i=1}^{n}(y_i-(\overline{y}-\beta_1\overline{X})-\beta_1X_i)X_i &= 0 \\
\sum_{i=1}^{n}\left[y_iX_i-\overline{y}X_i-\beta_1(X_i-\overline{X})X_i\right] &= 0 \\
\sum_{i=1}^{n}X_i(y_i-\overline{y})-\beta_1\sum_{i=1}^{n}X_i(X_i-\overline{X}) &= 0 \\
\frac{\sum_{i=1}^{n}X_i(y_i-\overline{y})}{\sum_{i=1}^{n}X_i(X_i-\overline{X})} &= \beta_1 \\
\frac{\sum_{i=1}^{n}(X_i-\overline{X})(y_i-\overline{y})}{\sum_{i=1}^{n}(X_i-\overline{X})^2} &= \beta_1
\end{aligned}$$

NOTE the following:

$$\begin{aligned}
\sum(y_i-\overline{y})X_i &= \sum y_iX_i-\overline{y}\sum X_i \\
&= \sum y_iX_i-n\overline{yx} \\
&= \sum y_iX_i-n\overline{yx}-n\overline{yx}+n\overline{yx} \\
&= \sum\left[y_iX_i-\overline{y}X_i-\overline{X}y_i+\overline{y}\overline{X}\right] \\
&= \sum(y_i-\overline{y})(X_i-\overline{X})
\end{aligned}$$

## C. Take away messages

1. For $\epsilon_i$ assumed to be *iid* $N(0,\sigma 2)$, the least squares solution is the maximum likelihood solution
2. Each $\beta_j$ for $j=1,...,p$ is a linear function of $\underset{\sim}{y}$:

$$\hat{\beta}_j=\sum_{i=1}^{n}w_{ij}(X_i)y_i$$

When $p=1$ (i.e. SLR), $w_{ij}(X_i)=\frac{(X_i-\overline{X})}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$.

3. $\underset{\sim}{\hat{\beta}}$ is not robust to outliers
4. Observations with large $(X_i-\overline{X})$ have greater weights, or more leverage.

5. At $X_i = \overline{X}$, what is $\hat{y}_{\overline{X}}$?

$$\begin{aligned}
\hat{y}_{\overline{X}} &= \hat{\beta}_0 + \hat{\beta}_1 \overline{X} \\
&= \overline{y} - \hat{\beta}_1 \overline{X} + \hat{\beta}_1 \overline{X} \\
&= \overline{y}
\end{aligned}$$