Multiple Linear Regression

Lecture 4

Alexandra Chronopoulou

I ILLINOIS

COLLEGE OF LIBERAL ARTS & SCIENCES

Department of Statistics
101 Illini Hall, MC-374
725 S. Wright St.
Champaign, IL 61820-5710

Learning objectives

In this lecture we will:

- introduce Multiple Linear Regression (MLR)
- Derive LS estimators in the general case
- Discuss the geometric representation of MLR

Single Predictor

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \;\; n = 1, \ldots, n$$

One Response y vs. *One Predictor* x

Multiple Predictors?

- x_1, x_2, ..., x_p be p predictors of a response y.
- The data will be of the form:

$$\begin{array}{cccccc} y_1 & & x_{11} & x_{12} & \cdots & x_{1p} \\ y_2 & & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ y_n & & x_{n1} & x_{n2} & \cdots & x_{np} \end{array}$$

Model Equation

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n$$

where we denote $\mathbf{x_i} = (x_{i1,\ldots,x_i p})^T$, with $x_{i1} = 1$.

- $(\beta_1, \beta_2, \ldots, \beta_p; \sigma^2)$ are unknown true parameters.
  - β_1 is the **intercept**.
  - $\beta_2, \beta_3, \ldots, \beta_p$ are partial slopes.
  - σ^2 is the **error variance**
- $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are the random errors. They usually assumed to satisfy the same conditions as in simple linear regression:
  - **zero mean**: $\mathbb{E}(\varepsilon_i) = 0$
  - **uncorrelated**: $\mathrm{Cov}(\varepsilon_\mathrm{i}, \varepsilon_\mathrm{j}) = 0,\ \mathrm{i} \neq \mathrm{j})$, and
  - **homoscedastic**: $\mathrm{Var}(\varepsilon_\mathrm{i}) = \sigma^2$ does not depend on i).

Matrix Representation

Define:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

So, the model equation can be written as:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Multiple Linear Regression (MLR)

Matrix Representation of the MLR Model:

$$\begin{array}{ccccccc} \mathbf{y}_{n\times 1} & = & \mathbf{X}_{n\times p} & \beta_{p\times 1} & + & \varepsilon_{n\times 1} \\ \uparrow & & \uparrow & \uparrow & & \uparrow \\ \text{Response} & & \text{Design} & \text{Coefficients} & & \text{Error} \\ & & \text{Matrix} & & & \text{Term} \end{array}$$

- n: sample size
- p: number of predictors or columns of $\mathbf{X}$
- By default the intercept is included in the model in which case the first column of $\mathbf{X}$ is a vector of 1's.

Least-Squares Estimation

Goal: Parameter Estimation

- We want to estimate β, i.e. obtain:

$$\hat{\beta} \;=\; \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\right)^T$$

- The LS estimator of β minimizes the sum of squared residuals:

$$RSS \;=\; ||y - \mathbf{X}\beta||^2 = (y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta)$$

In order to minimize $RSS = (y - \mathbf{X}\beta)^T(y - \mathbf{X}\beta)$, we take derivatives with respect to β's and set to zero (as before).

$$\frac{\partial RSS}{\partial \beta} = -2\, \mathbf{X}^T_{p\times n}(y - \mathbf{X}\beta)_{n\times 1} = \mathbf{0}_{p\times 1}$$

$$\mathbf{X}^T(y - \mathbf{X}\beta) = \mathbf{0} \;\longrightarrow\; \text{Normal Equations}$$

$$(\mathbf{X}^T\mathbf{X})\,\beta = \mathbf{X}^T\,y$$

$$\boxed{\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,y} \;\longrightarrow\; \text{LS Estimators}$$

Remarks

1. We assume that the rank of $\mathbf{X}$ is p, i.e. no columns of $\mathbf{X}$ is a linear combinations of the other columns of $\mathbf{X}$.

2. Since $\mathbf{X}$ has rank p, the inverse of $(\mathbf{X}^T\mathbf{X})$ exists.

Single Predictor Model

- Recall that the single predictor model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\;\; n = 1, \ldots, n$$

- If we re-write it in matrix format, we have:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- Use the formula $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\, y$ to obtain the estimator from the previous lecture.

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & \sum_i x_i^2 \end{pmatrix}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n\sum_i x_i^2 - (n\bar{x})^2} \begin{pmatrix} \sum_i x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix}$$

$$\mathbf{X}^T y = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} n\bar{y} \\ \sum_i x_i y_i \end{pmatrix}$$

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$
$$= \frac{1}{n\sum_i x_i^2 - (n\bar{x})^2} \begin{pmatrix} \sum_i x_i^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} \begin{pmatrix} n\bar{y} \\ \sum_i x_i y_i \end{pmatrix}$$

So, $\hat{\beta}_1$ is given by

$$\hat{\beta}_1 = \frac{-n^2\bar{x}\bar{y} + n\sum_i x_i y_i}{n\sum_i x_i^2 - (n\bar{x})^2} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{S_{XY}}{S_{XX}}$$

and similarly we can recover the formula for $\hat{\beta}_0$.

Fitted Values

$$
\begin{aligned}
\hat{y}_{n\times 1} &= \mathbf{X}\hat{\beta} \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\, y \\
&= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\, y := \mathbf{H}_{n\times n} y_{n\times 1}
\end{aligned}
$$

$\mathbf{H}_{n\times n}$ is called the *hat matrix*, since it returns the "y-hat" values.

Residuals

$$\mathbf{r}_{n\times 1} = y - \hat{y} = y - \mathbf{H}y = (\mathbf{I} - \mathbf{H})y$$

The residuals $\mathbf{r}$ are used to estimate the *error variance*:

$$\hat{\sigma}^2 = \frac{1}{n-p}\sum_i r_i^2 = \frac{RSS}{n-p}$$

Normal Equations & Residuals

- The LS estimator is the β that satisfies the *normal equations*, that is

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

- This implies the following properties for the residuals, $r_{n\times 1} = y - \mathbf{X}\hat{\beta}$:

1. The cross-products between the residual vector r and *each column of* $\mathbf{X}$ are zero, i.e.

$$\begin{aligned}\mathbf{X}^T\, r &= \mathbf{X}^T(y - \mathbf{X}\hat{\beta}) = \mathbf{X}^T y - \mathbf{X}^T\mathbf{X}\hat{\beta} \\ &= \mathbf{X}^T y - (\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\, y = 0\end{aligned}$$

2. The cross-product between the fitted value $\hat{y}$ and the residual vector r is zero, i.e.

$$\hat{y}^T\, r = \hat{\beta}^T X^T\, r = 0$$

This implies that the residual vector r is orthogonal to each column of X and $\hat{y}$.

Properties

- Let c be any linear combination of the columns of $\mathbb{X}$, then

$$\mathbf{H}c = c$$

  since $\mathbf{HX} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{X}$.

- *Symmetric*, since $\mathbf{H}^T = (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{H}$.

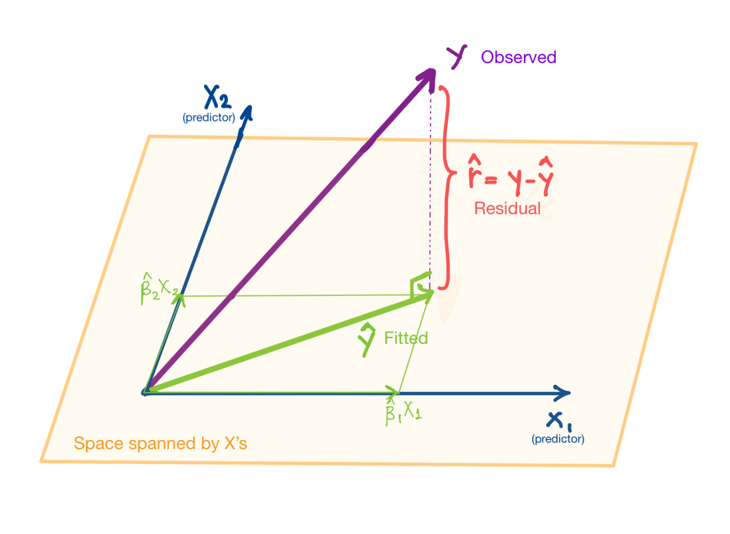
- *Idempotent*, i.e. $\mathbf{HH} = \mathbf{HH}^T = \mathbf{H}^\mathbf{T}\mathbf{H} = \mathbf{H}$. Indeed,

$$\mathbf{HH} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = H$$

  This also implies that $\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{0}_{n\times n}$.

- trace($\mathbf{H}$) = p, the number of LS coefficients we estimated.

Geometric Representation of LS



Estimation Space

- The columns of $\mathbf{X}$ span a p-dimensional subspace in $\mathbb{R}^n$. This is a subspace that consists of vectors that can be written as linear combinations of the columns of $\mathbf{X}$.
- The LS squares estimator $\hat{\beta}$ is obtained by minimizing the Euclidean distance between the vectors $\mathbf{y}$ and $\hat{\mathbf{y}}$, i.e. $||y - \hat{y}||^2$. $\hat{y}$is the projection of y onto the estimation space.
- $\mathbf{H}_{n \times n}$, projection/hat matrix is symmetric, unique, and idempotent.

Error Space

- The error space is an $(n - p)$-dimensional space that is orthogonal to the estimation space. The *projection matrix* of the error space is $(\mathbf{I} - \mathbf{H})$.
- The residual $\mathbf{r}$ is the projection of $\mathbf{y}$ onto the error space, orthogonal to the estimation space. So, $\mathbf{r}$ is orthogonal to *any* vector in the estimation space, including each column of $\mathbf{X}$.
- When the intercept is included in the model, then

$$\sum_{i=1}^{n} r_i = 0$$

In general, $\sum_{i=1}^{n} r_i X_{ij} = 0$, $j = 1, \ldots, p$ due to the normal equations.

Goodness of Fit: R-Square

- A measure of how well the model fits the data is the R-square or the so-called *coefficient of determination* or *percentage of variance explained*:

$$R^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{RSS}{TSS}$$

- An equivalent definition is

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

- $0 \leq R^2 \leq 1$

Rank deficiency

- The design matrix $\mathbf{X}$ is an $n \times p$ matrix[1]. If this matrix is not of full rank (i.e., its columns are not linearly independent), the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ can not be inverted (singular matrix).
- If the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$ is singular the LS solutions is not unique (identifiability problem)
- **R** can cope well with this problem. To solve the LS equations **R** uses the QR decomposition. You can read more on this in the supplemental material.

[1] You can use function *model.matrix(.)* in **R** to extract the model matrix of a fitted model