

Module 9: Linear regression

Siyue Yang

06/17/2022

Outline

In this module, we will review linear regression.

Linear regression

- Model:

$$Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \epsilon_{n \times 1}$$

- Equivalently:

$$y_i = \mathbf{x}_i^T \beta + \epsilon_i, \quad i = 1, \dots, n$$

Linear regression

- Model:

$$Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \epsilon_{n \times 1}$$

- Equivalently:

$$y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \dots, n$$

- Standard assumptions

- y_i independent (equivalently ϵ_i independent)
- $\mathbb{E}(\epsilon_i) = 0$
- $\text{var}(\epsilon_i) = \sigma^2$, constant
- x_i known, β to be estimated

Linear regression

- Model:

$$Y_{n \times 1} = X_{n \times p} \beta_{p \times 1} + \epsilon_{n \times 1}$$

- Equivalently:

$$y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \dots, n$$

- Standard assumptions

- y_i independent (equivalently ϵ_i independent)
- $\mathbb{E}(\epsilon_i) = 0$
- $\text{var}(\epsilon_i) = \sigma^2$, constant
- x_i known, β to be estimated

- More concisely:

$$\mathbb{E}(Y \mid X) = X\beta, \quad \text{var}(Y \mid X) = \sigma^2 I$$

Interpretation of β_j

- Effect on the expected response of a unit change in j th explanatory variable, all other variables held fixed

Least squares estimation

- Definition (minimize the residuals)

$$\hat{\beta}_{\text{LS}} := \min_{\beta} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2$$

- Equivalently,

$$\hat{\beta}_{\text{LS}} := \min_{\beta} (y - X\beta)^T (y - X\beta)$$

- Equivalently (L2 distance),

$$\hat{\beta}_{\text{LS}} := \min_{\beta} \|y - X\beta\|_2^2$$

- Equivalently, $\hat{\beta}$ is the solution of the score equation

$$X^T (y - X\beta) = 0$$

- Solution

$$\hat{\beta}_{\text{LS}} = (X^T X)^{-1} (X^T \mathbf{y})$$

Another interpretation

- think of this is as the projection of Y onto the linear subspace spanned by the columns of \mathbf{X}

Least squares estimation (cont'd)

Assume X is fixed,

- Expected value

$$\mathbb{E}(\hat{\beta}_{LS}) = (X^T X)^{-1} X^T \mathbb{E}(y) = (X^T X)^{-1} (X^T X) \beta = \beta$$

- Variance

$$\begin{aligned}\text{var}(\hat{\beta}_{LS}) &= (X^T X)^{-1} X^T \text{var}(y) X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1}\end{aligned}$$

Assumptions for ordinary least squares

- **Linearity**: the expectation of Y is linear in $X_1 \dots X_p$
- **Independence**: the ϵ_i are independent
- **Mean zero errors**: the ϵ_i have mean zero, i.e. $E[\epsilon_i] = 0$
- **Equal variance (homoscedasticity)**: the ϵ_i have the same variance, i.e. $\text{Var}[\epsilon_i] = \sigma^2$

What about normal distribution?

- If we further assume $\epsilon_i \sim N(0, \sigma^2)$ (and independent across i), then
- $y \mid X \sim N(X\beta, \sigma^2 I)$, and
- likelihood function is

$$L(\beta, \sigma^2; y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) \right\}$$

- log-likelihood function is

$$\ell(\beta, \sigma^2; y) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta)$$

- maximum likelihood estimate of β is

$$\hat{\beta}_{ML} = (X^T X)^{-1} X^T y = \hat{\beta}_{LS}$$

What about normal distribution? (cont'd)

- distribution of $\hat{\beta}$ is normal

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

- distribution of $\hat{\beta}_j$ is

$$N \left(\beta_j, \sigma^2 (X^T X)^{-1}_{jj} \right), \quad j = 1, \dots, p$$

- maximum likelihood estimate of σ^2 is

$$\frac{1}{n} (y - X\hat{\beta})^T (y - X\hat{\beta})$$

- but we use

$$\tilde{\sigma}^2 = \frac{1}{n-p} (y - X\hat{\beta})^T (y - X\hat{\beta})$$

Maximum likelihood estimation vs. OLS

- We did not place any distributional assumptions on the outcome,
 - We only required that $E[\epsilon_i] = 0$ with constant variance
 - In other words, OLS is a semiparametric method

Maximum likelihood estimation vs. OLS

- We did not place any distributional assumptions on the outcome,
 - We only required that $E[\epsilon_i] = 0$ with constant variance
 - In other words, OLS is a semiparametric method
- Sometimes, people assume that $\epsilon_i \sim N(0, \sigma^2)$, which means

$$Y_i \sim N(\beta_0 + \beta_1 X_{i1} + \dots + \beta_p X_{ip}, \sigma^2)$$

- If this additional assumption is made, then we can instead use maximum likelihood estimation for β
- This connects to a whole other class of models called generalized linear models (GLMs)
- Interestingly, in this case, you will end up with the same estimates for β

Resources

This tutorial is based on

- Nancy Reid's STA2101 Methods of Applied Statistics [links]
- Harvard's Biostatistics Preparatory Course Methods [links].