Module 7: Linear regression

Jianhui Gao

July 18, 2025

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:

$$y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad i = 1, \ldots, n$$

Linear regression

- Model:

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

- Equivalently:

$$y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad i = 1, \ldots, n$$

- Standard assumptions
  - y_i independent (equivalently ϵ_i independent)
  - $\mathbb{E}(\epsilon_i) = 0$
  - $\operatorname{var}(\epsilon_i) = \sigma^2$, 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:

$$y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad i = 1, \ldots, n$$

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

- More concisely:

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

Interpretation of β_j

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

Least squares estimation

- Definition (minimize the residuals)

$$\hat{\beta}_{\mathrm{LS}} := \min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^{\mathrm{T}} \beta \right)^2$$

- Equivalently,

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

- Equivalently (L2 distance),

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

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

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

- Solution

$$\hat{\beta}_{\mathrm{LS}} = \left(X^{\mathrm{T}} X \right)^{-1} \left(X^{\mathrm{T}} \boldsymbol{y} \right)$$

Another interpretation: the projection of Y onto the linear subspace spanned by the columns of $\mathbf{X}$

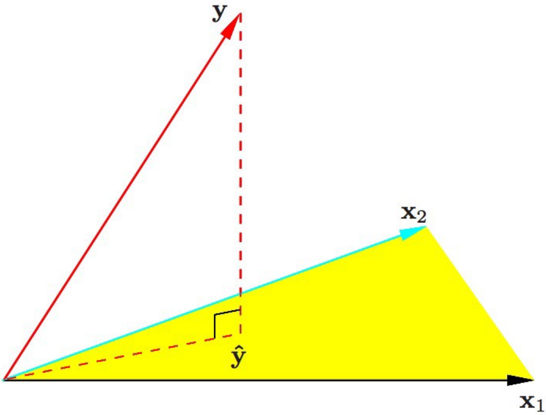

FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector $\mathbf{y}$ is orthogonally projected onto the hyperplane spanned by the input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions*

Least squares estimation (cont'd)

Assume X is fixed,

- Expected value

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

- Variance

$$\begin{aligned}\mathrm{var}\left(\hat{\beta}_{LS}\right) &= \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}\,\mathrm{var}(y)X\left(X^{\mathrm{T}}X\right)^{-1}\\ &= \left(X^{\mathrm{T}}X\right)^{-1} X^{\mathrm{T}}\sigma^2 IX\left(X^{\mathrm{T}}X\right)^{-1}\\ &= \sigma^2\left(X^{\mathrm{T}}X\right)^{-1}\end{aligned}$$

Assumptions for ordinary least squares

- **Linearity**: the expectation of Y is linear in $X_1 \ldots X_p$
- **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 $\epsilon_i \sim N\left(0, \sigma^2\right)$ (and independent across i), then
- $y \mid X \sim N\left(X\beta, \sigma^2 I\right)$, and
- likelihood function is

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

- log-likelihood function is

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

- maximum likelihood estimate of β is

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

What about normal distribution? (cont'd)

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

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

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

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

- maximum likelihood estimate of σ^2 is

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

- but we use

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

Maximum likelihood estiamtion 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 estiamtion vs. OLS

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

$$Y_i \sim N\left(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_1 X_{ip}, \sigma^2\right)$$

  - 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].