Methods of Data Analysis II

Derek Li

Contents

1 Review

1.1 Linear Model

Recall the general linear model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2).$$

Thus,

$$Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(X_i^T \beta, \sigma^2) \text{ and } Y \sim \mathcal{N}_n(X\beta, \sigma^2 I).$$

Recall the assumptions of linear regression:

1. Errors are independent (observations are independent);
2. Errors are identically distribute with $\mathbb{E}[\varepsilon_i] = 0$;
3. Homoscedasticity: $\text{Var}[\varepsilon_i] = \sigma^2$;
4. A straight-line relationship exists between ε_i and y_i.

1.2 ANOVA

Analysis of variance (ANOVA) is used to test differences between two or more means and to test general rather than specific differences among means. We test

$$H_0 : \mu_1 = \cdots = \mu_n \text{ against } H_1 : \text{At least one mean is different from the others.}$$

Recall the assumptions of ANOVA:

1. Errors are independent;
2. Errors are normally distributed with $\mathbb{E}[\varepsilon_i] = 0$;
3. Homoscedasticity: $\text{Var}[\varepsilon_i] = \sigma^2$.

Note that the normality assumption can be relaxed if sample size is large (Central Limit Theorem). The normality assumption is most important when n is small, highly non-normal or small effect size.

In fact, ANOVA is just a specific case of the general linear model.