Lecture 4: Continuing statistical simulations

Last time

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

Question: How important is it that $\varepsilon_i \sim N(0, \sigma^2)$? Does it matter if the errors are *not* normal?

- simulate data from different distributions for ε_i (normal, χ^2, exponential, ...) and different sample sizes n
- constructed 95% CI intervals, compared actual coverage w/ desired coverage (95%)

ADEMP: A useful framework for simulation studies

- **Aims:** Why are we doing the study?
- **Data generation:** How are the data simulated?
- **Estimand/target:** What are we estimating for each simulated dataset?
- **Methods:** What methods are we using for model fitting, estimation, etc?
- **Performance measures:** How do we measure performance of our chosen methods?

ADEMP

For the normal errors simulation study:

- **Aims:** Assess the importance of the normality for ε_i in linear regression model
- **Data generation:**
- **Estimand/target:** β_1
- **Methods:** fit linear regression of Y on X (using lm function in R), calculate a 95% CI for β_1 ($\hat{\beta}_1 \pm t^* SE(\hat{\beta}_1)$)
- **Performance measures:** observed coverage of 95% CIs for β_1

$n = 100$ or $n = 10$ or $n = 1000$

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

$\beta_0 = 0.5$

$\beta_1 = 1$

$X_i \sim \text{Uniform}(0,1)$

$\varepsilon_i \sim \text{Normal}$

or $\varepsilon_i \sim \text{Gamma}$

or $\varepsilon_i \sim \chi^2$

(could use other distributions too)

} want to compare performance for these different distributions

Another question

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

Question: How important is it that ε_i have constant variance?

With a neighbor, discuss the ADEMP steps you might use to answer this question (some of them will be similar to the normal simulation!). Then we will discuss together as a group.

ADEMP steps

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

Question: How important is it that ε_i have constant variance?

A : assessing the importance of the constant variance assumption

D $\varepsilon_i \sim N(0, 1)$ (satisfies constant variance) or or $\varepsilon_i \sim N(0, |X_i|)$ $\varepsilon_i \sim N(0, X_i^2)$ etc..

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

$\beta_0 = 0.5 \quad \beta_1 = 1 \quad X_i \sim \text{Uniform}(0, 1)$

E : β_1

M : fit linear regression model (lm in R), calculate 95% CI

P : observed coverage of "95%" CIs

Class activity

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

How important is the constant variance assumption?

https://sta279-s24.github.io/class_activities/ca_lecture_4.html