Simulation Study Lab

1 Objective

The objective of this assignment is to guide you through designing and conducting a simulation study. You will investigate how overfitting and underfitting a regression model can impact the bias and variance of coefficient estimates.

2 Background and Motivation

In statistical modeling, selecting the appropriate predictors is crucial. Including too many irrelevant variables (overfitting) or omitting important ones (underfitting) can lead to misleading results.

Example Scenario:

Suppose we are studying the effect of two medications, Drug A and Drug B, on blood pressure reduction. The true effect on blood pressure (y) depends on both drugs:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where:

- x_1: Dosage of Drug A
- x_2: Dosage of Drug B
- ϵ: Random error term

If we only include x_1 in our model (omitting x_2), we might misestimate the effect of Drug A due to the omitted variable bias. Similarly, adding an unrelated variable x_3 (e.g., daily water intake, which has no effect) might inflate variance without improving the model.

3 Your Task

Design a simulation study to explore how overfitting and underfitting affect the estimates of regression coefficients.

Before moving on to the next page, try to think of how you would set up this simulation study

3.1 Possible Steps to Follow

1. **Define the True Model:**
   Decide on the true underlying model that generates the data. For example:

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

   Choose realistic values for β_0, β_1, β_2, and the variance of ϵ.

2. **Generate Predictor Variables:**

   - Decide how to generate x_1 and x_2. For instance, you might draw them from a normal or uniform distribution.
   - If you include an irrelevant variable x_3, decide how to generate it.

3. **Simulate Data:**
   Use your chosen parameters and generated predictor variables to simulate the response variable y.

4. **Fit Different Models:**

   - **Model 1 (Correct Model):** Regress y on both x_1 and x_2.
   - **Model 2 (Underfit Model):** Regress y on x_1 only.
   - **Model 3 (Overfit Model):** Regress y on x_1, x_2, and an additional irrelevant variable x_3.

5. **Analyze Coefficient Estimates:**

   - Record the estimated coefficients $\hat{\beta}_1$ from each model.
   - Compare these estimates to the true β_1.

6. **Repeat Simulations:**
   Repeat the simulation multiple times (e.g., 10,000 iterations) to assess the variability and bias of the estimates.

7. **Evaluate the Impact:**

   - Calculate the bias and variance of $\hat{\beta}_1$ for each model.
   - Analyze how overfitting and underfitting influence these metrics.

3.2 Considerations

- **Choice of Parameters:**
  Selecting different values for β coefficients and error variance can affect your results. Consider experimenting with multiple sets of parameters.
- **Data Generation for Predictors:**
  Decide whether x_1 and x_2 are correlated. Correlation between predictors can impact multicollinearity and estimation.
- **Statistical Tests:**
  Consider using statistical tests or confidence intervals to assess whether the differences in estimates are significant.
- **Visualization:**
  Use plots (e.g., histograms, boxplots) to visualize the distribution of $\hat{\beta}_1$ across simulations for each model.