Project 4: Monte Carlo Simulation Proposal

Sanjana Sasmal

1. Scientific Question

I want to understand how sample size affects the reliability of my logistic regression model from Project 3. Specifically, I want to know how stable the coefficient estimates are and whether the model performs better or worse when trained on larger or smaller samples.

This is important because the original dataset was from a real survey, but if I wanted to apply the model to a new population or smaller subgroup, I would need to know how much confidence I can have in the predictions based on the sample size.

2. Data

We simulate data using a simplified logistic regression model that mimics the key predictors from Project 3. Each simulated person has three binary predictors:

- X_1: Changed behavior since Me Too (1 = Yes, 0 = No)
- X_2: Uses body language to gauge consent (1 = Yes, 0 = No)
- X_3: Asks for verbal confirmation of consent (1 = Yes, 0 = No)

The outcome variable Y is whether someone says they feel at greater risk of being falsely accused. It is generated using the following logistic model:

$$\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

We assume the following parameter values for the simulation:

- $\beta_0 = -1$
- $\beta_1 = 1.5$
- $\beta_2 = 1$
- $\beta_3 = 0$

These values reflect the idea that changed behavior and reading body language are positively associated with feeling at risk, while verbal confirmation is neutral.

3. Estimates

From each simulation, we estimate the logistic regression coefficients:

- $\hat{\beta}_1$: Estimated effect of changed behavior
- $\hat{\beta}_2$: Estimated effect of reading body language
- $\hat{\beta}_3$: Estimated effect of verbal consent

Across many repetitions, we summarize:

- **Bias**: $\text{Bias}(\hat{\beta}_j) = \mathbb{E}[\hat{\beta}_j] - \beta_j$
- **Variance**: $\text{Var}(\hat{\beta}_j)$
- **Mean Squared Error (MSE)**: $\text{MSE}(\hat{\beta}_j) = \text{Bias}^2 + \text{Variance}$

This tells us how well the model recovers the true effects from the simulated data.

4. Methods

The method being tested is logistic regression. For each simulated dataset, we:

1. Simulate binary outcomes using a logistic model with known coefficients.
2. Fit a logistic regression model to the simulated data.
3. Extract the estimated coefficients for each predictor.

We're testing how well this model can recover the true underlying effects used in the simulation. We're especially interested in the consistency of the estimates (variance), how close they are to the true values (bias), and overall performance (MSE).

We are not comparing multiple models, but focusing on whether logistic regression gives reliable estimates under the assumptions used in Project 3.

5. Performance Criteria

After running many simulations, we'll evaluate how well the logistic regression model performs using:

- **Bias**: The average difference between the estimated coefficient and the true coefficient used in simulation.
- **Variance**: The variability of the coefficient estimates across simulations.
- **Mean Squared Error (MSE)**: A combination of bias and variance, calculated as the average of (estimate - true value)^2 for each predictor.

These metrics will help us understand how accurate and stable the model is when applied to data similar to what we saw in Project 3.

6. Simulation Plan

We'll simulate data 1000 times using the same binary outcome and predictor structure as Project 3. Each simulation will follow these steps:

1. Generate a fake dataset of 1000 observations with the same binary predictors used in the original model.
2. Use fixed probabilities to assign values to each predictor, mimicking real patterns from the Project 3 dataset.
3. Simulate the outcome variable using the logistic model from Project 3, with chosen true coefficients.
4. Fit a logistic regression to the simulated data and extract the estimated coefficients.
5. Store the estimates across all simulations.

At the end, we'll have 1000 estimates for each coefficient, and we'll calculate bias, variance, and MSE for each.

This code will be a slightly modified version of the Project 3 model code, with a loop added to repeat the process and save results.

7. Anticipated Challenges or Limitations

One challenge is making sure the simulated data realistically reflects the patterns in the actual survey data. If we oversimplify the data-generating process, our results might not say much about real behavior.

Another issue is that with binary predictors and a binary outcome, small sample sizes could lead to unstable estimates in some simulations. We're using 1000 observations per simulation to help reduce that, but randomness will still cause some variation.

There's also a limitation in assuming the same coefficients apply across all simulations, when real-world relationships might shift depending on context. In future versions, we could explore what happens when those coefficients vary or when interaction effects are added.
