Scientific Question:

The question this experiment aims to answer is: How accurate is the model used by the dataset when predicting whether a text is written by a human vs. generated by Artificial Intelligence? Specifically, this experiment is designed to test how well the model performs when it is applied to synthetically generated data instead of the original dataset. This places emphasis on the "native" variable, which indicates (0 or 1) whether the author of the text (if human) is a native English speaker or not.

The modern day implications of this experiment and research question are profound, with AI generated text becoming more and more prevalent by the day. Tools that can reliably distinguish whether text is human or AI written will be invaluable, especially in educational and professional environments where a misclassification can be catastrophic. By simulating text data based on the original dataset (detectors.csv), the model's strengths and weaknesses become increasingly clear. Setting up a controlled experiment allows for testing of the model through the changing of parameters. The simulation allows for identification of bias, variance, and classification errors.

Data:

We simulate the data using a logistic regression model that follows the structure of the original dataset.

$$\text{logit}(P(Y=1 \mid X)) = \beta 0 + \beta 1 \cdot \text{native}$$

native ~ Bernoulli(p) where p = 0.6. This represents the proportion of native english speakers in the population.

β0 and β1 are fixed values.

Estimates:

The main quantities of interest in the simulation are the prediction accuracy, bias of coefficients, and the Mean Squared Error (MSE).

MSE = E[(p̂-p)^2]
P̂: predicted probability of AI generated text.

Methods

The experiment evaluates the logistic regression model by creating a new variable:
Detector Accuracy = | kind - predAI |

“Detector accuracy (0: wrong, 1: right) equals the absolute value of kind (0: human, 1: AI) minus predicted to be AI (0: human, 1: AI)”

We then factor in the “native” variable to see if there is a disproportionate amount of non-native text being classified as AI generated.

We will generate synthetic datasets and fit the logistic regression model to each dataset. Then, the parameter accuracy will be generated and recorded.

Performance Criteria

This experiment will use mean and SD of $\beta 0$ and $\beta 1$ hat, the variance and bias of coefficient estimators, and MSE of predicted probabilities.

Simulation Plan

Number of simulations: 1000
Sample size per simulation: 500
Parameters: $\beta 0$ and $\beta 1$ in sensitivity analysis
Recorded outputs: Coefficient estimates, predicted probabilities, classification labels, accuracy, MSE
Changes in the code: Replace the real dataset with the generated one

Anticipated Challenges

The real world data and the simulated data might not follow the same distributions. Furthermore, in my case specifically, this might take a long time to run per repetition because of reticulate errors and the lack of a conda environment. The estimates might also be noisy which masks true effects.