Project 4, Analysis Plan

Joseph Froelicher

December 5, 2021

Introduction

The purpose of this analysis is to perform simulation for two experimental cases, each with a single sub aim. The aims are as follows:

1. 250 Subjects with 20 variables in a linear regression model with five significant variables. The coefficients for these variables will range from 0.17 to 0.83.

Case A. All variables to be considered independent
Case B. Variables to be correlated

2. 500 subjects with 20 variables in a linear regression model with five significant linear predictors. The coefficients for these variables will range from 0.17 to 0.83.

Case A. All variables to be considered independent
Case B. Variables to be correlated

With the results of this simulation, we would like to compare the linear regression model selection methods by P-value, AIC, BIC, LASSO, and Elastic Net. These selection methods are going to be compared by their accuracy in selecting significant variables in the model selection. This will be measured by several statistical benchmarks. This includes the bias of the parameter estimates, the coverage of the 95% confidence interval, the false positive rates of the 15 insignificant simulated variables, and the true positive rate of each of the 5 significant variables individually.

Selection Statistics

Bias:

$$b_\beta = E[\frac{1}{M}\sum_{m=1}^{M}\beta_m] - \theta$$

$$\beta_m : \text{estimated coefficient}$$

$$\theta : \text{true parameters value}$$

Coverage:

$$c = \frac{r}{n}$$

$$r : \text{number of coefficient estimates that fall within the CI}$$

$$n : \text{number of simulated data sets}$$

$$CI = \theta \pm SE(\theta) * Z$$
$$SE(\theta) = \sqrt{E[\theta - E[\theta]]^2]}$$
$$Z = 1.959$$

True Positivity:

$$TPR = \frac{p}{s}$$
$$p : \text{number of times a variable is in the final model}$$
$$s : \text{number of simulated data sets}$$

False postivity:

$$FPR = \frac{q}{15s}$$
$$p : \text{number of insignificant predictors in the final model}$$
$$s : \text{number of simulated data sets}$$

Algorithm

```
@model_selction(data):
  run linear regression model
  select by p-values
    store coefficients with variable names
  select by AIC
    store coefficients with variable names
  select by BIC
    store coefficients with variable names
  select by LASSO
    store coefficients with variable names
  select by Elastic Net
    store coefficients with variable names

  return p-value_coeffs, aic_coeffs, bic_coeffs, lasso_coeffs, el_net_coeffs

@selection_statistics(coefficients)
  calculate bias
  calculate coverage
  calculate false negativity
  calculate true positivity

  return bias, coverage, false_negativity, true_positivity


for each aim:
  for each simulated iteration:
    if aim = 1 or 2 n = 250 else n = 500
    if aim = 1 or 3 rho = 0 else rho = correlation value
      create new data set based on conditions
    if variable is selected store the values else store a 0
    initializae data frame to store coefficients (20 by number of iterations)
    data_frame[iteration] = @model_selection(new_data_set)

for each selection method:
  bias, coverage, false_negativity, true_positivity = @selection_statistics(data_frame)
```