Bootstrap Methods In Case of Categorical Response Variable

Bharat Jambhulkar

Department of Statistics,
Savitribai Phule Pune University

Introduction

Suppose $(x_i, y_i),\ i = 1, \dots, n$ is a bivariate sample on variables x and y such that $y_i,\ i = 1, \dots, n$ is a binary variable and $x_i,\ i = 1, \dots, n$ can be continuous or categorical. Suppose y is a response variable and x is a regressor. The aim is to construct a bootstrap sample for this data.

Non-Parametric Method

Suppose $O_i = (x_i, y_i),\ i = 1, \dots, n$.

1. Fix B as a large number.
2. From the original sample $O_i,\ i = 1, \dots, n$, draw a random sample with replacement of size n until B bootstrap samples are generated.
3. Compute $\theta_1^*, \theta_2^*, \dots, \theta_B^*$ based on each of these B bootstrap samples, where:
$$\theta_i^* = (\hat{\beta}_{0i}^*, \hat{\beta}_{1i}^*)$$
4. The set $\{\theta_1^*, \theta_2^*, \dots, \theta_B^*\}$ forms a sample of size B from the distribution of $\hat{\theta}^*$.

Note that the ordered pair $(x_i, y_i),\ i = 1, \dots, n$ should remain unchanged. Why is this method called non-parametric? The obvious reason is that we have not specified or assumed any model to obtain the bootstrap sample.

Example: Iris Data Set

The Iris dataset is one of the most well-known and widely used datasets in statistical analysis. It consists of 150 samples of iris flowers from three species: *Iris setosa*, *Iris versicolor*, and *Iris virginica*. Each sample includes four features: sepal length, sepal width, petal length, and petal width. The target variable represents the species of the iris flower and has three classes. In this exercise, only two species *Iris setosa* and *Iris versicolor* are considered. Logistic regression is performed using sepal width as the single predictor to classify the species. The exercise involves generating 2000 bootstrap samples, estimating the regression coefficients from each sample, and constructing confidence intervals for the estimated coefficients.

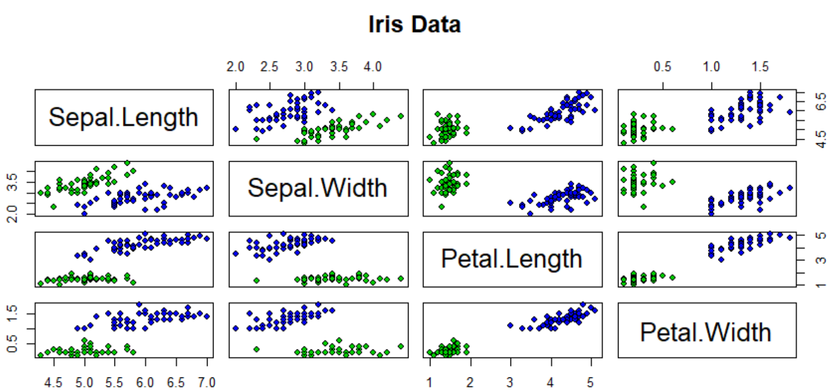

fig: Pairs Plot

Following the algorithm described in the non-parametric methods section, generate a bootstrap sample and obtain the estimates of

$$\theta_i^* = (\hat{\beta}_{0i}^*, \hat{\beta}_{1i}^*)$$

$i = 1, \ldots, B$.
The histogram below displays the distribution of estimates of the regression coefficients.

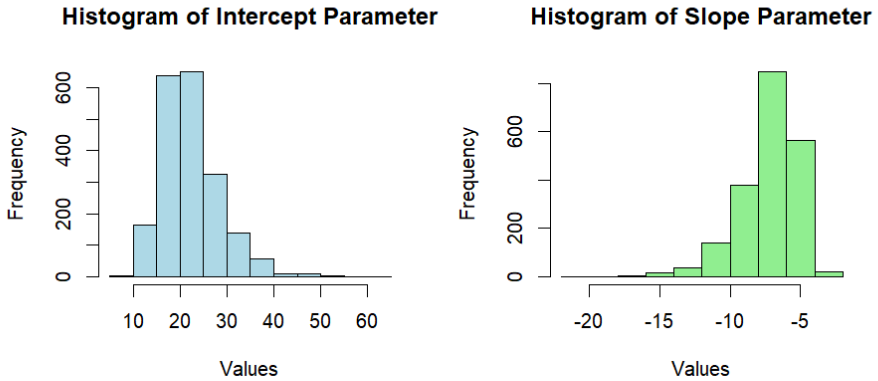

fig: Histogram of Parameter Estimates

The quantile-based and normal distribution approximation-based confidence intervals are given as -
95% quantile-based confidence interval: for the intercept (13.1281, 37.0507), and for slope (-12.0685, -4.3066).
95% normal-approximation based confidence interval: for the intercept (9.9929, 34.6119), and for slope (-11.2141, -3.2392).

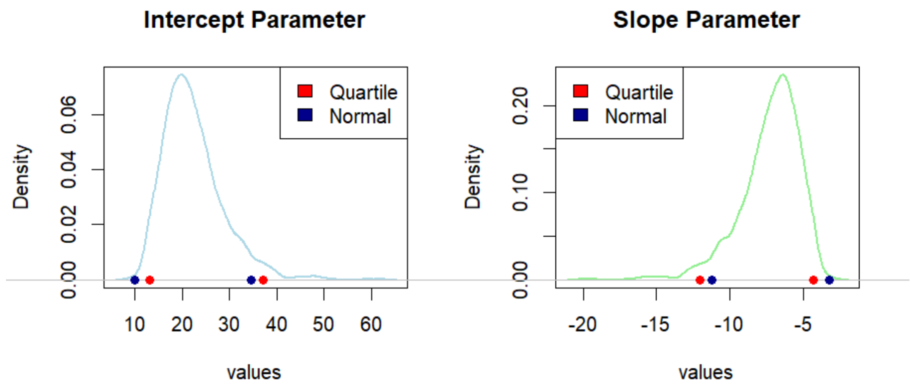

fig: Density Plot of Parameter Estimates

Semi-Parametric Method

Suppose $O_i = (x_i, y_i),\ i = 1, \dots, n$.

1. Fix B as a large number.

2. Fit the logistic regression model to the original sample $O_i, i = 1 \dots, n$.

3. Obtain the fitted values and classify the observations based on a threshold.

4. Compute the residuals as:
$$e_i = y_i - \hat{y}_i$$
where e_i represents the residual for the ith observation, y_i is the observed value, and $\hat{y}_i$ is the predicted class.

5. Draw a sample of size n with replacement from the residuals. Denote the sampled residuals as:
$$e_1^*, e_2^*, \dots, e_n^*$$

6. Calculate the new response value as:
$$y_i^* = \hat{y}_i + e_i^*;$$
$i = 1, \dots, n$.

Note that, in logistic regression, since we are classifying observations into two categories, residuals will take values 0, 1, or -1. The values $\{1, -1\}$ indicate misclassification. Also, $\hat{y}_i$ will be either 0 or 1. It is possible that y_i^* takes values less than 0 or greater than 1. In this situation, y_i^* is bounded as follows: if $y_i^* < 0$, then $y_i^* = 0$; and if $y_i^* > 1$, then $y_i^* = 1$. Why is this a semi-parametric method? In the second step, we use a model to predict the classes and obtain residuals from the predictions. While constructing y_i^*, we again use an equation. These two things represent the parametric part. After obtaining the residuals, we use random sampling with replacement on the residuals. We use predicted classes and sampled residuals to construct bootstrap observations, which is the non-parametric part.

Example: Iris Data Set

The data description remains the same as above. Using the algorithm explained in the semi-parametric method section, generate 2000 bootstrap samples, estimate the logistic regression model coefficients, and construct confidence intervals for the estimated coefficients.
The histogram below displays the distribution of estimates of the regression coefficients.

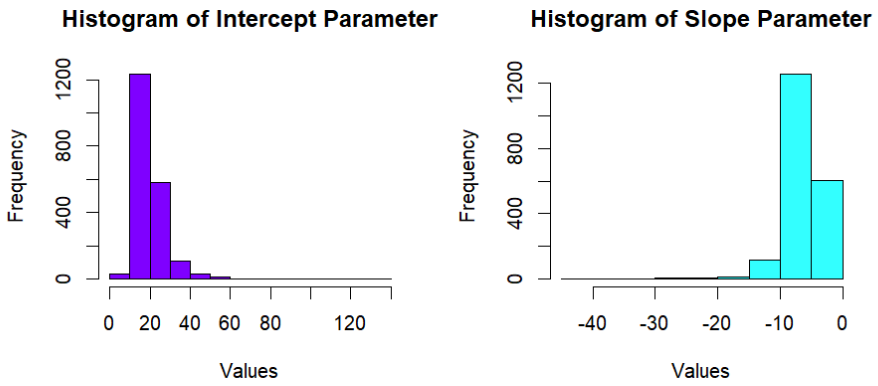

fig: Histogram of Parameter Estimates

The quantile-based and normal distribution approximation-based confidence intervals are given as -
95% quantile-based confidence interval: for the intercept (10.5238, 38.5474), and for slope (-12.4273, -3.4279).
95% normal-approximation based confidence interval: for the intercept (3.7426, 35.6117), and for slope (-11.5752, -1.1822).

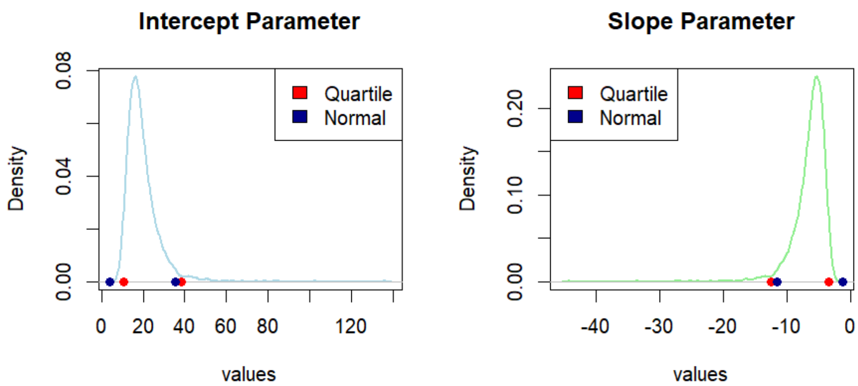

fig: Density Plot of Parameter Estimates

Assignment Challenges

- As discussed in the case of a continuous response variable, adding variation to the response variable is straightforward since the error is symmetrically distributed around zero. However, in the case of logistic regression, where the loss function is binary, introducing variation into the response variable using the semi-parametric method is challenging.