Bootstrap Applications in Model Validation for Logistics Regression

STAT573 Fall 2020

Haeyoon Chang
Computer Science Department
Portland State University
Portland, OR USA
haeyoon@pdx.edu

***Abstract*—Bootstrapping technique is applied to validate the stability of logistics regression coefficients estimates and to validate the model predictive accuracy rate. Each of coefficient estimates were checked against the reference distribution generated from bootstrapping samples. Also, the bootstrap method detected optimism bias in predictive accuracy and the corresponding bias was corrected to report true accuracy rate. The train and test data split method and cross validation method also detected optimism bias, and the amount of bias from these two methods were very close to the one detected from the bootstrap method.**

***Index Terms*—Bootstrap, model validation, cross validation**

I. Introduction

Model validation and performance validation are important steps when building a model to solve the given set of problems. Model validity can be judged by the stability and reasonableness of the coefficients and ability to generalize to the new data set. There were two goals for this project. The first goal was to judge whether the logistic regression coefficients are reasonable using the bootstrap simulation. The second is to validate the performance of the model using the bootstrap method. This project demonstrates how to validate the regression coefficients estimates and predictive accuracy using bootstrap method with the data on heart diseases patients. Bootstrap methods are also compared against two other validation methods - train and test data split and cross validation.

II. Data

Heart diseases patients data was from UCI Machine Learning Repository and the original data set were contributed by the four different hospitals. The data contains a total of 303 patient information with 14 attributes including the label column indicating the presence of heart disease in the patient. The label is integer valued from 0 (no presence) to 1. This data set was balanced as 165 patients had heart diseases and 138 patients did not. The 13 predictors are as follows:

- age
- sex
- chest pain type
- resting blood pressure
- serum cholestoral in mg/dl
- fasting blood sugar > 120 mg/dl
- resting electrocardiographic results
- maximum heart rate achieved
- exercise induced angina
- ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels colored by flourosopy
- thal

III. Methodology

For this project, logistic regression was used as main model. Logistic regression is a statistical model that uses a logistic function to model a binary dependent variable. In this project, logistic regression model estimated the parameters of 13 attributes or predictors, and binary values were assigned based on probability calculated based on the estimated parameters and logistic function.

A. Stability of Regression Coefficients using Bootstrap

As part of model validation, bootstrapping technique was used to construct the reference distribution of each of seven coefficients. This is useful for situation where one should infer the p-value or confidence interval of the estimated coefficients without making any assumptions about the underlying distribution of these statistics. Based on the reference distribution from bootstrapping, I tested the stability of regression coefficients estimates. Next step was to calculate the prediction accuracy based on the coefficients estimates.

B. Validation of Model Prediction Accuracy

Since there were no more data set available outside the 303 patients information, three common techniques were used to validate the model prediction accuracy. First method was to use the bootstrap technique to adjust the accuracy rate to get an idea of its hypothetical performance in predicting the outcomes for additional data sampled from a similar population. Second method was to divide the data into two groups - train and test groups. Train data were used to estimate the regression coefficients and then the model were validated using the test data set. The smaller the accuracy rate gap between train

and test data, the more confident that the model generalizes. Third, k-fold cross validation were used to validate the model accuracy. Cross validation is a method to split the data into k groups. The method iterates the data K times and in each iteration, one group becomes the test set and the remaining groups were used to train the model. Cross validation sounded like the combination of bootstrap and test-train split methods. However, the major difference between bootstrap and cross validation was that bootstrap method allows replacement while cross validation does not.

IV. Results

Initially, all 13 variables were included in the model to fit a simple logistic regression that estimated the linear effect of the logit of those variables on the probability of having heart diseases or not. Out of the 14 variables, the following seven variables were statistically significant at 0.1 level.

- sex
- chest pain type
- resting blood pressure
- maximum heart rate achieved
- exercise induced angina
- number of major vessels colored by flourosopy
- thal

The coefficients of these seven variables were statistically significant at 0.05 level when the model excluded variables that were not statically significant. See "Fig. 1". Experiments going forward used the following seven predictors only because it yielded higher accuracy rate for new data and for interpretation.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.454986   1.797750   0.809 0.418321
sex1        -1.508238   0.401857  -3.753 0.000175 ***
cp1          1.503752   0.518385   2.901 0.003722 **
cp2          1.895602   0.421492   4.497 6.88e-06 ***
cp3          1.654810   0.584799   2.830 0.004659 **
trestbps    -0.021894   0.009251  -2.367 0.017947 *
thalach      0.031485   0.008808   3.575 0.000351 ***
exang1      -1.045425   0.388873  -2.688 0.007181 **
ca          -0.745972   0.170593  -4.373 1.23e-05 ***
thal        -0.898006   0.267527  -3.357 0.000789 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig. 1. Logistic regression coefficients estimates for seven variables of choice

A. Stability of Regression Coefficients using Bootstrap

To validate the model's assumptions about the shape of the estimated parameters' distributions, I constructed reference distributions for each of seven estimated coefficients. For example, Fig. 2 showed the reference distribution of estimated coefficient of resting blood pressure per second. Null hypothesis was that the resting blood pressure coefficient in Fig. 1. was the same as the mean of the reference distribution. The

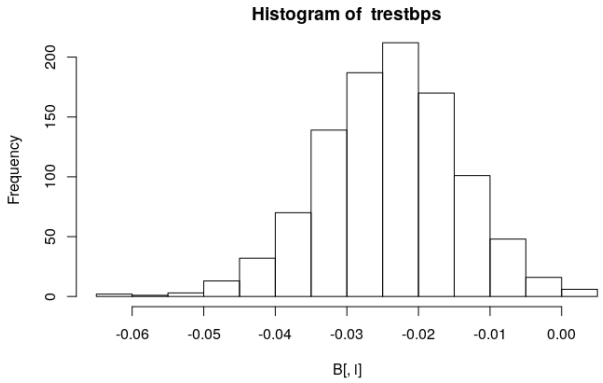

Fig. 2. Reference distribution for tresbps coefficient distribution using bootstrap method

null hypothesis was not rejected because p-value of -0.02 was 0.428. Therefore, -0.02 seemed a stable estimate.

Same experiments were done for all seven coefficients, and each of coefficients hypothesis were not rejected as their p-value were not small enough at 0.1, 0.05, or 0.01 level. See Table I.

TABLE I
P-value of Coefficients estimates

	p-value
intercept	0.527
sex1	0.43
cp1	0.55
cp2	0.584
cp3	0.544
trestbps	0.428
thalach	0.56
exang	0.464
ca	0.443
thal	0.45

B. Prediction Accuracy

The Receiver Operating Characteristic (ROC) curve for the model were plotted in Fig. 3. The Area under the ROC curve (AUC) is widely-used measure of accuracy for many classification models. The closer AUC is to 1, the higher the accuracy. The baseline of the model accuracy would be AUC = 0.5 because at this point it indicates the model is as accurate as guessing at random with probability of 0.5. Based on the coefficients estimates, the accuracy rate of the logistic regression model on heart diseases prediction was 0.908.

C. Validation of Model Prediction Accuracy

It was natural that the logistic regression model's accuracy rate was high because the predictions were made on the same data set that the model were trained on. To get more realistic measure of the model's predictive accuracy, three different methods were used to detect whether there is any bias in prediction accuracy - bootstrap method, train-test split, and cross validation.

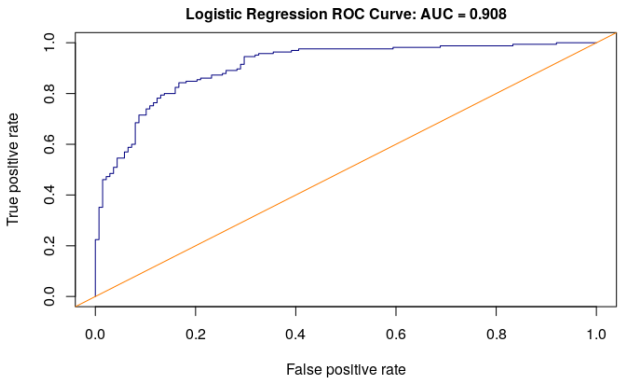

Fig. 3. ROC curve on prediction of all 303 data points

1) Boostrapping Method: In order to calculate the optimism bias, the average accuracy rate derived from original data and that from bootstrap samples data were compared. First, I generated 100 bootstrap samples from the 303 patients data, and estimated coefficients based on each bootstrap sample in each iteration. Then, I applied the fitted model to the original data to give 100 estimates AUC. Also, I applied the fitted model to the 100 bootstrap samples to obtain another set of 100 estimates AUC on bootstrap samples. The average AUC on original sample was 0.901 where as the average AUC on bootstrap samples was 0.914. The difference between two accuracy rate is the optimism bias, which in this case was 0.013. The optimism bias is the amount which the testing against already trained model overestimates the true prediction accuracy. That is, it would be reasonable to deduct 0.013 from the accuracy calculated above (0.908), leading to estimated "true" accuracy rate of 0.896.

2) Train-Test Data Split: The original data was split into two groups - test and train data. The 80 percent of the data were picked as train data and the model was fitted into the train data. The remaining 20 percent was later used to validate that the model performed well for the data it has not seen before. The AUC for train and test data were 0.909 and 0.898, respectively. See the Fig 4 and Fig 5.

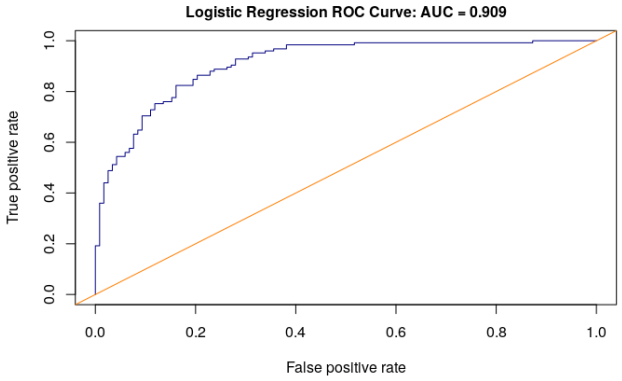

Fig. 4. ROC curve on prediction of 243 data points

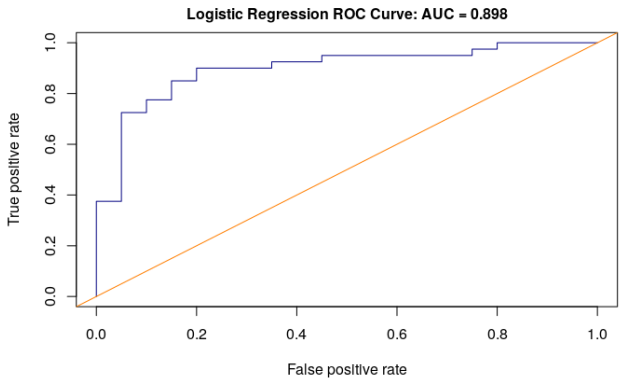

Fig. 5. ROC curve on prediction of 60 data points

3) Cross Validation: Cross validation refers to a set of methods for measuring the performance of the predictive model on new test data set. This is expansion of what I did in test-train split above and also similar to bootstrap method, but train and test data were selected without replacement. For heart diseases data, I assigned the data points evenly into 5 folds. For each iteration one block is set aside as test data while the remaining four block were used to train the model. In general, the prediction accuracy was slightly higher for train data than test data except for Block 1. See Fig II. On average, the AUC for train data and test data using cross validation were 0.91 and 0.896, respectively.

TABLE II
AUC FOR TRAIN AND TEST DATA
CORRESPONDING BLOCK WAS TEST AND
THE REMAINING WAS TRAIN SET

	AUC train	**AUC test**
Block 1	0.894	0.962
Block 2	0.928	0.829
Block 3	0.915	0.871
Block 4	0.907	0.898
Block 5	0.907	0.922

D. Comparison of Three Validation Methods

Interestingly, all three validation methods showed similar results, supporting that there is about 0.011 – 0.15 optimism bias and bias correction was needed to estimate "true" accuracy rate. Overall, "true" accuracy rate of this experiment would be around 0.896 – 0.898. One caveat is that all three validation methods use the same 303 patients data in the end and the performance was never tested against truly new data.

V. CONCLUSION

Through this project, it is demonstrated that the bootstrap method is useful tool to validate the model and its accuracy rate along with other model validation techniques. In the abundance of data and increasing needs for analysis, number of estimation and classification methods such as logistic regression are used. In this process, model validation and confidence

in model accuracy rate are crucial parts and the bootstrap can be one of tools or solutions to effectively validate the results.

REFERENCES

[1] Efron, B., & Tibshirani, R. (1993). "An Introduction to the Bootstrap". Chapman and Hall.

[2] https://rpubs.com/vadimus/bootstrap

[3] Lecture notes from STAT573 Computation Intensive Methods in Statistics