SESUG Paper 191-2019

Bootstrapping Regression Models using PROC SURVEYSELECT

Bryce Whitehead, M.S., University of Northern Colorado, Austin Brown, Ph.D., Kennesaw State University

ABSTRACT

When constructing regression models, it is commonplace for a researcher to be interested in assessing the relationship between categorical predictor variables and some response variable. It is common for a categorical predictor to be dichotomous and coded as “0” or “1” in the dataset. However, when a large proportion of the observations fall into either category (i.e., greater than 80% or 90%) parameter estimation can become unreliable as the standard error of the estimator may become either inflated or deflated. Such a data situation may occur in observational types of analyses. One way of addressing this concern could be through taking a random sample from the larger group to match the sample size of the smaller group and then fitting the desired model. To efficiently use the total sample, this procedure could be performed multiple times using a bootstrapping technique. Here, several models are fit and the means of parameter estimates along with their standard errors are taken to be the final estimators. Using an example dataset containing final letter grades of domestic and international introductory statistics students over the course of several semesters, the aforementioned bootstrapping procedure will be demonstrated for a logistic regression using PROC SURVEYSELECT in SAS®. Sampling techniques and assessing model fit will also be discussed.

INTRODUCTION

One common research question in a variety of fields is that of group comparison. For example, educational administrators may be interested in determining if students in an undergraduate introductory statistics course for whom English is a second language (i.e., ESL students) perform comparably to native English speakers. By doing this, administrators could ascertain if additional resources need to be given to ESL students as well as instructors to maximize the opportunity for student success. In such a case, a simple way of assessing the research question could be collecting historical data on both ESL and native English-speaking students, possibly controlling for other covariates such as academic standing (e.g., Freshman, Sophomore, Junior, Senior), academic term (Fall or Spring), and academic year (2014, 2015, etc.). The researchers are primarily interested in determining if ESL students and native English-speaking students have different rates of completing the course successfully, which was operationalized to mean earning a “B” grade or better. The nature of the available data combined with the research question at hand lends itself to a logistic regression analysis.

Generally, it is desirable to have similar sample sizes with respect to the levels of a categorical predictor in a logistic regression model. The reason for this is because it is hoped that there are several observations from these levels which overlap in the response. In the example given, it is desired to have multiple ESL students successfully completing and not successfully completing the course as well as having multiple native English-speaking students having the same characteristics. The reason for this is because of a phenomenon in logistic regression model fitting referred to as “perfect separation,” which is a special case of the overfitting problem in general model fitting. Perfect separation implies that all members of one level of a categorical predictor fall into one category of the binary response. In this case, a logistic regression model would suffer from perfect separation if all ESL students successfully passed the course and none of the native English-speaking students did or vice versa. Certainly, it is unlikely these results can be generalized to the larger populations. However, problems with model fit may still arise in the situation when very few observations overlap. For example, a dataset could be comprised of 95 observations from one group and only 5 from the other. The nature of the dataset certainly suggests that direct comparison with any practical meaning would be limited with such group disparity. Comparing the two groups using a logistic regression model, while computationally possible, would not be advisable as the power of the test would be low. Of course, in many situations this is avoidable with good study design. However, in some situations it may not be avoidable, as would be the case when working with secondary data. The question then arises: if it is unavoidable to have groups of vastly differing sizes, how can comparisons be more reasonably be made?

BOOTSTRAPPING

One potential solution to the aforementioned problem of comparison could be to take a sample from the larger group of a comparable size to that of the smaller group and build a model using the smaller subset. Now, the issue of lower statistical power still persists, but the comparisons would be more “fair” in terms of estimated variance. However, in performing this a single time, the bulk of the observations from the larger group are being ignored and is a clear misuse of valuable information. Additionally, the observations sampled from the larger group, because of random sampling, may not necessarily be representative of the larger sample. In the teaching method study example, the university from which these data were gathered is a predominately white institution, which means that the number of ESL students with respect to the whole population is quite small. Thus, if a small sample from the native English-speaking population is taken to make a comparison, and the students randomly sampled just so happened to be excellent students who performed exceptionally well while the ESL students had middling performance, then the researchers may unfairly conclude that ESL students are not as well equipped to be successful in this particular course. Conversely, the opposite conclusion may also be made.

A possible remedy in this scenario is through resampling. Instead of taking a single sample from the larger group to make the comparison or fit the model, several samples are iteratively taken, several models are fit, and the mean values of the estimated parameters are taken to be the “true” fitted logistic regression model. This technique is commonly referred to as “boostrapping” (Rizzo, 2007). To reiterate, boostrapping in this case will not solve the problem of reduced statistical power, but it does allow more fair comparisons and more appropriate conclusions with respect to inherent uncertainty to be made. With all of this in mind, how can this be done in SAS? While there are a variety of ways this could be done, one such way is using PROC SURVEYSELECT.

USING PROC SURVEYSELECT BY EXAMPLE

After importing and cleaning the data so that it is ready for analysis, it was decided that the sample sizes for ESL and native English-speaking students should be equal for each term (e.g., Fall 2014). In order to do this, the larger dataset was subset into smaller year and term datasets, and then further split into ESL and native English-speaking students. As an example:

```
/* Spring 2014 */


data AY_S2014;
set students4;
if Year ne 2014 then delete;
run;

data AY_S2014;
set AY_S2014;
if Term1 ne 'Spring' then delete;
run;


/* Subset Native Speakers & ESL Speakers */


data AY_S2014_Native;
set AY_S2014;
if esl = 'Y' then delete;
```

```
run;


data AY_S2014_ESL;
set AY_S2014;
if esl = 'N' then delete;
run;
```

Then, it can be determined using PROC FREQ how many ESL students took the course in the particular year and term of interest:

```
/* Determine Number of ESL Students */


proc freq data=AY_S2014;
table esl;
run;
```

This yields Figure 1.

esl	Frequency	Percent	Cumulative Frequency	Cumulative Percent
N	437	97.76	437	97.76
Y	10	2.24	447	100.00

Figure 1. Results of PROC FREQ

Now, we know that in Spring 2014, there were 10 ESL students. This means, using the bootstrapping technique outlined in the previous section, simple random samples can be taken from the native English-speaking group using PROC SURVEYSELECT.

PROC SURVEYSELECT is a robust tool which has a variety of options and uses. Only some will be outlined through this example, and the reader is directed to the SAS documentation for other functionality not described here. Before getting into the SAS syntax, one practical question may arise: How many samples is enough samples to appropriately describe the relationship. Of course, the answer is that it depends on a number of factors, but in general, more is better. In this example, 10,000 samples were taken using the below code:

```
/* Obtain 10000 Simple Random Samples of Native Speakers of Size n = 10 */
/* Using PROC SURVEYSELECT */

proc surveyselect data=AY_S2014_Native
method = srs n = 10 out = AY_S2014_SRS reps=10000;
run;
```

Going through the code, after calling the procedure and specifying the dataset to be sampled, there are a few options used here. The first is specification of the sampling method to be used. Here, simple random sampling without replacement as denoted by `method = srs`. However, there are a variety of sampling techniques available including probability proportional to size techniques, systematic random sampling, and stratified random sampling. Next, the size of the sample must be specified, and in this case 10 observations are desired to match the number of ESL students for this year and term. The third option in the second line is the outputted dataset. This is followed by the number of samples to take, which in this case is 10,000.

After this step, and in order to fit 10,000 logistic regression models, the ESL student dataset must be replicated 10,000 times for the BY step in PROC LOGISTIC to work as desired. To do so, the following code in the DATA step was used:

```
/* Replicate ESL Dataset 10000 Times */

data ESL_S2014_Dup(drop=i);
do i = 1 to 10000;
do j = 1 to n;
set Ay_S2014_esl nobs=n point=j;
output; end; end; stop;
run;
```

Now, the two datasets, `AY_S2014_SRS` and `ESL_S2014_Dup` need to be merged. However, these two datasets need to be merged in a way that each sample from the bootstrapped dataset is matched with one of the replicated datasets of ESL students. The convenient aspect of outputting the dataset in PROC SURVEYSELECT is that it creates a new variable which specifies which iteration or replicate of the bootstrap this particular set of observations came from, as shown by Figure 2.

Obs	Replicate
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	2
12	2
13	2
14	2
15	2
16	2
17	2
18	2
19	2
20	2

Figure 2. Output from Bootstrapped Dataset

However, in order to effectively utilize this convenient aspect of PROC SURVEYSELECT, the observations of the duplicated ESL student dataset need to be appended with replicate numbers as well. Since the bootstrapped dataset and the duplicated dataset are the same dimension, the replicate variable in the bootstrapped dataset can simply be isolated and bound to the duplicated dataset by:

```
/* Separate Out Replicate Number Variable */

data rep_S2014;
set Ay_S2014_srs;
keep Replicate;
run;
/* Add Replicate Variable to ESL Dataset */

data ESL_S2014_Dup;
merge ESL_S2014_Dup rep_S2014;
run;
```

Now, the bootstrapped dataset and the duplicated dataset can be simply “stacked” on top of each other to form one large dataset for the whole year and term. In the case of Spring 2014:

```
/* Merge ESL & Native Bootstrapped Datasets */

data Spring_2014;
set Ay_S2014_srs ESL_S2014_Dup;
run;
```

This procedure was performed for Fall 2013 through Spring 2018 terms. Then, all of the merged datasets from each year and term were merged together into one large dataset:

```
/* Merge All Bootstrapped Terms */

data big_file;
set Fall_2013 Spring_2014 Fall_2014 Spring_2015
      Fall_2015 Spring_2016 Fall_2016 Spring_2017
      Fall_2017 Spring_2018;
run;
```

FITTING LOGISTIC REGRESSION MODELS USING PROC LOGISTIC

After merging all the datasets together for analysis, now the logistic regression models can be fit. Conveniently, one of the procedures with which a logistic regression model can be fit, PROC LOGISTIC, has the option of fitting multiple models across one of the levels of a particular variable in the BY statement. Thus, after sorting the replicate variable the logistic regression model can by fit:

```
/* Sort Replicate Number */

proc sort data=big_file;
by Replicate;
run;

/* Run Logistic Regression BY Replication Number */

proc logistic data=big_file noprint outest=big_mod;
by Replicate;
class ESL Student__Classification Term1;
model Final_Grade(Event='1') = ESL Student__Classification Year Term1 ;
run;
```

The above code will fit 10,000 logistic regression models and will output the parameter estimates to the dataset named `big_mod`. Now, using PROC MEANS, coefficient estimates and their standard errors can be estimated, the results of which are given by Figure 3:

```
/* Determining Bootstrapped Coefficient Estimates & Standard Errors */

proc means data=big_mod mean stderr;
var intercept eslN Student__ClassificationFreshman
Student__ClassificationJunior Student__ClassificationSenior
      term1Fall year;
run;
```

Variable	Label	Mean	Std Error
Intercept	Intercept: Final_Grade=0	328.8558037	1.6928133
eslN	esl N	0.1644031	0.0012550
Student__ClassificationFreshman	Student Classification Freshman	-0.5168739	0.0019340
Student__ClassificationJunior	Student Classification Junior	-0.1046786	0.0016133
Student__ClassificationSenior	Student Classification Senior	0.7966417	0.0017861
term1Fall	term1 Fall	-0.0413496	0.0012045
year		-0.1628531	0.000839812

Figure 3. Bootstrap Coefficient Estimates and Their Respective Standard Errors

CONCLUSION

To conclude, in this paper, an example of how to and when to perform a bootstrapped logistic regression was shown using PROC SURVEYSELECT. In this case, the primary purpose of performing the bootstrap was to alleviate concerns arising from the large sample size discrepancy between the ESL group and the native English-speaking group. However, there may be other circumstances when a similar tactic could be employed. For example, imagine a large national retailer wanted to examine and compare consumer behavior trends for some subset of their product offerings using a traditional statistical technique, such as ANOVA. If this retailer has hundreds of thousands of unique customers in each of the comparison groups, an ANOVA model using all available data would be overpowered and all differences, however slight, would be significant. Thus, it may be better to bootstrap many ANOVA models using smaller (although not too small) sample sizes. However, as is the case with any statistical modelling, there are limitations to bootstrapping analysis which should be noted and considered alongside the results it produces.

REFERENCES

Rizzo, M. L. (2007). *Statistical computing with R*. Boca Raton, FL: Chapman & Hall/CRC.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Bryce Whitehead, M.S.
University of Northern Colorado
Bryce.whitehead@unco.edu

Austin Brown, Ph.D.
Kennesaw State University
Abrow708@kennesaw.edu