Fall 2013, Volume 3, Issue 3

THE STATSWHISPERER

The StatsWhisperer Newsletter is published by staff at StatsWhisperer.™ For many more free resources in learning statistics, including webinars and subscribing to this newsletter, visit us on the web at: www.statswhisperer.com

Bootstrapping: It's Not Just for Footwear Anymore

Ah yes, I am sure there was a time for all of us when our idea of bootstrapping was buying some new footwear and fastening them to our feet. However, now that we have entered the world of statistics the term has a double meaning. Don't worry, you can still keep that footwear that suggests "I'm with the band" or "All my exes live in Texas," but now you will also have an incredibly useful statistical procedure to go along with them.

How useful is this procedure? Well, the term "bootstrapping" actually refers to the utility of the technique. Specifically, it has been suggested the term is derivative of the phrase, "To lift yourself up by your bootstraps." This phrase suggests that one is doing the seemingly impossible, as it is unlikely that you will be able to lift yourself into the air by tugging at the straps attached to your boots. Similarly, the use of the bootstrapping technique often seems that you are being afforded a seemingly impossible benefit, which we will discuss in this newsletter issue.

The focus of this newsletter has always been to suggest answers to challenges encountered in data analysis, and bootstrapping is a key solution to many.

INSIDE THIS ISSUE

What is Bootstrapping in Statistics?

Before describing the bootstrap technique, let's first discuss a common problem in statistics that bootstrapping may be used to address. Specifically, most researchers are eventually tasked with performing an inferential statistical analysis using a relatively small sample of study participants. A small sample often presents several challenges.

For example, many times when using a somewhat small sample many of the assumptions of parametric testing, such as a normal distribution of scores, are not met. Additionally, widely used methods of treating the data in response to these challenges, such as data transformations (e.g., log odds), may not be effective. Subsequently, the researcher is left in a situation where the analysis must be done, but the data do not seem to meet the needs of the analysis. At this point the researcher ponders how to use this somewhat small sample, which does not meet the assumptions of the parametric tests he or she wished to use, to produce an effective analysis.

What is Bootstrapping in Statistics? (Continued)

In this situation, bootstrapping may provide a viable solution to the challenge the researcher is facing. Specifically, when a sample is not large enough to facilitate a straightforward statistical inference, bootstrapping provides a means of accounting for the distortions caused by a somewhat small sample. Specifically, bootstrapping is a statistical procedure used to estimate statistical perimeters by sampling with replacement from the original sample. The purpose of bootstrapping is very often to derive robust estimates of statistical values such as standard errors and confidence intervals regarding a population parameter, such as the mean, median, odds ratio, correlation coefficient, or regression coefficient. However, as suggested above, bootstrapping is also used as a robust alternative to straightforward inference based on parametric assumptions when those assumptions are dubious.

How does the bootstrapping procedure work?

Essentially, the bootstrap method takes a sample of your data (e.g., your sample of study participants), then another sample, then another sample, and so on up to thousands of times, to empirically estimate the statistical values (see Efron & Tibshirani,1993 and Hesterberg et. al., 2005). Not surprisingly, bootstrapping is from a family of statistical tests known as *resampling* methods. The idea behind bootstrap is to use the data available to you via your sample as a "surrogate population", for the purpose of approximating the sampling distribution of a statistic. In other words, the bootstrap is resampling with replacement from the sample you do have to create a significant number of "phantom samples" that we call bootstrap samples. The software then computes a sample summary based on each of the bootstrap samples.

For example, suppose you are interested in the average (i.e., the mean) weight of people in the United States. Of course, it might not be possible for you to weight every person in the country to uncover this number. Subsequently, in true research form you instead sample a small cross-section of the country to obtain an estimate. Subsequently, let's assume you obtain a sample size of 1,000 study participants, which of course would produce only one mean value of weight. However, in order to draw estimates about the entire population of the United States, you will need a sense of the variability in the mean weight score computed. The bootstrap method will take the original 1,000 weight values in your dataset and via your software program will sample from that distribution to form a new sample, which is known as the bootstrap sample or re-sample.

The bootstrap sample is derived from the original sample of 1,000 weights using sampling and replacement, so each bootstrap sample is not identical to the original sample. The bootstrap sample process is repeated a significant amount of times (usually varying from 1,000 or 10,000 times based on various factors). Each bootstrap sample will incorporate a mean value score, which would be known as a bootstrap estimate.

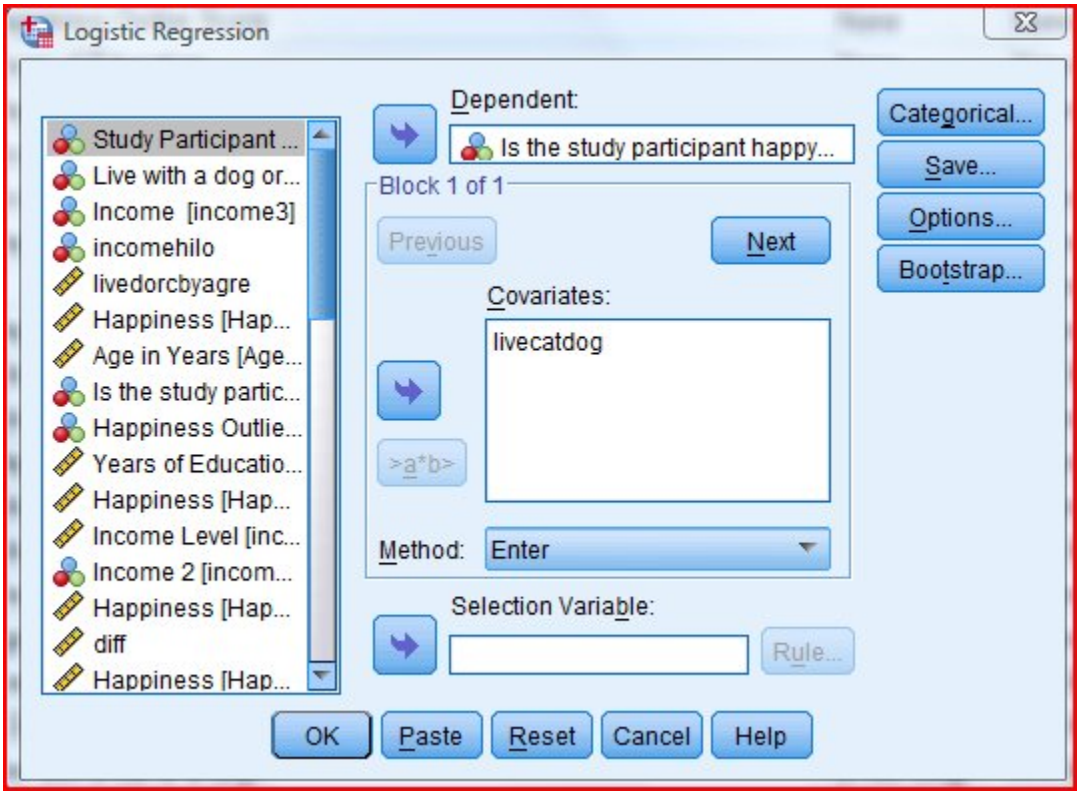
Based upon the distribution of mean value scores within each bootstrap sample, you would then have a larger distribution of mean value scores based upon the multiple bootstrap means. These values would then provide an estimate of the shape of the distribution of mean scores from which you can now answer questions regarding how much the mean score regarding weight within your sample varies.

How Can Bootstrapping Methods Be Implemented?

Here is even more good news! You do not need specialized statistical software to conduct the bootstrapping method. Specifically, in recent years the bootstrap method has been made available in commonly used statistical software, including SPSS. Thus, this method is quite accessible to many people. Below, to illustrate the bootstrap method, we have employed the technique in binary logistic regression.

To employ bootstrapping in binary logistic regression using SPSS, we would first go to: **Analyze→ Regression→ Binary Logistic**. A dialogue box will open that looks similar to the image below. You will see in the image below the dependent variable we have entered (within the box labeled Dependent) is the variable *Is the study participant happy* (1=Yes, 0=No). The independent variable (within the box labeled Covariates) is the variable *Does the study participant live with a cat or a dog?* (0=cat, 1=dog).

Before we conduct this analysis, there are two boxes we need to check. First, click the box in the upper right hand corner labeled *Options*. Within the dialogue box the opens, click the box next to the term *CI for Exp(B)*, which will produce the 95% confidence interval for the odds ratio estimate. This is the range in which we are 95% confident the true odds ratio within the population lies. Then click the button marked *Continue*.

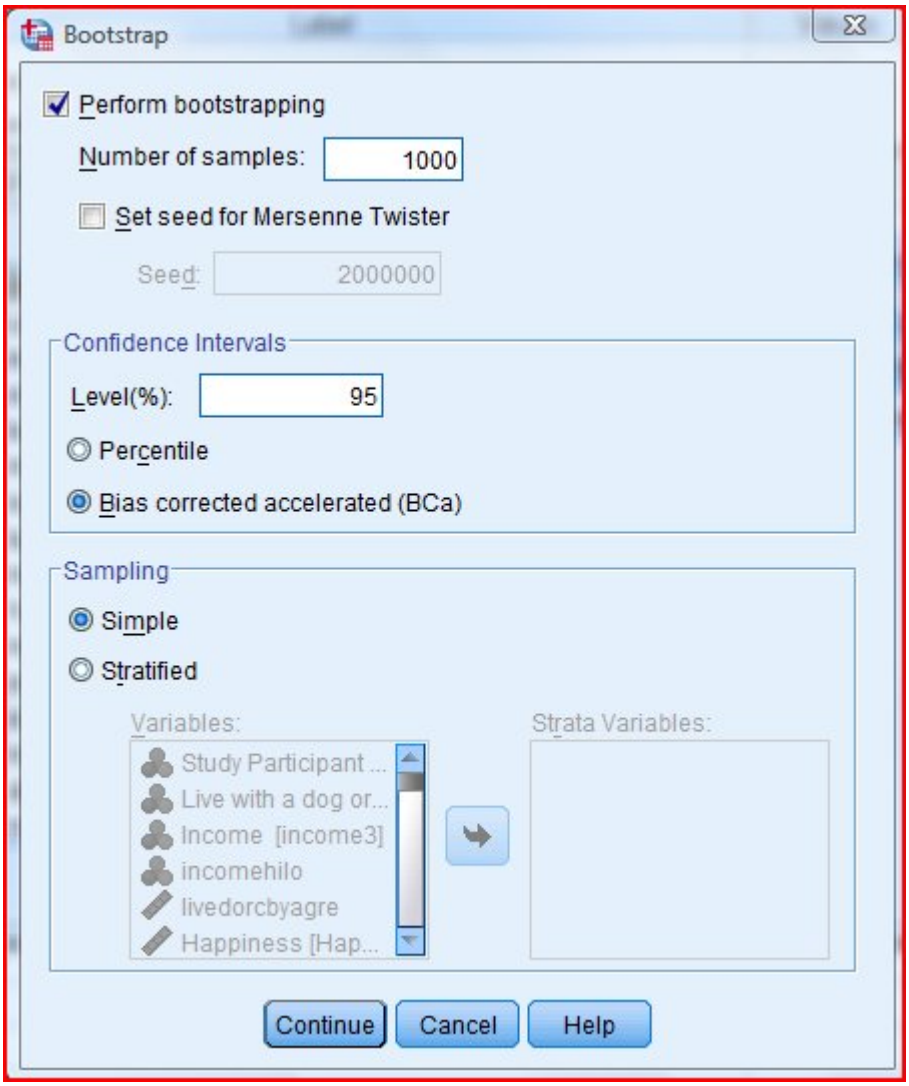


How Can Bootstrapping Methods Be Implemented?

Then, in the upper right hand corner of the original dialogue box, click the button labeled *Bootstrap*. A dialogue box will open that looks similar to the image below. Check the box in the upper left hand corner marked *Perform bootstrapping*. The box below labeled *Number of samples* will be set by default at 1,000. For our purposes here that is fine. As I mentioned before this number may vary by certain study traits/needs.

Then under the label *Confidence Intervals* click *Bias corrected accelerated*. Then under the term *Sampling* click *Simple*. Then click the button labeled *Continue*.

Then in the original dialogue box, click the button marked *OK* or *Paste* (to save the syntax).



How Can Bootstrapping Methods Be Implemented? (cont.)

At the very end of the statistical output for the binary logistic regression, the next to last box will describe the relationship between the independent variable *Does the study participant live with a dog or cat?* (0=cat, 1=dog) and the dependent variable *Is the study participant happy?* (1=yes, 0=No) as per the regular analysis.

We see that the odds ratio indicates that those who live with a dog are over five times (OR=5.391) more likely to be happy (happy=yes) relative to study participants living with a cat (the column labeled *Exp(B)* presents the odds ratio).

We can also see that the relationship is statistically significant at the $p<.001$ level (Sig.=.000).

Lastly, we see the 95% odds ratio is 2.095 to 13.866. Again this is our way of affirming that we do not know the actual odds ratio in the population. Specifically, we are affirming that in the actual population it is unlikely that people that live with a dog are exactly 5.391 times more likely to be happy relative to those living with a cat, as is indicated within our analysis. However, we are 95% confident that the true odds ratio in the population lies in between 2.096 to 13.866.

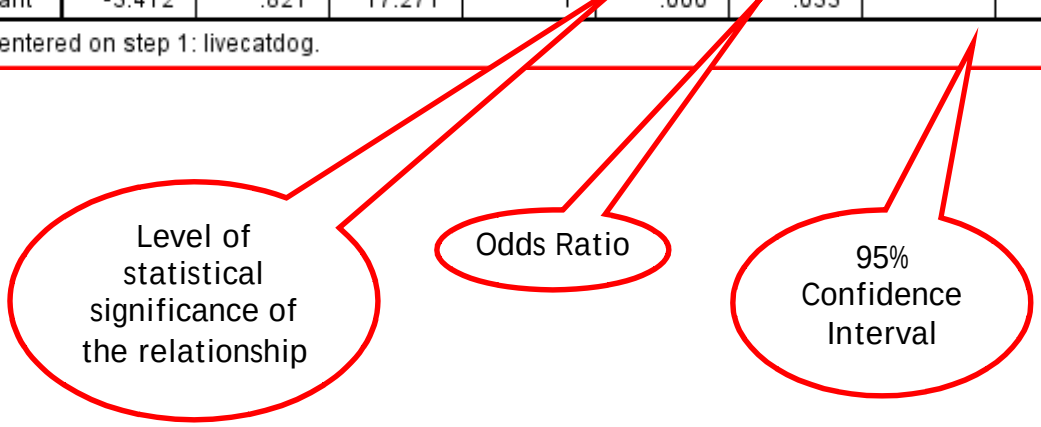
This confidence interval is rather wide. Specifically, the confidence interval suggests that the true effect size between the independent and dependent variables might be on the smaller side (OR=2.096) or it may be on the very large side (OR=13.886).

Suppose we wanted a better estimate of the odds ratio, which might more clearly convey what the effect size might be. In such a case, we could simply observe the final box within the SPSS output for this procedure which presents the bootstrap estimated values regarding this analysis.

Variables in the Equation

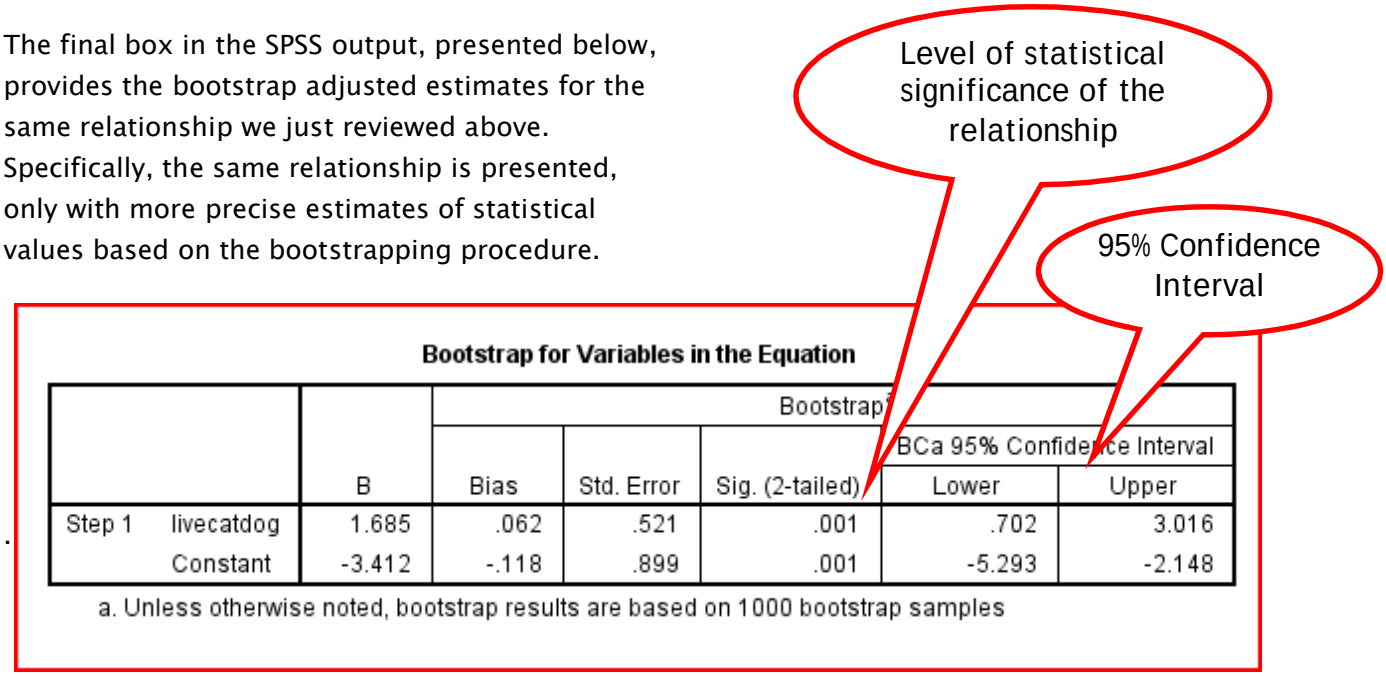
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	livecatdog	1.685	.482	12.214	1	.000	5.391	2.096	13.866
	Constant	-3.412	.821	17.271	1	.000	.033		

a. Variable(s) entered on step 1: livecatdog.



How Can Bootstrapping Methods Be Implemented? (cont.)

The final box in the SPSS output, presented below, provides the bootstrap adjusted estimates for the same relationship we just reviewed above. Specifically, the same relationship is presented, only with more precise estimates of statistical values based on the bootstrapping procedure.



Bootstrap for Variables in the Equation

		B	Bootstrap				
			Bias	Std. Error	Sig. (2-tailed)	BCa 95% Confidence Interval	
						Lower	Upper
Step 1	livecatdog	1.685	.062	.521	.001	.702	3.016
	Constant	-3.412	-.118	.899	.001	-5.293	-2.148

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

.

The first thing that might interest us is the level of statistical significance regarding the relationship. We see that the level of statistical significance regarding the relationship has changed very little, from .000 to .001. Thus, the relationship is essentially at a very similar level of statistical significance. However, the estimate of the 95% confidence interval is quite a different story. Recall, in the box above the odds ratio is 5.391 with a 95% confidence interval of 2.095 to 13.866, suggesting a significant small to very large effect might exist between the independent and dependent variables.

However, the 95% confidence interval is now .702 to 3.016. The first thing to notice is that the confidence interval now includes the null value 1.00. The null value is the number that indicates that there is not a statistically significant relationship between the independent and dependent variables. Thus, the bootstrapping procedure has suggested that this relationship might not be significant in the actual population as the number 1.00 is included in the 95% confidence interval as the lowerbound estimate is .702 and the upperbound estimate is 3.016. Furthermore, the upperbound estimate is 3.016. An odds ratio of 3.016 is considered a medium size effect between variables. Note, this upperbound effect is close to the lowerbound effect size in the first analysis we examined of 2.095.

Thus, the bootstrap procedure has suggested that the actual effect size in the population is not small (OR=2.095) to very large (OR=13.866) as indicated in the first box within the output we examined, but rather not significant (including the null value of 1.00) to at most a medium size effect (OR=3.016).

Final Comments

Obviously, this is a very simplistic presentation of bootstrapping. At more advanced levels there are different types of procedures and important rules to follow and assumptions that need to be met. However, our presentation today is a more of a "get your feet wet" sort of discussion. My goal was to draw attention to the procedure to put people on the road to learning.

Doubtlessly, the procedure can be of great use in many ways. For example, while we briefly examined the use of bootstrapping in binary logistic regression, we did not examine how we can apply the procedure within linear regression (also available in SPSS) to address the challenges related to violated test assumptions we mentioned earlier.

However, as I say regularly, a big challenge in statistics is not a problem accessing information, but a lack of awareness regarding what procedures exist which you need to learn about. Hopefully, this small discussion of bootstrapping will be useful in this respect.

Dr. William M. Bannon, Jr., the founder and CEO of StatsWhisperer can be reached at:

wb@statswhisperer.com

Thanks for your interest in our newsletter!

For information on books, products, and consultation go to:

http://www.statswhisperer.com/products/

For information on webinars and seminars go to:

http://www.statswhisperer.com/seminars/

REFERENCES

Efron, B., & Tibshirani, R.J., 1993. *An introduction to the bootstrap*. Boca Raton: Chapman & Hall/CRC.

Hesterberg, T. C., Moore, D. S., Monaghan, S., Clipson, A., and Epstein, R. (2005). "Bootstrap methods and permutation test". In David S. Moore and George McCabe. *Introduction to the Practice of Statistics*.