Paper 49

Tricks and Tips for Using the Bootstrap in JMP Pro 11

Jason Brinkley, East Carolina University, Greenville, NC;
Jennifer Mann, The Ohio State University, Columbus, OH

ABSTRACT

The bootstrap has become a very popular technique for assessing the variability of many different or unusual estimators. Starting in JMP Pro 10 the bootstrap feature was added to a wide variety of output options; however, there has not been much development as to the possible uses of this somewhat hidden feature. This paper will discuss a handful of uses that can be added to routine analyses. Examples include confidence interval estimates of the 5% trimmed mean, validation of covariates in regression analysis, comparing the differences in Spearman correlation estimates across two groups, and eigenvalues in principal components analysis. The examples will show the extra depth that can be easily added to routine analyses.

INTRODUCTION

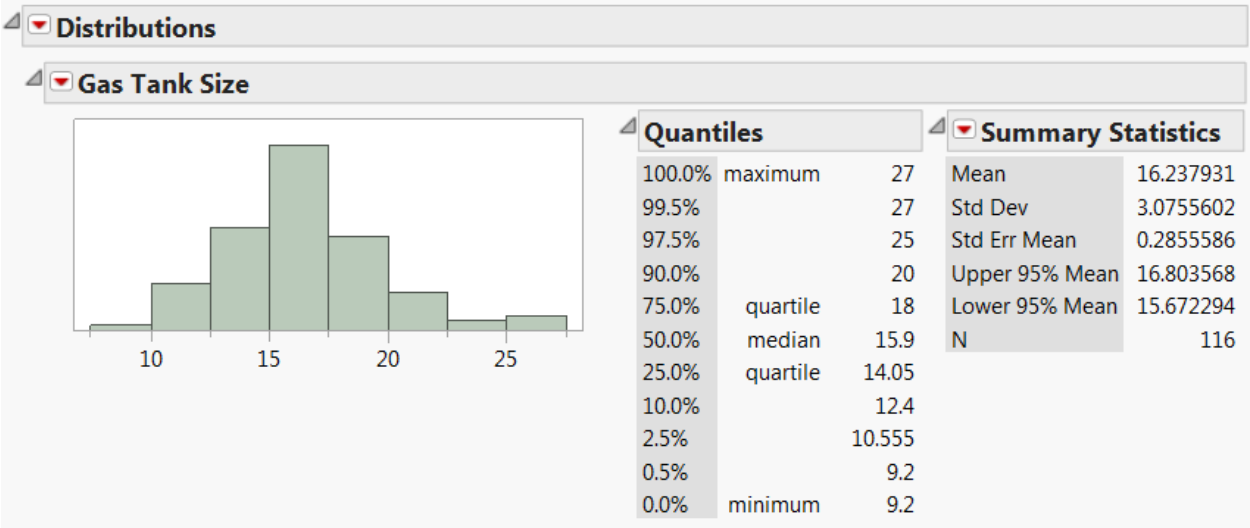
The bootstrap has become a popular technique for statistical analysis for a wide variety of metrics. Bootstrapping is the process of repeated sampling with replacement from a given dataset. The technique is powerful and becoming increasingly popular in applied analytics. First, it does not rely on parametric assumptions or large sample mathematics, which are traditionally the most common methods in variance and inferences of standard statistics (such as the mean and standard deviation). The technique relies more on the observed data and computational acumen rather than assumptions about the underlying structure or statistical model for the data.

The simple or naïve bootstrap for the mean is a relatively simple procedure. Starting with an original set of observations, denoted here as $X_1, X_2, \ldots, X_n$, create a new sample of observations, denoted here as $X_{11}, X_{12}, \ldots, X_{1n}$ by sampling the original dataset. Note that the naïve bootstrap creates a resampled version of the data whose size is the same as the original sample (*n*). To keep the samples from being exactly the same, the bootstrapped sample has been created with replacement, which means that one X_i in the original data may appear many times in the bootstrapped sample. The general idea is that the behavior of the bootstrapped sample mimics features of the original sample but is potentially different. The power and utility of bootstrap comes into play when one creates not one resampled version of the data but many different resampled datasets, thereby creating a way to explore sample to sample variation of different measures of interest. The reader is encouraged to look at Efron and Tibshirani (1993) and Chernick (2007) as excellent sources of a complete overview of bootstrapping.

JMP introduced bootstrapping as a standard option in many different analyses in JMP Pro 10. The goal of this paper is to explore the use of the bootstrap in non-standard settings to examine practical ways to utilize the bootstrap to gain additional insights and analyses. There will be no deviation from the standard options that JMP uses to bootstrap and examine the data. The goal here is to provide some examples and ideas to motivate the reader into using this feature in their day–to-day work. The examples will be pulled from the JMP Sample Data archive found under the Help menu in JMP. Visuals consist of screen captures of software options and output from JMP Pro 11. Sections are organized based on metrics of interest and grouped by sample dataset. Individuals following this guide should note that bootstrap resampling relies on random number generation so the values obtained by others may not match exactly those printed in the visuals here but should be reasonably close.

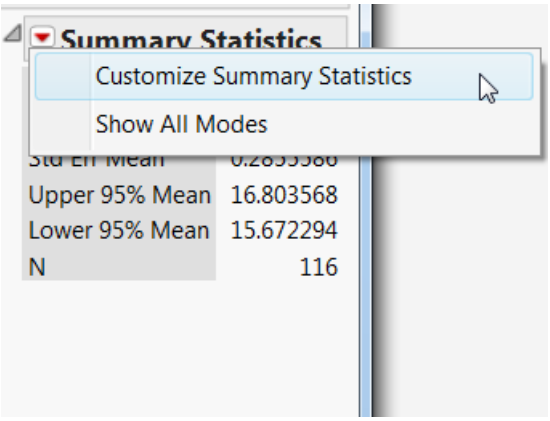
THE MEAN AND TRIMMED MEAN

Starting with an examination of the mean and trimmed mean, the first dataset under consideration is the "Car Physical Data" file in the sample data archive. The data was collected in 1990 and consist of 116 different car models from manufacturer's, which are grouped into three geographic regions (USA, Japan, Other). The data also list vehicle type (Large, Medium, Compact, Small, Sport) and vehicle metrics for weight, turning circle displacement, horsepower and gas tank size. Display 1 below illustrates standard JMP output for the distribution of Gas Tank Size.



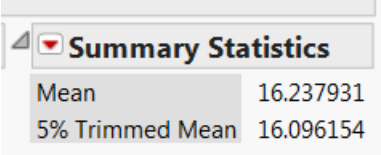
Display 1. Standard Distributions Output in JMP Pro 11

We see that the mean gas tank size is 16.23 with a 95% confidence interval (15.67, 16.80). That interval estimate is usually based on either distributional assumptions about the mean and/or large sample mathematics. To explore the use of the bootstrap, let's first add the trimmed mean as a summary statistic. The trimmed mean is calculated by removing a portion of the highest and lowest observations to provide an estimate of center less dependent on potential outliers. See Display 2 below for details.



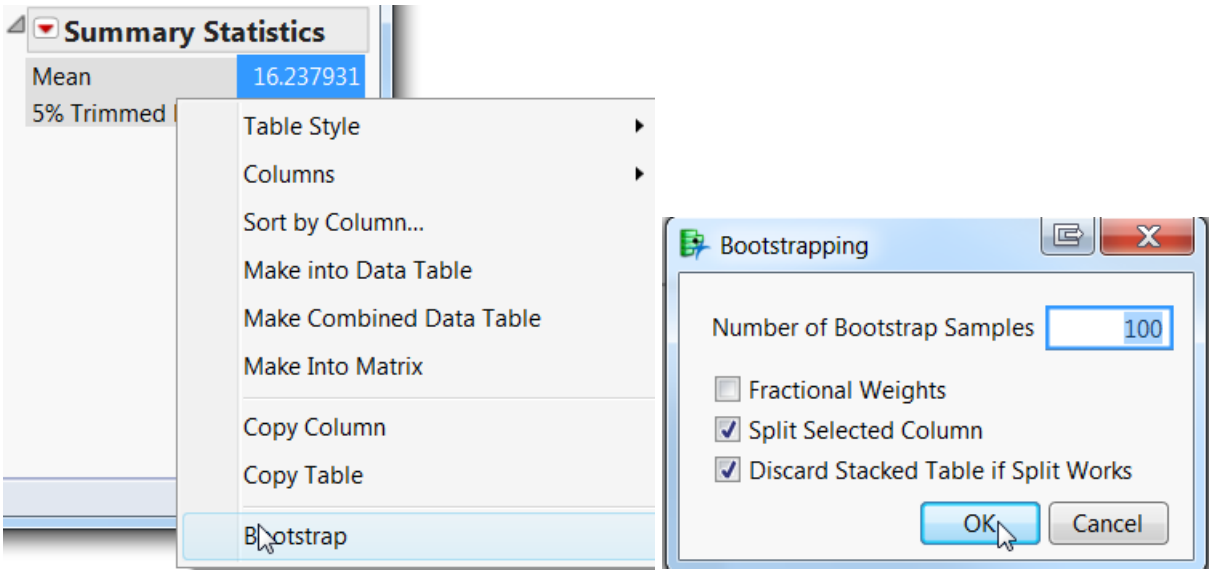
Display 2. Left Click on Red Tab for Summary Statistics

Choosing only Mean and Trimmed Mean as options, the following output is obtained.



Display 3. Mean and Trimmed Mean Estimate

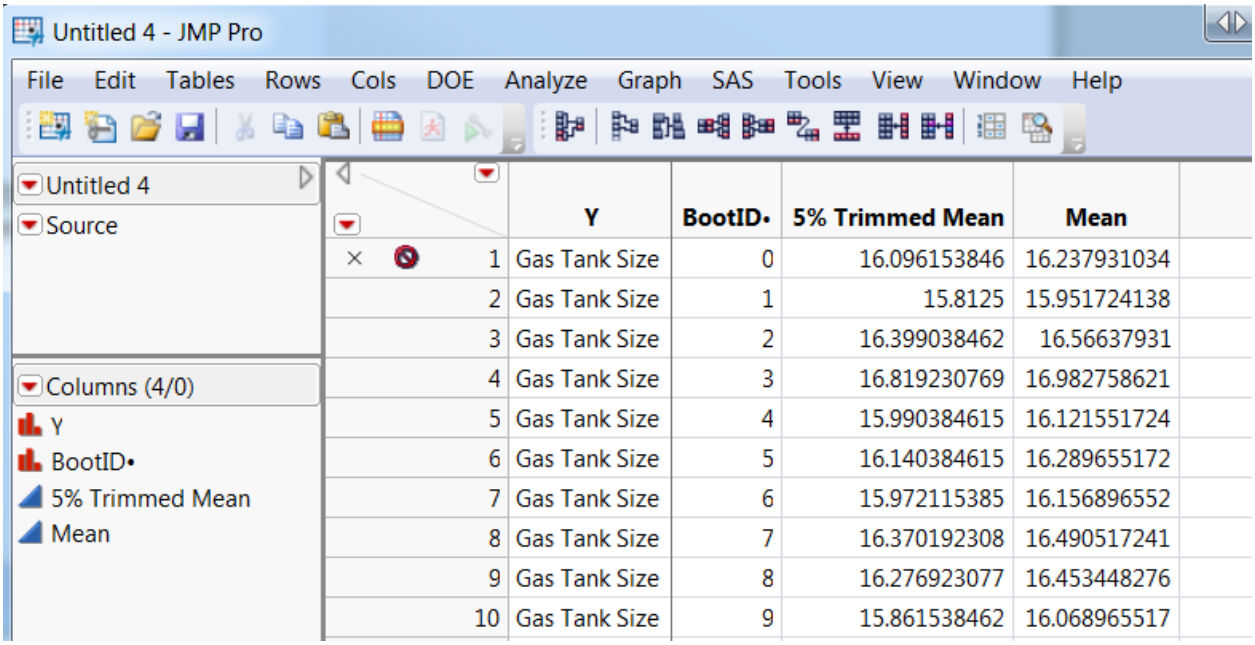
These are the two measures that we would like to perform bootstrapping on. To utilize the bootstrap in JMP Pro simply right click on the table desired for bootstrapping. Select the Bootstrap option. (See Display 4 below.)



Display 4. Bootstrap Option and Information Box in JMP

Note that there are multiple options in the Bootstrapping dialogue. To perform the naïve bootstrap simply leave the default options in place. One may be interested in increasing the number of samples if there is a need for more precision in the intervals. 100 samples is the default option but some users perfer several hundred resamplings, if not 1000 or more.

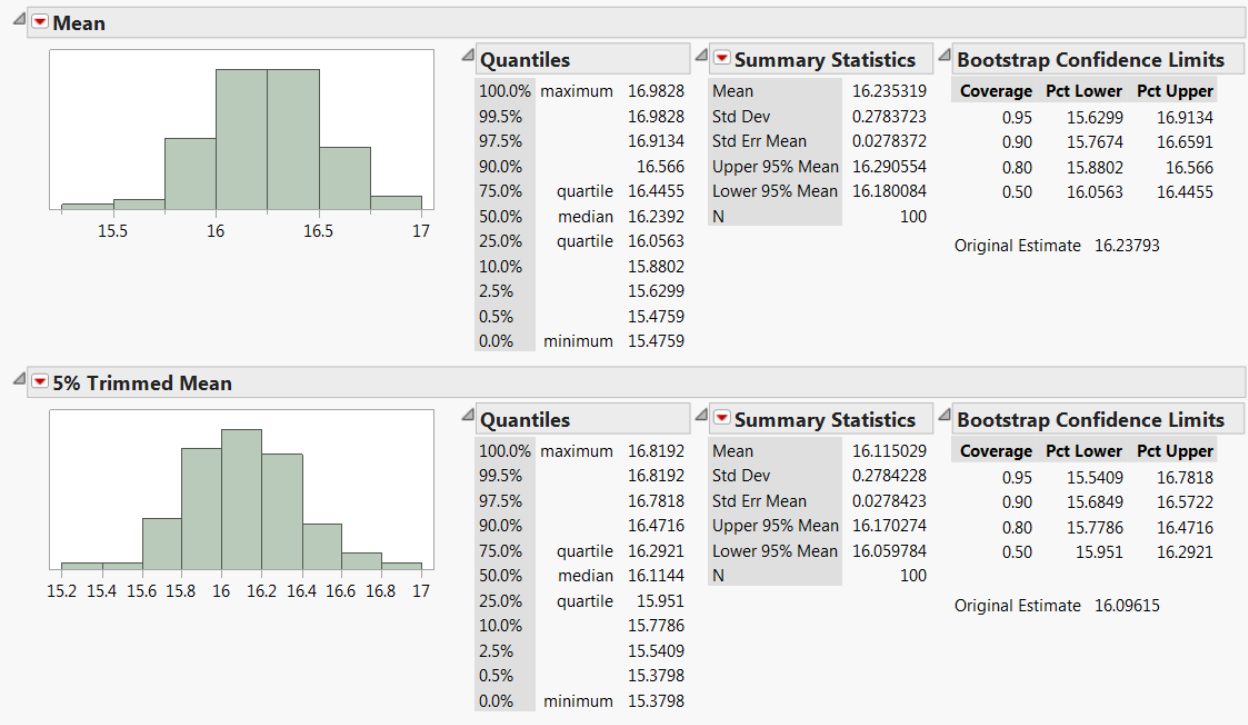
After selecting the appropriate options, a new dataset will emerge that contain the original metric values and the recalculated metric for each bootstrap sample. JMP bootstraps the data behind the scenes and users are given the recalculated metric for each resampling. See Display 5 as an example of an output dataset.



	Y	BootID•	5% Trimmed Mean	Mean
1	Gas Tank Size	0	16.096153846	16.237931034
2	Gas Tank Size	1	15.8125	15.951724138
3	Gas Tank Size	2	16.399038462	16.56637931
4	Gas Tank Size	3	16.819230769	16.982758621
5	Gas Tank Size	4	15.990384615	16.121551724
6	Gas Tank Size	5	16.140384615	16.289655172
7	Gas Tank Size	6	15.972115385	16.156896552
8	Gas Tank Size	7	16.370192308	16.490517241
9	Gas Tank Size	8	16.276923077	16.453448276
10	Gas Tank Size	9	15.861538462	16.068965517

Display 5. Example of Bootstrapped Output for 5% Trimmed Mean and Mean

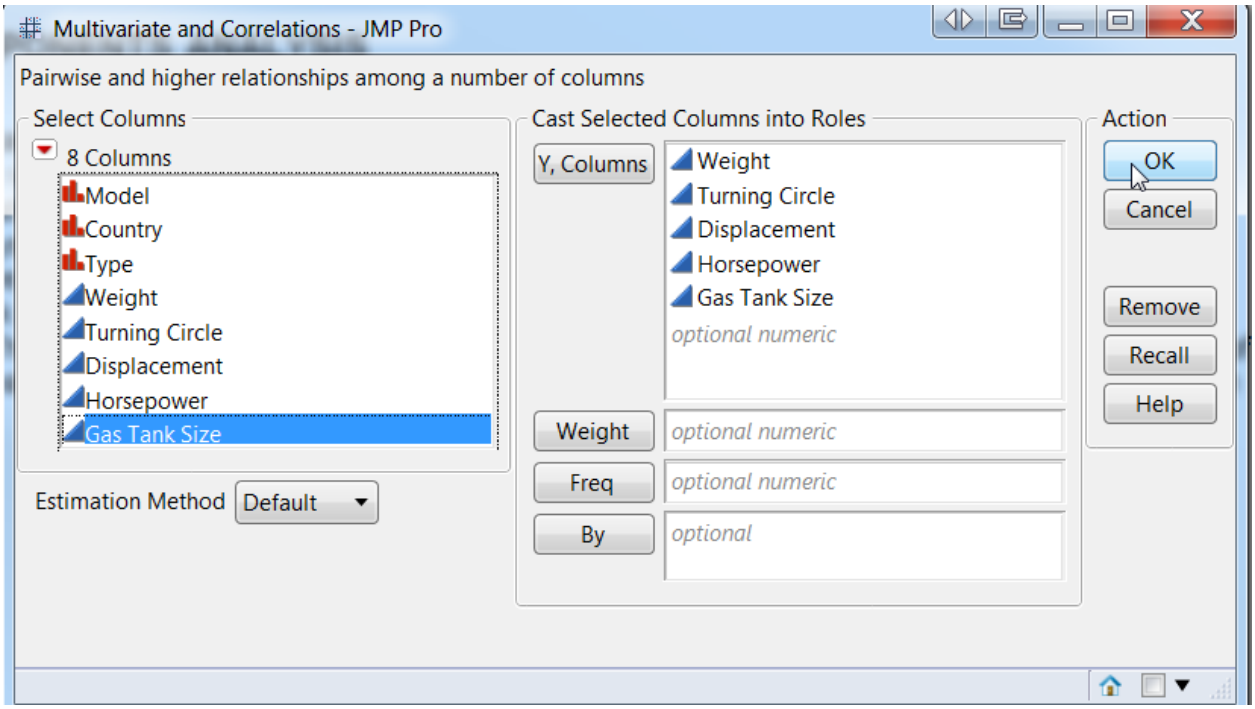
Now users simply run the distribution option again, this time choosing the Trimmed Mean and the Mean as their variables of choice. JMP recognizes that this is a bootstrapped sample and provides different output in support of the bootstrapping. Recall the original output had a mean gas tank size of 16.23 with a 95% confidence interval (15.67, 16.80). Each resampled dataset contained its' own mean gas tank size, the mean of which is listed in the output here as 16.24. Naïve bootstrapping generally relies on percentile based estimates for confidence limits, that is to say, the lowest 2.5 percentile and the highest 97.5 percentile are used to form a 95% bootstrap interval estimate for the population parameter of interest. Here we see the bootstrap intervals are a little more conservative with a wider interval (15.63, 16.91) than the one that relied on distribution assumptions or large samples. However, the true utility in bootstrapping comes into play when one examines the interval estimate for the trimmed mean. JMP does not provide confidence intervals for the trimmed mean so users would need to find other options if they wanted such an interval estimate. The output also provides the estimate from the original data so that users can compare the bootstrap sample statistics to the original data. See Display 6 below.



Display 6. Mean and Trimmed Mean Summary Statistics and Confidence Limits

PRINCIPAL COMPONENTS ANALYSIS

Principal components analysis (PCA) is a popular data reduction technique. The general idea is to reduce the number of quantitative variables under consideration by taking a smaller set of weighted averages that retain a certain amount of variation from the original data. The idea is that a smaller set of variables will be easier to work with for follow-up analysis. The interested reader should see Jolliffe (2002) for a full discussion of PCA. Returning to the "Car Physical Data," there are five quantitative measurements for each car (Weight, Turning Circle, Displacement, Horsepower, Gas Tank Size), many of which are obviously related. Larger cars must be heavier and use more fuel and power to move. Therefore there is some redundancy in this data as different measures 'explain' similar facets of the cars. PCA is a great technique to elucidate how much common variation is in data and how much reduction can be performed. JMP performs PCA using the Multivariate option found in the Multivariate Methods section (under Analyze; see Display 7).



Display 7. Multivariate Menu

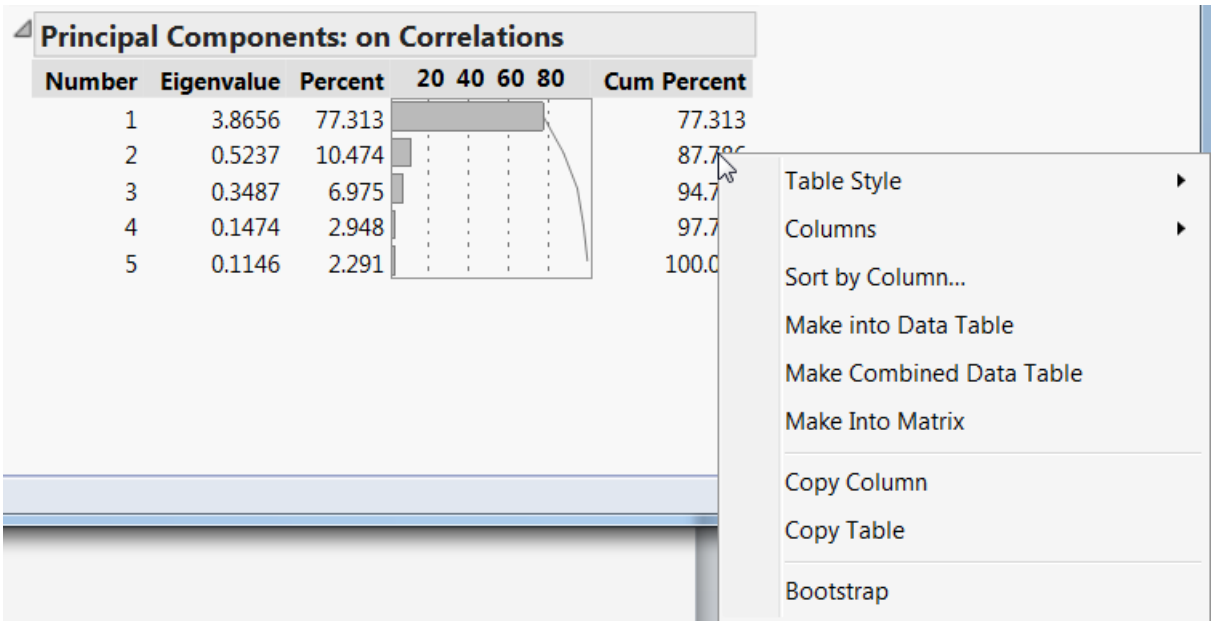
Performing PCA on the correlations in this dataset yielded the following output:

Principal Components: on Correlations

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	3.8656	77.313		77.313
2	0.5237	10.474		87.786
3	0.3487	6.975		94.761
4	0.1474	2.948		97.709
5	0.1146	2.291		100.000

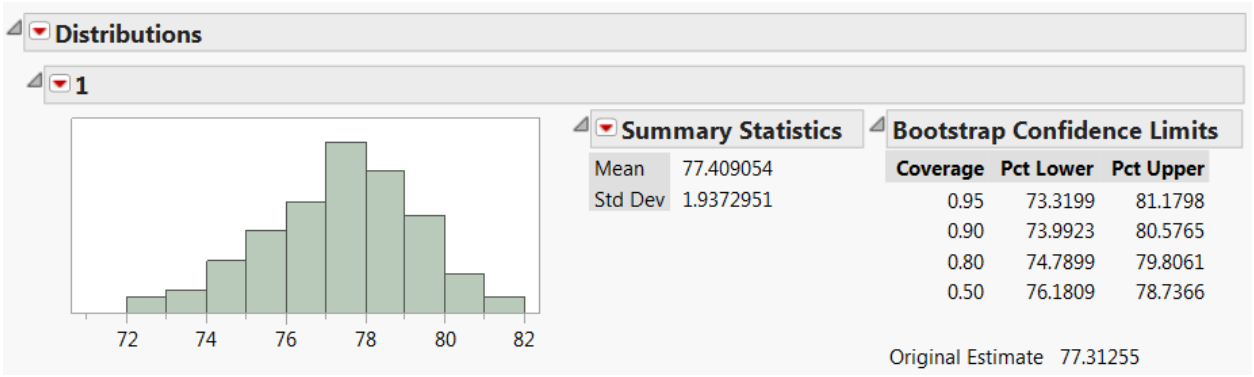
Display 8. Principle Components Analysis Output on Car Physical Data

PCA suggests that 77% of the total variation in all five measures can be summarized in one weighted sum measure. Indeed this suggests that the five variables are separate measures but are mostly aspects of only one underlying trait or quantity. Many individuals refer to such a quantity as a latent trait, because while unobserved directly in the data, there are many surrogates of it in the things that have been quantified. Now suppose one wanted an interval estimate of that percentage. What is a 95% interval estimate for the true proportion of variation that these variables explain in one latent trait or quantity? For that one can simply right click on the table and perform a bootstrap analysis in the exact same way as in the previous section. See Display 9 below for assistance.



Display 9. Selecting Bootstrap Option from PCA Output Table

Here bootstrapping the data will give you interval estimates for all 5 principal components. In general it is advisable to only examine the first couple of components because the percentages must sum to 100% and the interval estimates for each are calculated separately. Note from the output below that we obtain an interval estimate of the true proportion, which indicates that between 73.32% to 81.18% of the variation is explained. By relying on percentiles in the confidence limits, we are guaranteed to have values that fall between 0% to 100% because the interval relies on actual calculated values from resampled datasets.

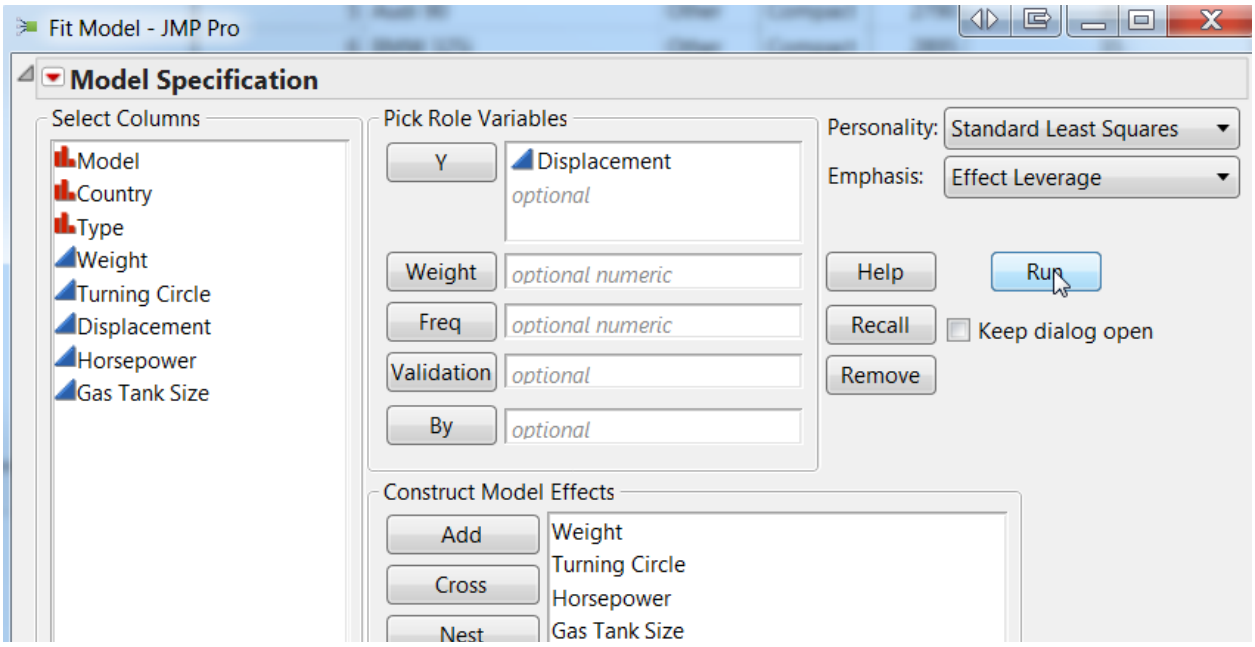


Display 10. Output from Bootstrap of PCA Data

REGRESSION ANALYSIS

Continuing to utilize the current dataset, let's consider bootstrapping with another popular analysis, regression. While there are many different types of regression analyses based on different outcomes, bootstrapping has become a popular technique for both fitting regression models and validating particular choices of models. Simply put, regression modeling allows individuals to create statistical models which fit several potential predictor variables to a particular response of interest. The goal is to look at the simultaneous impact of all these different predictor variables on the outcome. The interested reader is highly encouraged to read Harrell (1992) for a complete discussion of regression modeling. Continuing with the current example, start by fitting a regression model for the outcome

(chosen here to be Displacement) using the other quantities as predictors (Weight, Turning Circle, Horsepower, Gas Tank Size). Using JMP's Fit Model dialogue we start with the model listed in Display 11 below.



Display 11. Options for Regression Example Using Car Physical Data

This leads, in turn, to the output in Display 12 below. (Note that 95% interval estimates were added and are not part of the original standard output; to duplicate, simply right click and add those columns to the output table).

Summary of Fit

RSquare	0.797586
RSquare Adj	0.790291
Root Mean Square Error	27.66359
Mean of Response	158.3103
Observations (or Sum Wgts)	116

Analysis of Variance

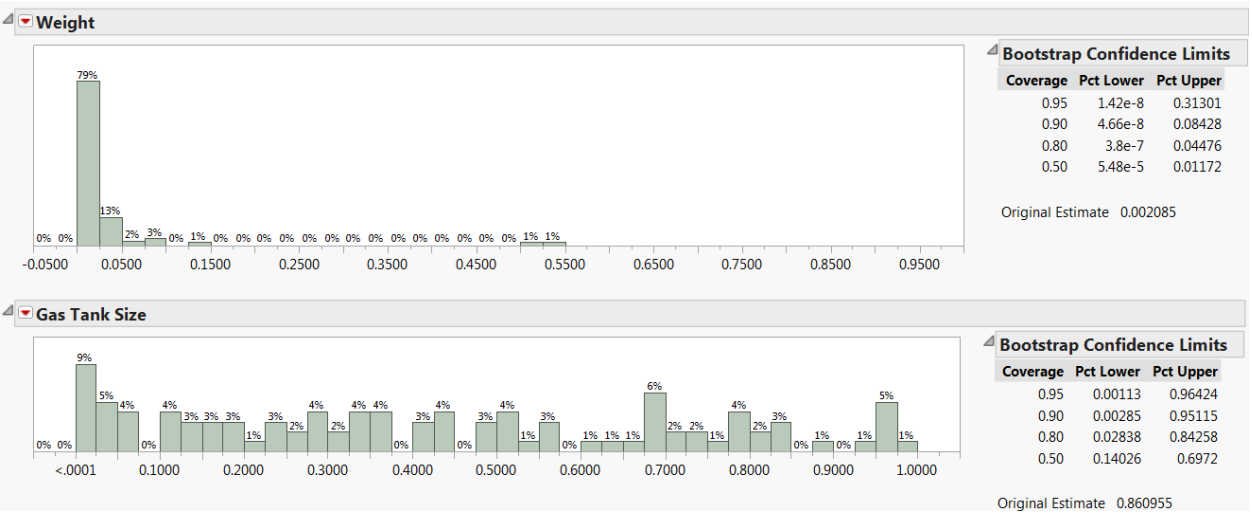
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	334715.42	83678.9	109.3450
Error	111	84945.41	765.3	Prob > F
C. Total	115	419660.83		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	-263.6616	35.38467	-7.45	<.0001*	-333.7787	-193.5445
Weight	0.0364195	0.011554	3.15	0.0021*	0.0135237	0.0593152
Turning Circle	6.2736766	1.25664	4.99	<.0001*	3.7835608	8.7637925
Horsepower	0.5894934	0.093497	6.30	<.0001*	0.4042233	0.7747635
Gas Tank Size	-0.281546	1.603658	-0.18	0.8610	-3.459301	2.8962081

Display 12. Regression Output for Car Physical Data on Displacement as Response Variable

There are a large number of different uses for the bootstrap here; for example, one could follow the ideas of the last section and obtain a 95% bootstrap interval estimate of the R-Square value for this model. Or one could ignore the large sample confidence intervals for the regression slope parameters and instead obtain 95% bootstrap intervals of those same slope parameters. The latter may be preferred if the variables in the model come from unusual or unstable distributions of data. However, as to not replicate techniques discussed in previous sections, this example will look at something different. As a validation technique, consider taking bootstrap replications of the p-values from the above model. The question of interest is, in what proportion of bootstrap resamplings do we see a statistically significant p-value for each variable in the model? If one or a handful of unusual influential observations are driving this model, then those variables may not stand up to such further scrutiny. Bootstrapping just the p-value column (Prob > |t|) and looking at the distribution of p-values for Weight and Gas Tank Size, we see the following (output adjusted using standard JMP options) in Display 13 below.



Display 13. Bootstrap Output for P-values of Weight and Gas Tank Size

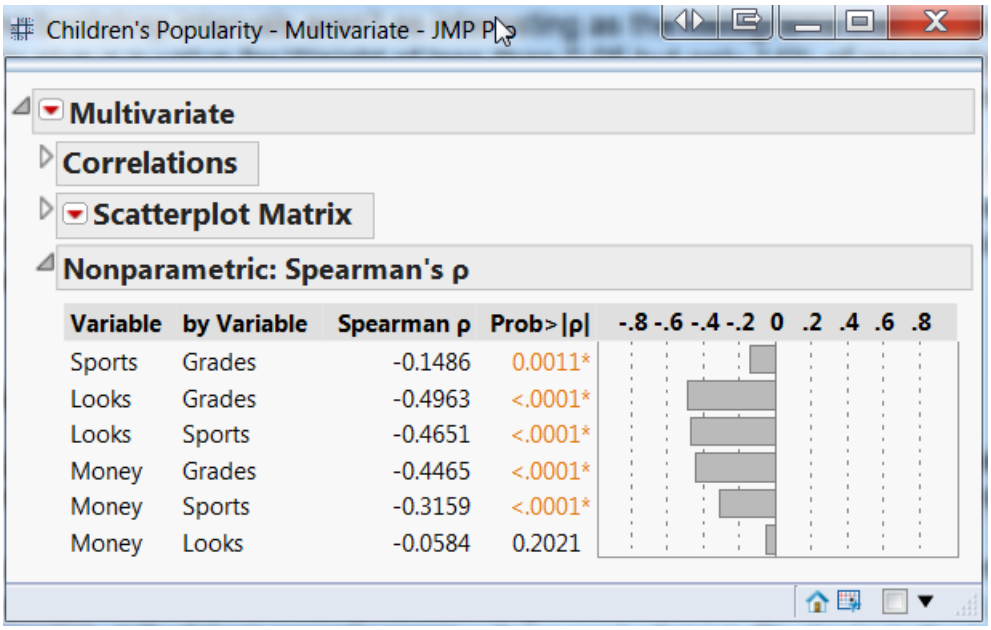
Here the bootstrap intervals aren't as interesting as the histograms themselves. Note that 92% of resampled datasets give a p-value for Weight of less than 0.05 but only 14% of resamplings indicate Gas Tank as a significant predictor. Here we see the impact of resampling. The original model suggested that Gas Tank Size is not a significant predictor, but a small amount of resamplings do show significance. This is a simple and informative diagnostic regarding the stability of different predictors in a statistical model. It is easy to employ and provides an extra layer of analyses regarding whether the predictors in a given model have some evidence of stability. There is no hard and fast rule as to how what proportion of resamplings should have a significant predictor, but it is easy to speculate that it should be at least 0.50 (better than coin flips).

SPEARMAN CORRELATIONS

Turning to the last example, a different sample dataset is needed. The sample dataset "Children's Popularity" contains 480 observations from a study by Chase and Dummer (1992). JMP notes showing the following description:

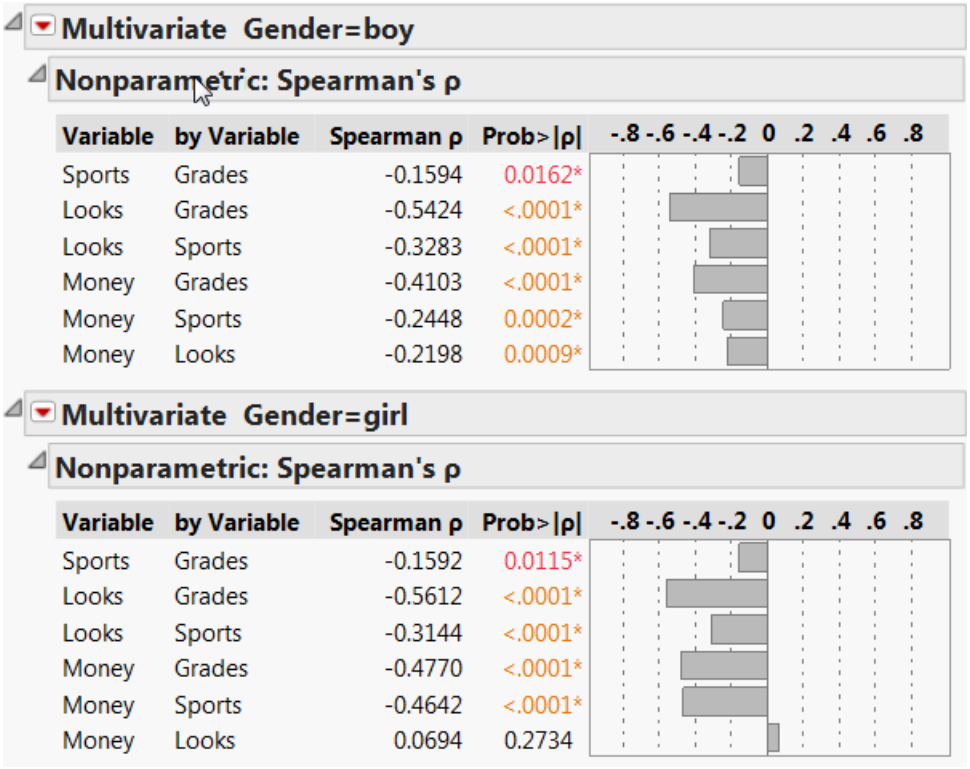
> "*Subjects were students in grades 4-6 from three school districts in Ingham and Clinton Counties, Michigan. Chase and Dummer stratified their sample, selecting students from urban, suburban, and rural school districts with approximately 1/3 of their sample coming from each district. Students indicated whether good grades, athletic ability, or popularity was most important to them. They also ranked four factors: grades, sports, looks, and money, in order of their importance for popularity. The questionnaire also asked for gender, grade level, and other demographic information.*"

The ranked factors are the values of primary interest. Using the multivariate option (see visuals in PCA section) to look at correlations between the variables Grades, Sports, Looks, and Money the following output is derived. (See Display 14 below.)



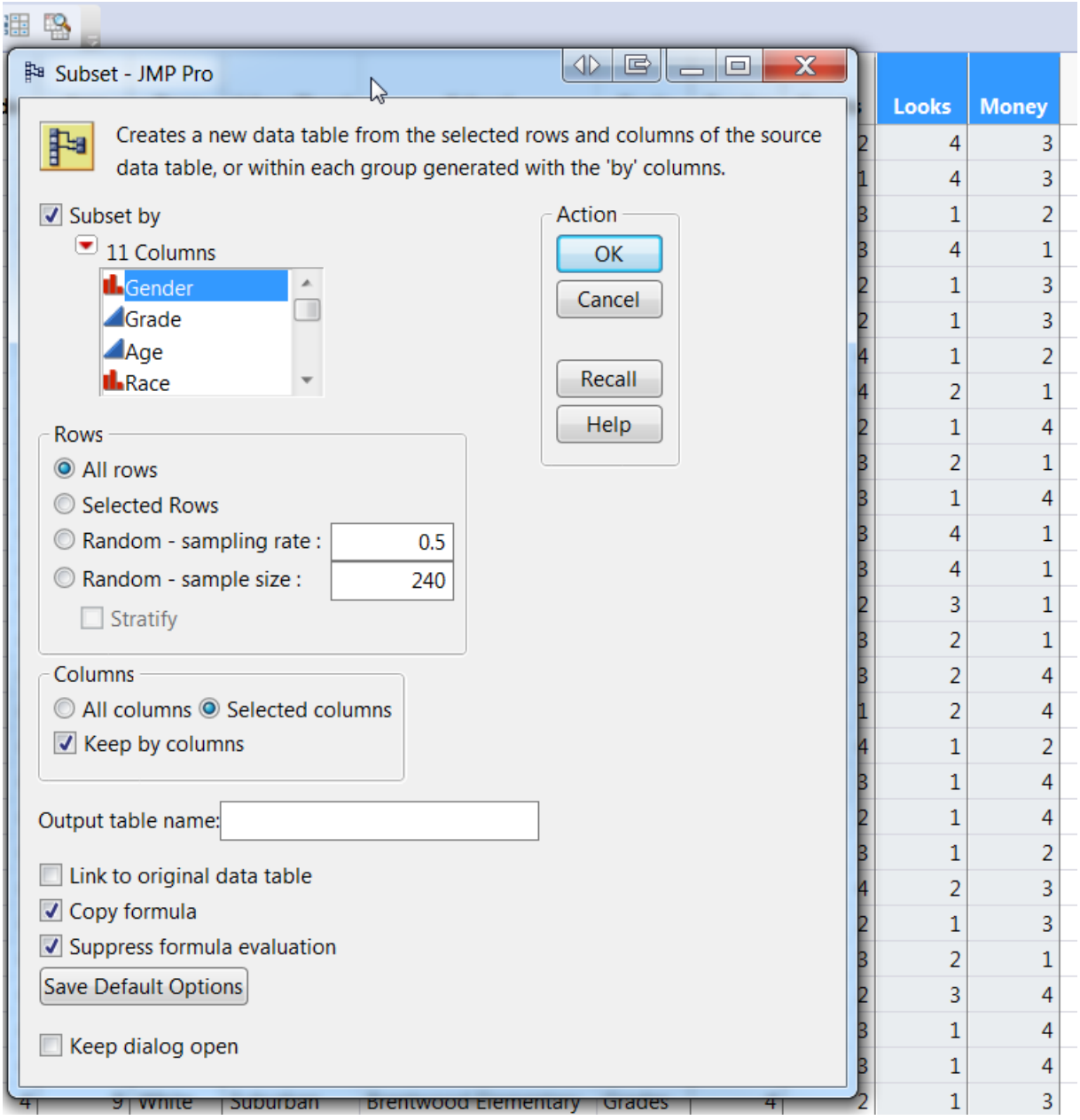
Display 14. Correlation Output from Ranked Variables in Children's Popularity Dataset

Spearman correlations illustrate that Looks and Money both seem to correlate weakly with Grades; however, Money and Looks have virtually no correlation. Consider a stratified analysis that considers the Spearman correlation by Gender. (See Display 15 below.)



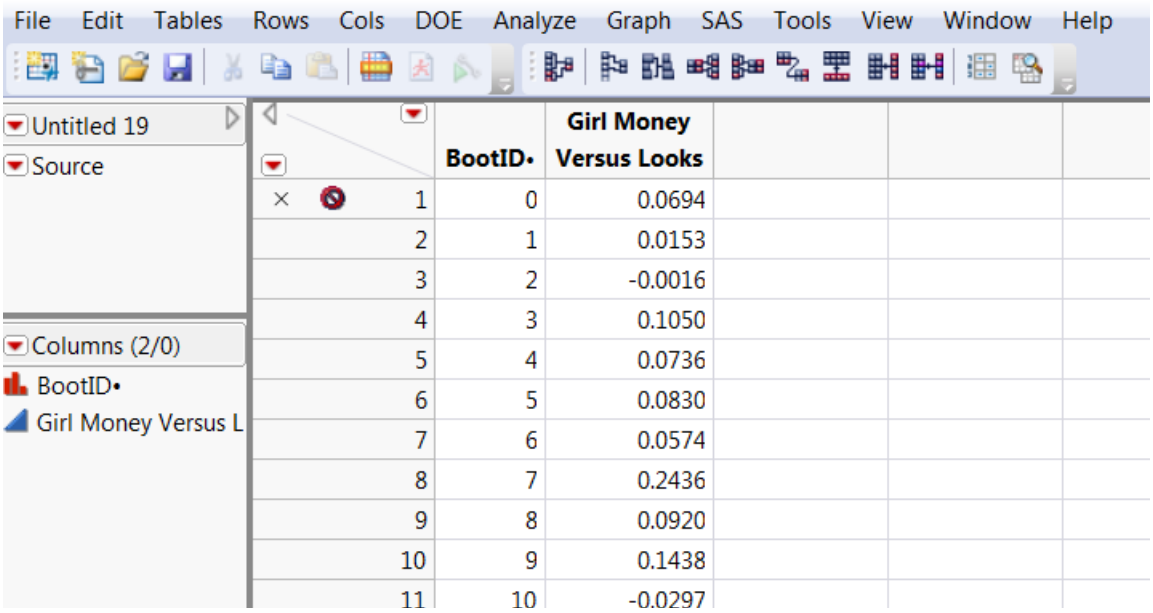
Display 15. Stratified Analysis of Spearman Correlation by Gender on Children's Popularity Dataset

Here we see something of a different constellation of correlations between the genders. With the measured association between Money and Looks to be -0.2198 for boys and 0.0694 for girls, the question that arises is whether the correlation is significantly higher for boys than girls. We have an estimate of that difference to be -0.2892, but can one find a 95% bootstrap confidence interval for that difference to determine if it contains zero? First note that the analysis here is stratified by Gender and JMP will NOT take a bootstrap sample of any table that has used the "By" option; therefore one must manually split the data into a subset for this analysis. Start by creating separate data sheets for boys and girls. Here we will focus on just the Gender, Money, and Looks variables. (See Display 16 below for visuals of this process.)



Display 16. Creating Subsets of Money and Looks Variables By Gender

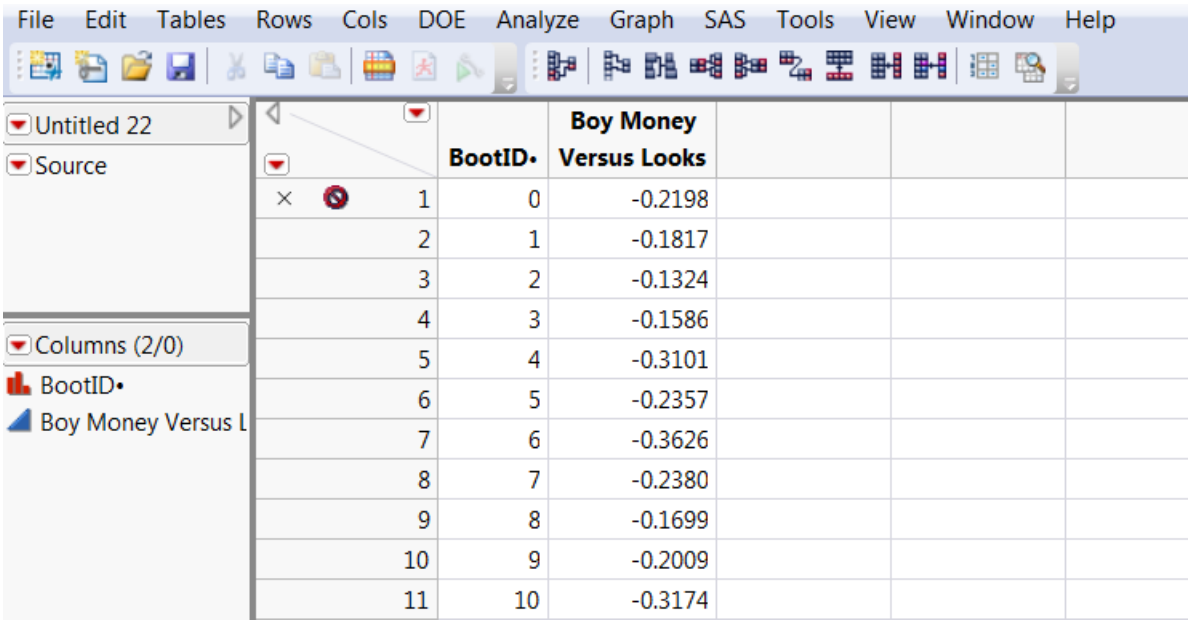
Then, for each subset, find the Spearman Correlation between Money and Looks and bootstrap that value. Renaming the lead column as Girl Money Versus Looks, the output should look similar to Display 17below.



	BootID•	Girl Money Versus Looks
1	0	0.0694
2	1	0.0153
3	2	-0.0016
4	3	0.1050
5	4	0.0736
6	5	0.0830
7	6	0.0574
8	7	0.2436
9	8	0.0920
10	9	0.1438
11	10	-0.0297

Display 17. Bootstrap Output of Spearman Correlations on Money Versus Looks for Girls

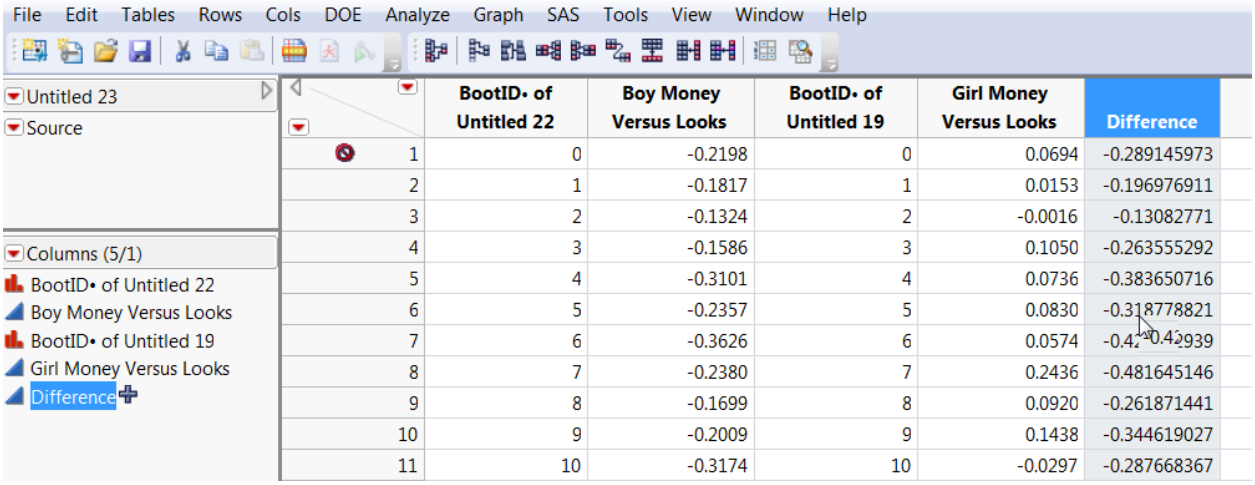
Now create a similar bootstrapped data for the Boys. (See Display 18 below.)



	BootID•	Boy Money Versus Looks
1	0	-0.2198
2	1	-0.1817
3	2	-0.1324
4	3	-0.1586
5	4	-0.3101
6	5	-0.2357
7	6	-0.3626
8	7	-0.2380
9	8	-0.1699
10	9	-0.2009
11	10	-0.3174

Display 18. Bootstrap Output of Spearman Correlations on Money Versus Looks for Boys

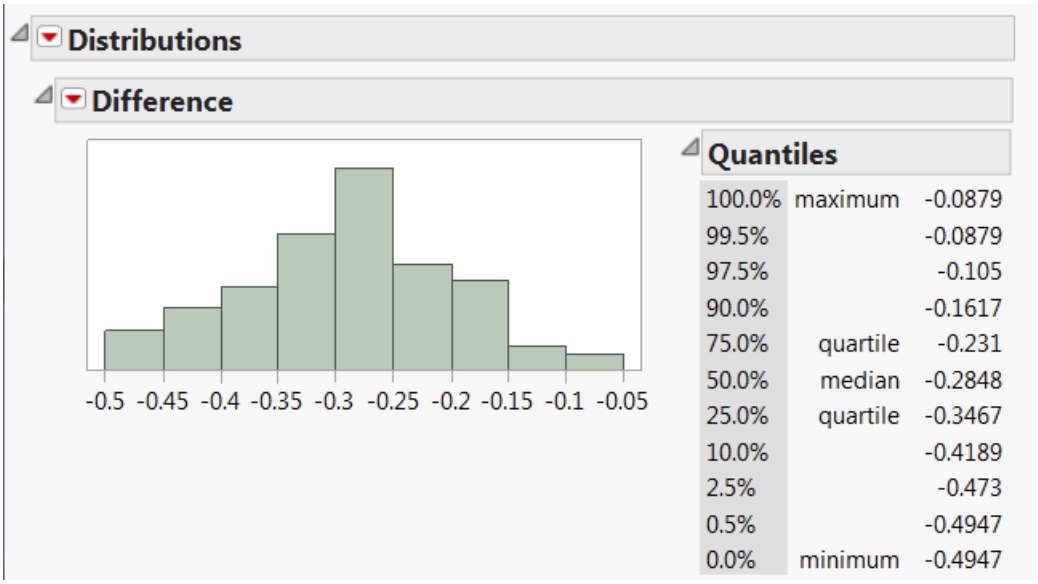
Join those tables by BootID and create a new column for the difference in the correlations. Exclude BootID = 0 since it is the original value. The output for this is given in Display 19 below.



	BootID• of Untitled 22	Boy Money Versus Looks	BootID• of Untitled 19	Girl Money Versus Looks	Difference
1	0	-0.2198	0	0.0694	-0.289145973
2	1	-0.1817	1	0.0153	-0.196976911
3	2	-0.1324	2	-0.0016	-0.13082771
4	3	-0.1586	3	0.1050	-0.263555292
5	4	-0.3101	4	0.0736	-0.383650716
6	5	-0.2357	5	0.0830	-0.318778821
7	6	-0.3626	6	0.0574	-0.42[illegible]939
8	7	-0.2380	7	0.2436	-0.481645146
9	8	-0.1699	8	0.0920	-0.261871441
10	9	-0.2009	9	0.1438	-0.344619027
11	10	-0.3174	10	-0.0297	-0.287668367

Display 19. Merged Dataset, by BootID, Excluding Original Values, With Calculated Difference Between Spearman Correlations

Fitting the distribution of the Difference column and manually and examining the 2.5% percentile to 97.5% percentile yields a 95% bootstrap interval for the difference in Spearman correlation values of (-0.4947, -0.105), indicating that the boys correlation between Money and Looks is stronger than the corresponding females. See the output in Display 20 below.



Display 20. 95% Bootstrap Interval for the Difference in Spearman Correlation Values, Found by Examining the 2.5% to 97.5% Quantiles

CONCLUSION

The goal of this paper is to illustrate several different accessible examples of how a standard user can add bootstrapping to their routine analyses. Whether one wants interval estimates for non-standard measures, to further explore some aspect of the data, provide some measure of reliability/validity to existing work, to compare subgroups of the data, the examples here illustrate bootstrapping as an easy to use and flexible tool.

We do want to conclude with some simple cautions to the reader. There are some cases in which bootstrapping fails. The idea of the bootstrap 'failing' usually revolves around the idea that interval estimates may not cover the true parameter values. In some cases, the bootstrap is not conservative enough. See the text by Chernick (2007) for more details. However, many of these examples entail situations where the measure of interest is not smooth (i.e. may contain a lot of indicator functions). Do proceed with caution and always check the distribution of the values that you are bootstrapping both in the original data and in the bootstrap metrics. Also note that it is possible that the software will provide bootstrap estimates for metrics for which the bootstrap is not appropriate.

REFERENCES

Chase, Melissa A. and Dummer, Gail M. 1992. "The Role of Sports as a Social Determinant for Children" *Research Quarterly for Exercise and Sport*. 418-424. Reston, Virginia. American Alliance for Health, Physical Education, Recreation & Dance. Available from the Data and Story Library at http://lib.stat.cmu.edu/DASL/Datafiles/PopularKids.html.

Chernick, Michael R. 2007. *Bootstrap Methods: A Guide for Practitioners and Researchers, 2nd Edition*. Hoboken, New Jersey. John Wiley & Sons, Inc.

Efton, Bradley and Tibshirani, Robert J. 1993. *An Introduction to the Bootstrap*. New York, New York. Chapman and Hall/CRC.

Harrell, Frank Jr. 2002. *Regression Model Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Secaucus, New Jersey. Springer-Verlag.

Jolliffe, I.T. 2002. *Principal Component Analysis*. New York, New York. Springer.

CONTACT INFORMATION

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

Name: Jason Brinkley, Ph.D.
Enterprise: East Carolina University
Address: Mail Stop 668, 2435 Health Sciences Building
City, State ZIP: Greenville, NC 27834
E-mail: brinkleyj@ecu.edu