Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 18, Number 12, September 2013 ISSN 1531-7714

Normality of residuals is a continuous variable, and does seem to influence the trustworthiness of confidence intervals : A response to, and appreciation of, Williams, Grajales, and Kurkiewicz (2013)

Jason W. Osborne
University of Louisville

Osborne and Waters (2002) focused on checking some of the assumptions of multiple linear regression. In a critique of that paper, Williams, Grajales, and Kurkiewicz correctly clarify that regression models estimated using ordinary least squares require the assumption of normally distributed errors, but not the assumption of normally distributed response or predictor variables. They go on to discuss estimate bias and provide a helpful summary of the assumptions of multiple regression when using ordinary least squares. While we were not as precise as we could have been when discussing assumptions of normality, the critical issue of the 2002 paper remains – researchers often do not check on or report on the assumptions of their statistical methods. This response expands on the points made by Williams, advocates a thorough examination of data prior to reporting results, and provides an example of how incremental improvements in meeting the assumption of normality of residuals incrementally improves the accuracy of confidence intervals.

Let's start with this assertion: that our goal as researchers and scholars is to understand or reveal truth. In our narratives, we attempt to pull strands of data, observation, intuition, scholarship, theory, experience, and reality together for a greater purpose. It is my belief that the ultimate goal of our scientific narrative is to understand better a small portion of the world we care deeply about. If we start with that premise, and pursue it in good faith, I think we are all better for it. Why is this important? Because it is easy in works such as the original article being discussed (Osborne & Waters, 2002), or in articles that respond to those articles (Williams, Grajales, & Kurkiewicz, 2013) to lose sight of important goals, focusing rather on minutiae that rarely influence the majority of statistical research practice.

What is important to me, and I assume to my colleagues who so aptly critiqued our earlier work, is that we help improve statistical practice, and thereby, improve the quality of the knowledge being produced by the legions of researchers around the world who use these techniques on a daily basis. Let's add a second assertion at this point: that a significant portion researchers in our fields fail to report basics like having tested assumptions and cleaned data. For example, a recent examination of top journals in several fields (Osborne, Kocher, & Tillman, 2012) summarized in Figure 1, show that authors in top journals do not have a good track record of reporting having attended to these issues. I think it is difficult to argue that we should not attend to, and report having attended to, basic data cleaning and testing of assumptions if in fact you are convinced that assumptions and data quality matters. I worry there is a not uncommon sentiment amongst researchers that data cleaning is not desirable and that assumptions are largely "robust" to violation, and as such, neither issue is much worth worrying about (Osborne, 2012).

I will first congratulate Williams et al. (2013) for a keen critique of our original work. It is a good clarification of our original work. They were correct in noting that we were not as precise as we could have

been when discussing assumptions of normality.[1] As I reflected on their points, what I find myself concerned with today is making sure researchers are motivated to expend effort to examine their data for illegitimately influential cases (e.g., outliers) that might bias results. As Cohen et al. note (Cohen, Cohen, West, & Aiken, 2002, p. 141), one of the primary reasons for examining

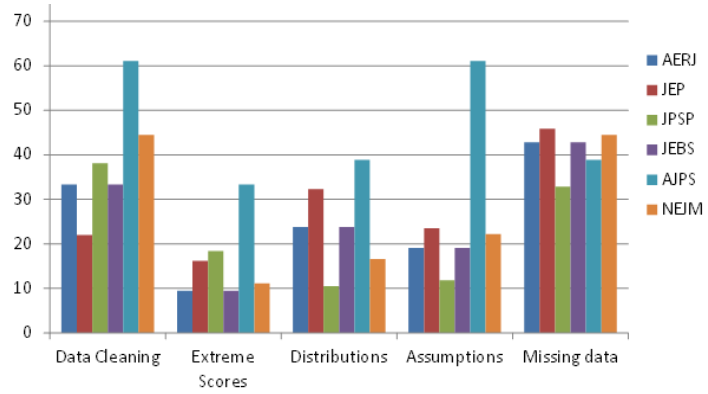

Figure 1. Percent of articles in prominent journals reporting basic aspects of data cleaning and assumptions. From Osborne, Kocher, & Tillman (2012).

normality of residuals is to identify model misspecification or inappropriately influential cases rather than the actual normality or non-normality of the residuals.[2] In fact, much of our narrative in the section of our original paper that Williams et al. (2013) objected to is devoted to identification of outliers (inappropriately influential cases).[3]

Non-normality is not always caused by influential cases or outliers, but non-normality of univariate distributions or residuals can be an initial indicator that there are potential data cleaning issues. Although perhaps inelegantly argued in our original piece, one of our intentions in advocating for exploring normality was to motivate routine examination of their data prior to analysis. Readers interested in this topic can refer to Osborne (2012) or Osborne and Overbay (2004).

Aside from initial screening for illegitimately influential (or just plain illegitimate) data points, it is important to meet assumptions and to have the tools necessary to deal with situations where assumptions are not reasonably met (as in the strictest sense, assumptions are almost never completely met). Providing researchers with practical solutions to common problems, and motivating them to examine the data and use these solutions where appropriate is critical, it seems. From this practical perspective, one common question from researchers exploring their residuals is: "How do I make the residuals more normal if I find this assumption seriously violated?" In my mind, if one has done a thorough job of examining and removing inappropriately influential data points, and the residuals are still non-normal enough to cause concern over the validity of the results of the analyses, I might suggest experimenting with some transformations of the original variables (interested readers can refer to Osborne (2002, 2010, 2012).

Williams et al. (2013) present an example where non-normally distributed variables produce normally-distributed residuals, further showing in the context of small samples that this subsequently produces trustworthy effect estimates and 95% confidence intervals. This is a good point, but made me wonder

[1] Note that this discussion is strictly related to OLS regression. In other types of regression (i.e., logistic regression) where assumptions are different, data cleaning is still important but there might not be any assumptions regarding distributions of the variables or the residuals. In other analyses, such as multivariate analyses or structural equation modeling (Byrne, 2010) multivariate normal distributions of the variables are critical, and dealing with individual variable non-normality and influential cases can help address violations of multivariate non-normality (although not always, as one can have universal univariate normality without multivariate normality, much as one can have normally distributed variables and non-normally distributed residuals in OLS regression).

[2] As we and many others have noted, most scholars have asserted that multiple regression analyses are "robust" to violations of the assumption of normal distribution of the residuals (except in very small samples, which are problematic for other reasons).

[3] Another possible critique of our original article might include the fact that we neglected the other half of Cohen et al.'s point: that non-normality of residuals could be due to model misspecification—leaving out a variable that should be modeled, neglecting to model curvilinearity or interactions, etc. These points are more well elaborated in my forthcoming book on logistic regression (Osborne, in press) and perhaps best presented regarding OLS regression by Aiken and West (1991).

what happens when the distribution of residuals does not so closely approximate a Gaussian distribution? I worry that readers of these articles, seeing assertions to the effect that normally distributed errors are not required for regression coefficients to be unbiased, consistent, and efficient will fall into the seductive trap of assuming regression analyses are "robust" to violations of assumptions, and thus might conclude that it is not necessary to test the assumption of normality—or even to examine the data for inappropriately influential data points. So in appreciation and support of the points made by Williams et al. (2013), let's examine an example that readers might encounter. Below I present real data with continuous variables (one of which is markedly non-normal), with reasonably normally distributed residuals. If we view normality as a continuous variable, and take seriously the point that less normal residuals can lead to less trustworthy confidence intervals (95%CIs), we should be able to demonstrate this effect. Further, I will propose two possible methods toward improving the normality of residuals: data cleaning of inappropriately influential cases and transformation of original variables.

An example

This example is borrowed from Osborne (2012, Chapter 8) and involves a data set from the American Association of University Professors (AAUP) listing data on over 1100 institutions of higher education in the United States from back in the 1990s. If you have worked in higher education in the US, and thought about faculty salaries, there are many obvious factors that influence salary—field of specialization, whether the university is public or private, faculty rank, and size of the university. In the USA, institution size and faculty salaries tend be reasonably well correlated. In this example, salary has a univariate distribution that is not markedly non-normal (skew of 0.35, and kurtosis of 0.12), but institution size is markedly non-normal (skew of 2.62 and a kurtosis of 8.90). The distributions of these variables are presented below in Figures 2 and 3.

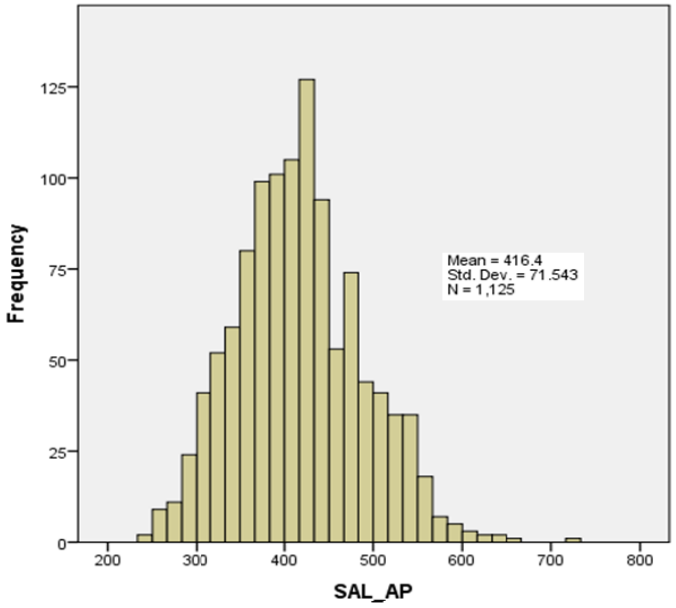

Figure 2. Distribution of Associate Professor Salaries in the US in the 1990s

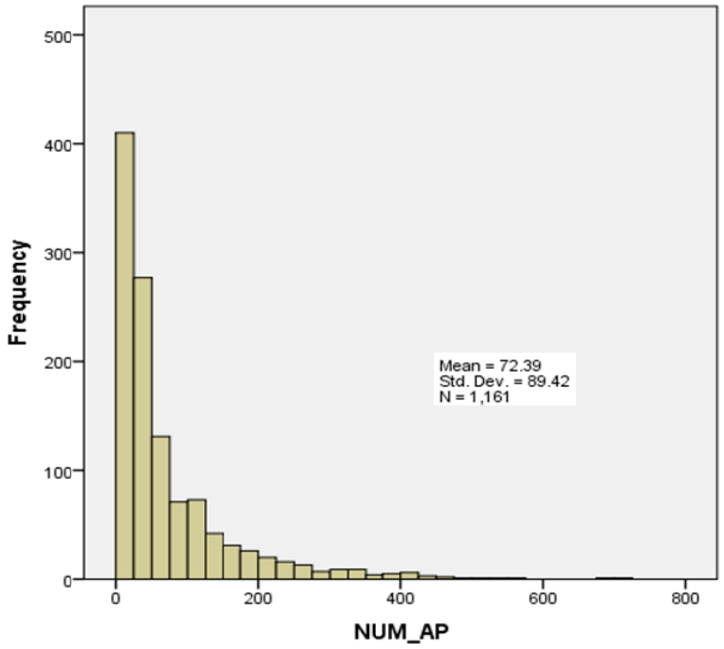

Figure 3. Distribution of institution sizes (# of faculty) in the 1990s

In accord with one of the points from Williams et al., a regression analysis predicting salary (SAL_AP) from institution size (NUM_AP) reveals residuals that

probably do not raise concerns about meeting the assumption of normality, as you can see in Figure 4:

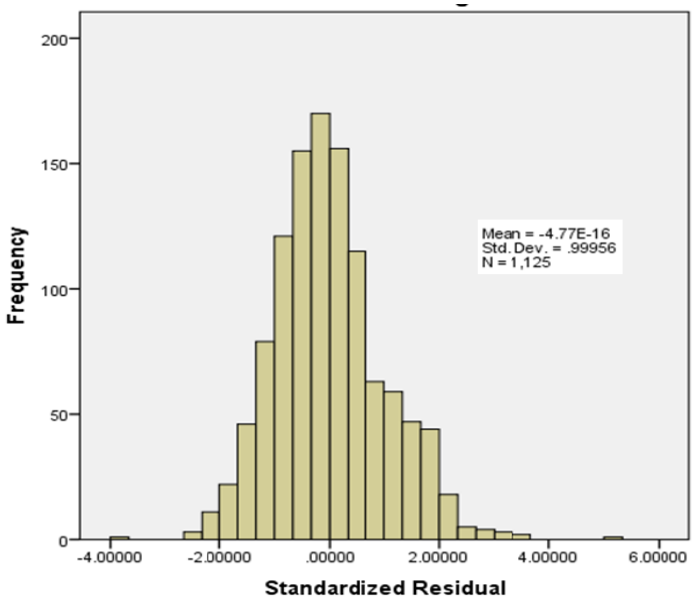

Figure 4. Residuals from regression equation predicting SAL_AP from NUM_AP

These residuals have a skew of 0.50 and a kurtosis of 0.89. Not terrible, but significantly different from Gaussian by most measures (e.g., Kolmogorov-Smirnov Z =1.91, $p < .001$). The results from this regression are: beta = 0.49, $p < .0001$, and an $R^2 = .24$. Given the point in Willams et al. (2013) about non-normally distributed residuals producing untrustworthy confidence intervals, it might be desirable to attempt to improve adherence to this assumption. So how to *easily* improve the normality of these residuals? No transformation of the independent variable resulted in marked improvement in normality of residuals,[4] but examination of Figure 4 reveals several cases that could be considered outliers—standardized residuals more than 3 standard deviations beyond the mean.[5] Removal of 7 cases with standardized residuals more than three standard deviations beyond the mean results in residuals that are closer to normal, as you can see in Figure 5, and improved normality (skew for the residuals dropped from 0.50 to 0.34 and kurtosis dropped from 0.89 to -0.06).

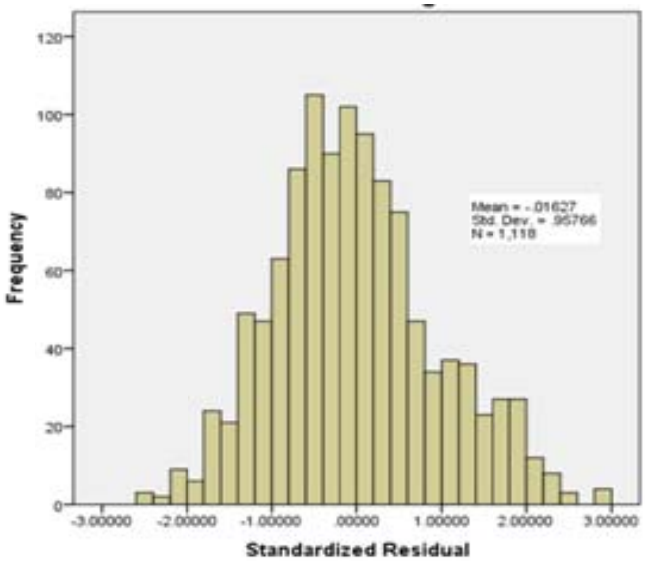

Figure 5. Standardized residuals following removal of extreme cases

Let's return our focus to normality and confidence intervals, and the extent to which they are trustworthy. As was pointed out previously, the literature suggests that parameter estimates should be trustworthy regardless of normality of residuals, whereas the trustworthiness of 95%CIs should be more influenced by the extent to which this assumption is met. Using this example, we have two analyses we can play with to demonstrate this issue. The first, the original analysis presented above, had some modest deviations from the Gaussian ideal distribution for residuals. The second, with seven cases removed, more closely met the

[4] For the reader's convenience, I have included SPSS syntax for performing a wide range of Box Cox transforms (users of other statistical software often have Box Cox transformations as part of the statistical package). Additionally, as it is desirable to explore the influence of a particular transform on the normality of the residuals, I have included Appendix B, which contains SPSS syntax for repeatedly performing a regression analysis on different transforms of a variable and then summarizing the normality of the residuals.

[5] When residuals are more than 3 SD from the mean, the probability of them coming from the population of interest is about 0.14%, which is *prima face* evidence that these cases are not representative of the population of interest. Removal tends to improve the accuracy of population parameter estimates (Osborne 2012). Furthermore, there are many different indices of influence, including DfBetas, Mahalanobis or Cook's Distance, etc.

assumption. Our expectation should be that: a) the parameter estimates should be relatively stable in both cases, but b) that the 95%CIs should be more "trustworthy" when we more closely meet the assumption of normality. In other words, the point I hope to illustrate is that most regression residuals will not be perfectly normally distributed, but by taking actions to *improve the normality of the residuals*, one can produce analyses that are more trustworthy.

These two data sets were each subjected to 10,000 bootstrap analyses to test the extent to which expectations A and B are met (as well as modeling an alternative method of empirically calculating 95%CIs when this assumption is not strictly met). With bootstrap analyses becoming more common, violations of assumptions (that might not be addressable by other means) might be addressed empirically by simulating thousands of bootstrap analyses and empirically generating confidence intervals (some good places to start exploring boostrap analyses are: DiCiccio & Efron, 1996; Efron & Tibshirani, 1994; Rodgers, 1999; Thompson, 1993) rather than relying upon calculated confidence intervals that might be untrustworthy. The results of the original analyses and the bootstrap analyses are presented in Table 1: the original regression predicting salary from faculty size, and the same analysis after the removal of 7 cases as detailed above.

Improving normality has little effect upon parameter estimates. Referring to Table 1, you can see that this expectation seems well-supported. As you can see in Analysis #3, the bootstrap analysis produced a point estimate that is very close to the original unstandardized regression coefficient from Analysis #1. Likewise, with Analysis #4, the bootstrap analysis closely replicated the original unstandardized regression coefficient from Analysis #2: 0.407 vs. 0.408, with slightly wider 95%CIs. This indicates that the initial parameter estimates in Analyses #1 and 2 are reasonable approximations of what a researcher might find drawing a different sample of similar size from a similar population.

Improving normality improves the trustworthiness of the confidence intervals. As reviewed above, one of the primary concerns regarding the non-normality of regression residuals (particularly in small samples) is trustworthiness of confidence intervals (e.g., Cohen et al., 2002). Although these samples are relatively large (N over 1000) and thus should be "robust" to violations of this assumption, the bootstrap analyses raise some interesting questions. For example, the regression residuals from Analysis #1 were not markedly non-normal (recall a skew of 0.50 and kurtosis of 0.89) but in the bootstrap analysis (Analysis #3, in Table 1) the empirical 95%CI is 0.121

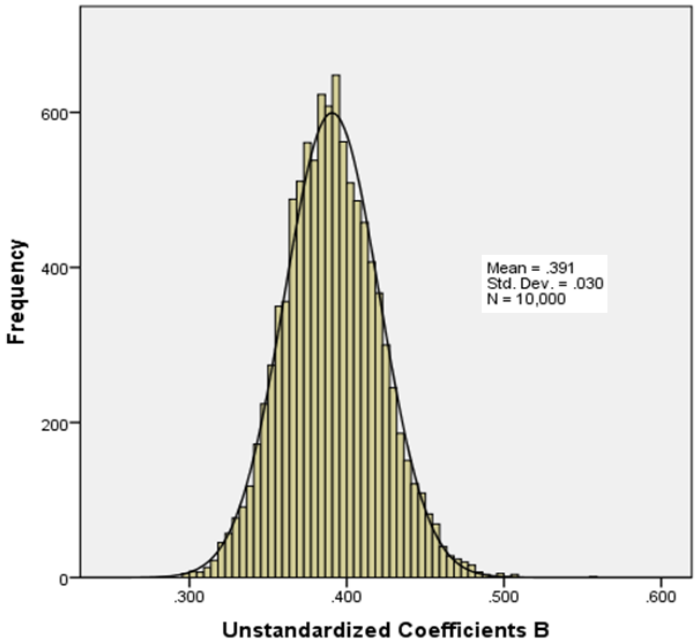

Figure 6: Distribution of unstandardized regression coefficients from original data. Skewed bootstrap analyses can be indicative of outliers – which are present in this data set.

in width as opposed to 0.081 from the original analysis (i.e., 49.38% larger). Removal of seven cases with relatively extreme residuals improved the normality of the regression residuals (skew= 0.34 and kurtosis =-0.06). Our expectation should be that the CIs should now be more trustworthy. Accordingly, the spread of the 95%CIs were smaller (0.079 for Analysis #2 and 0.106 for the bootstrap of that sample, Analysis #4). While the empirical CIs are still 35.18% larger than the calculated CIs, it was a closer match. Put another way, improving the extent to which our analyses met the assumption improved the extent to which the

calculated CIs matched the empirical CIs. Another index of trustworthiness of the calculated CIs was the percent of the bootstrapped parameter estimates that fell within the calculated CIs. When the residuals were less normal, fewer point estimates fell within the calculated CIs (82.7%) than when the residuals were more normal (85.7%).

Table 1: Comparison of parameter estimates before and after data cleaning, as well as from bootstrap analysis

	Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Analysis:	B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
NUM_AP	.389	.021	.488	18.761	<.001	.348	.429
NUM_AP	.407	.020	.514	20.004	<.001	.367	.446
Bootstrap 10,000 samples	.391	.030				.333	.454
Bootstrap 10000 samples post data cleaning	.408	.027				.358	.464

Predicting SAL_AP from NUM_AP. Analysis #1 has all cases. Analysis #2 has cases with standardized residuals > |3| removed, improving normality of residuals and parameter estimates. Analysis #3 is a 10,000 sample bootstrap of Analysis #1. Analysis #4 is a 10,000 sample bootstrap of Analysis #2.

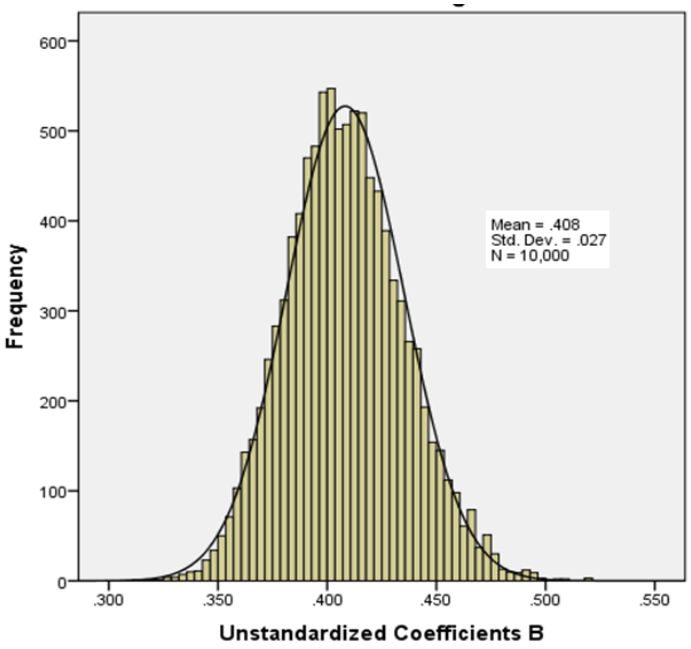

Figure 7: Distribution of unstandardized regression coefficients from Analysis #4

Conclusions

This simple example provides us with confirmation of several of the points from Williams et al. (2013) regarding the assumption of normal distribution of residuals in OLS regression, yet in the context of real data with continuous variables. First, the assertion that non-normality of residuals does not substantially bias parameter estimates is largely supported: improving the normality of the residuals via removal of several inappropriately influential cases altered the parameter estimate slightly but in each sample corresponded closely to the bootstrap estimates of the parameter. Secondly, it seems that bootstrap analyses indicate that the calculated 95%CIs are less trustworthy (even in relatively large samples) when this assumption is less well met. Conversely, when the assumption is more well met, the trustworthiness of the CIs improved.

Note that this is contrary to published guidance in that in large samples, this is supposed to be less of an issue. If one is to believe the value of bootstrap analyses, we might conclude that the calculated 95%CIs are under-estimated rather dramatically, even in large samples and even when residuals are relatively normally distributed-- particularly when outliers are present. This example, combined with that from Williams et al. (2013), underscores the importance of attending to assumptions, particularly in light of many organizations

(e.g., the American Psychological Association, journals in many fields) requiring or suggesting reporting of confidence intervals.

References

Aiken, L. S., & West, S. (1991). *Multiple Regression: Testing and Interpreting Interactions* Thousand Oaks, CA: Sage Publications.

Byrne, B. M. (2010). *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming.* New York, NY: Routledge.

Cohen, J., Cohen, P., West, S., & Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* Mahwah, NJ: Lawrence Erlbaum.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 189-212.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap* (Vol. 57): Chapman & Hall/CRC.

Nunnally, J. C., & Bernstein, I. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw Hill.

Osborne, J. W. (2002). Notes on the use of data transformations. *Practical Assessment, Research, and Evaluation., 8(2)*, Available online at http://pareonline.net/getvn.asp?v=8&n=2 .

Osborne, J. W. (2008). Sweating the small stuff in educational psychology: how effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices. *Educational psychology, 28*(2), 1 - 10.

Osborne, J. W. (2010). Improving your data transformations: Applying Box-Cox transformations as a best practice. *Practical Assessment Research & Evaluation, 15*(12), 1-9. Available online at http://pareonline.net/getvn.asp?v=15&n=12 .

Osborne, J. W. (2012). *Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data.* Thousand Oaks, CA: Sage Publications.

Osborne, J. W. (in press). *Best practices in logistic regression.* Thousand Oaks, CA: Sage.

Osborne, J. W., Kocher, B., & Tillman, D. (2012). *Many authors in top journals don't report testing assumptions: Why you should care and what we can do about it.* Paper presented at the Annual meeting of the Eastern Education Research Association, Hilton Head, SC.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research, and Evaluation, 9*(6). Available online at http://pareonline.net/getvn.asp?v=9&n=6.

Osborne, J. W., & Waters, E. (2002). Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research, and Evaluation, 8*(2). Available online at http://pareonline.net/getvn.asp?v=8&n=2.

Rodgers, J. L. (1999). The bootstrap, the jackknife, and the randomization test: A sampling taxonomy. *Multivariate Behavioral Research, 34*(4), 441-456.

Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education, 61(4), 361-377.*

Williams, Matt N., Grajales, Carlos Alberto Gómez, &Kurkiewicz, Dason (2013). Assumptions of Multiple Regression: Correcting Two Misconceptions. *Practical Assessment, Research & Evaluation*, 18(11). Available online: http://pareonline.net/getvn.asp?v=18&n=11

Appendix A

SPSS syntax to perform Box Cox analysis with expanded range over (Osborne, 2010), referenced at http://pareonline.net/getvn.asp?v=15&n=12

```
***BOX COX SPSS syntax.  Refer to http://pareonline.net/pdf/v15n12.pdf for
***information.  Anchor minimum value at 1.0 and change NUM_AP to name of
***variable you want to transform prior to running.
***  Examine TRANS frequency table to explore normality of transformations.  ***
LAM table tells you what lambda was used for each transformation in
***  TRANS table.

     COMPUTE var1=num_AP.
     execute.

     VECTOR lam(101) /tran(101).
     LOOP idx=1 TO 101.
     - COMPUTE lam(idx)=-5.1 + idx * .1.
     - DO IF lam(idx)=0.
     -   COMPUTE tran(idx)=LN(var1).
     - ELSE.
     -   COMPUTE tran(idx)=(var1**lam(idx) - 1)/lam(idx).
     - END IF.
     END LOOP.
     EXECUTE.


     FREQUENCIES VARIABLES=var1 tran1 to tran101
       /format=notable
       /STATISTICS= SKEWNESS KURTOSIS
       /ORDER=ANALYSIS.

     FREQUENCIES VARIABLES= lam1 to lam101
       /format=notable
       /STATISTICS= MINIMUM
       /ORDER=ANALYSIS.
```

Appendix B

SPSS syntax to run regression analyses using a variety of transformed variables. Macro syntax partially modeled on syntax found at Raynald's SPSS tools web site: http://www.spsstools.net/. In this syntax I performed analyses on a variety of transformed versions of NUM_AP that were reasonably normal (TRAN40- TRAN60 in this case). The macro also shows the skew and kurtosis of the residuals resulting from each analysis.

```
     DEFINE !regloop(nby=!TOKENS(1)).
     !DO !cnt=1 !TO !nby.
     REGRESSION
         /STATISTICS COEFF OUTS CI(95) R ANOVA CHANGE
         /DEPENDENT SAL_AP
         /METHOD=ENTER !CONCAT('tran',!cnt)
```

```
    /save resid.
!DOEND.
Frequencies variables=res_1 to !CONCAT('res_',!nby)
  /format=notable
  /statistics=skewness kurtosiss.
!ENDDEFINE.


*Call macro (replace 101 with something else if you use a different number of
transformations).
!regloop nby=101.
```

Acknowledgements and author notes:

The original article Williams et al. are responding to (Osborne & Waters, 2002) was inspired and produced through a close collaboration with Dr. Elaine Waters, who at the time was one of my doctoral students at the University of Oklahoma. I would like to acknowledge her valuable contributions to the original article. More importantly, I accept full responsibility for any errata or obfuscation that occurred in the original article.

Citation:

Osborne, Jason W. (2013). Normality of residuals is a continuous variable, and does seem to influence the trustworthiness of confidence intervals : A response to, and appreciation of, Williams, Grajales, and Kurkiewicz (2013). *Practical Assessment, Research & Evaluation*, 18(12). Available online: http://pareonline.net/getvn.asp?v=18&n=12

Author:

Jason W. Osborne,
Educational and Counseling Psychology, Counseling, and College Student Personnel
College of Education and Human Development
University of Louisville
Louisville KY 40292
Jason.osborne [at] louisville.edu