External Review of AMCI Report

Explanation

The report below is from an external reviewer (Thomas A. Gregory, PhD at University of Georgia). The body text is his direct review. The endnotes are the methodologist's (James Eric Gaskin, PhD at Brigham Young University) responses to each major comment.

Summary

This study evaluates revenue and expense differences between associations utilizing management companies and independently managed organizations (IMO), specifically IRS 501(c)(6) associations and IRS 501(c)(3) non-profit organizations. The analysis concluded that organizations served by association management companies (AMCs) demonstrated greater net revenue growth and net income growth between 2010–2012 compared with IMOs. Although these analyses were summarized as "not statistically significant" for identifying differences between groups in many cases (p. 5), the analysis did conclude a consistent trend existed, sufficient for drawing useful conclusions. I agree with this assessment.

Although the differences are often statistically uncertain, this is likely a consequence of small sample sizes and the analysis method used. Including additional organizations in the sample or utilizing different statistical methods (explained below) will most likely only serve to cement the study's conclusions.

Discussion

Data were suitably sampled from an appropriate population consistent with the requirements of the RFP. Financial data for the relevant time period (2010–2012) was extracted from IRS 990 forms for the sampled organizations. This archival data uses a trusted source and is likely to be reliable. Although the analysis describes removing outliers (p.5), this is not discussed in the data collection (p. 4). What effect removal of these outliers had on analysis is not described, but is likely minimal.[i]

For simple analyses, the t-test is a suitable method, and is robust against deviations from normality such as those identified in the samples. However, if the population, and thus, samples from the population, are assumed to be normally distributed (albeit slightly skewed), partitioning these samples into low/high revenue would virtually guarantee non-normal and highly skewed distributions in each partition. Author(s) describe utilizing Levene's homogeneity of variance test, which is appropriate, but the extent to which skewed or non-normal data affected the analysis is not revealed.[ii]

An alternative method, particularly given the identified skewness, might have been to utilize Wilcoxon rank-sum test (also known as the Mann–Whitney U test) or an analysis of variance (ANOVA)[iii] instead of the t-test. The latter in particular is often used for panel data.

The report describes the Mann–Whitney test as not a viable option due to differences in distribution shapes across groups (p.17). However, this comment is not itself persuasive as the differences in distribution shapes of the observed populations is not an assumption[iv] of the test. In addition, the Mann–Whitney test compensates for outliers, which the report indicates were a complication of the mean-based t-test. Further, it is not clear whether the differences that prevented use of Mann–Whitney refers to all partitions or just some. Despite this potential avenue of analysis, the repeated observations over multiple years (2010–2012) may have been more suitable for ANOVA.

Although outside of the scope of the RFP, I am left to wonder whether differences in grants paid (IRS Form 990, line 13), and fundraising expenses (IRS Form 990, line 16) contribute to the observed differences. However, I am given to understand that these are reported inconsistently between organizations (or not at all) and may not be suitable measurements for comparison.[v]

Another concern with the analysis is the drawing of conclusions for high *p*-values, especially those above 0.5. The p-value indicates the probability of a Type I error, namely, that a difference is incorrectly identified where none exists. The p-value typically falls as sample size increases, so some of this effect is due to the relatively small size of samples in each partition. However, for some of these results, the reported *p*-value is high enough that the likelihood of Type I error is closer to 1 than to 0. (Esp. *Liability Revenue (ratio)* in nearly all partitions*, Products/Services revenue (ratio)* in most partitions, 4.3 re *Net Total Revenue Growth* and *Net Income Growth*, to list some examples.) In no case is the likelihood of a Type II error reported, although that is typical with analysis such as these.[vi]

To conclude, despite these concerns I am persuaded that the differences described are likely to exist between IMOs and AMCs. Although the *p*-values are high, this is likely a consequence of small sample sizes. In addition, there is a consistent trend across all calculations, no matter the *p*-value, in support of the report's conclusion, which is confirmed by the consistent relative differences between mean and median values for the comparison of IMOs and AMCs.

—Thomas A. Gregory, PhD

End Notes (responses)

[i] Page: 1
To clarify, the outliers I removed were those outside of the $7.5 million budget range. There were a few (<5) that dealt with operating budgets of more than $7.5 million. In the report, we should not have labeled those as “outliers”. Instead they are simply “out of scope”.

[ii] Correct. The only other way this was addressed was by showing the difference between mean and median in revenue.

[iii] On reflection, we may have been able to leverage more power with ANOVAs rather than t-tests. The result may have been lower p-values in some cases.

[iv] This is Page: 1
correct (that it isn’t a strict prerequisite). Instead, it simply changes the potential of the test. If the distributions are different, then the MW test allows you to determine whether those distributions are statistically different. If the distributions are not different, then the MW test allows you to determine if the medians of those distributions are statistically different. It would have been prudent to run the test anyway to determine whether the distributions were statistically different. With the small sample size, we still probably would have observed no statistical differences. For the scientific community, this would have been absolutely necessary, but for the practitioner community, it may actually obfuscate the message.

[v] That is correct.

[vi] This is a good point. To improve the report, I should have indicated (maybe through color) which differences were significant at 0.500 or less, and which were not. Then also have a small discussion of type 1 and type 2 errors. Again, for the scientific community, this would have been absolutely necessary, but may serve only to distract the practitioner community.