Theme: Theory & Practice

OVERALL GOODNESS-OF-FIT STATISTICS IN ACADEMIC MARKETING RESEARCH

Work In Progress

Mark Uncles & Kelly Page

School of Marketing
University of New South Wales

Address for all correspondence:

Mark Uncles
Professor of Marketing
Head of School of Marketing
University of New South Wales
Sydney 2052
Australia

Phone: +612 9385 3385
Fax: +612 9663 1985

E-mail: m.uncles@unsw.edu.au
http://www.marketing.unsw.edu.au

OVERALL GOODNESS-OF-FIT STATISTICS IN ACADEMIC MARKETING RESEARCH

Work In Progress

ABSTRACT

A study of how overall goodness-of-fit statistics are used in academic marketing research suggests there are ways for improving the presentation of these statistics, for seeing them as essentially contingent on other statistical and non-statistical considerations, and for developing a broader view of what is meant by a "good fit".

AIMS AND OBJECTIVES

The aim is to:

(1) (re-)consider the role of overall goodness-of-fit statistics in academic marketing research.
(2) systematically and empirically assess current practices in academic marketing research.
(3) discuss possible explanations for, and implications of, these practices.
(4) consider alternative methods and approaches for assessing overall goodness-of-fit.

Many of these issues were considered in a paper by Sawyer and Peter (1983), but it appears analysts have not taken heed of their conclusions over the intervening years.

BACKGROUND FROM THE PERSPECTIVE OF ACADEMIC MARKETING RESEARCH

In conventional regression the overall fit of a model is measured by the R^2 statistic (or the coefficient of determination). This measures the proportion of the variation in the response variable accounted for or "explained" by the independent variables, and it ranges in value from 0 to 1. The closer R^2 is to 1, the better the fit of the model - ie. the better the independent variables are at accounting for or explaining the variation in the response variable. A development of the standard R^2 is adjusted-R^2; this may rise or fall with the introduction of new variables, and therefore at least partially eliminates the incentive for analysts to include numerous variables in their regression models.

We are also interested to know whether the relationship between the independent variables and response variable is statistically significant. That is, we want to compare the regression-associated variance with the error variance and assess whether the regression model has explained a statistically significant portion of the variance in the dependent variable. An F statistic is used for this purpose, with k-1 and N-k degrees of freedom, for a certain confidence level.

All elementary statistics and marketing research texts discuss these topics. If this conventional treatment of regression is accepted then the converse - of having a low R^2 - should be of at least some concern because:

- Doubts arise as to the *predictive use* of regression models. This goes to the heart of regression modelling given that prediction is a key objective, and without accurate prediction it is hard to envisage meaningful explanation.

- It also means a lot of unexplained variation remains in the dependent variable. This may be "random" variation of no substantive interest. Or it may reflect variation that could be explained by potentially important omitted variables. We just don't know.

- The problem here is not only that of ignoring potentially important variables through model mis-specification, but that the reported relationship might itself be jointly determined by another underlying variable.

Despite these concerns, reported R^2s range from .00 to .99 (rounded to two decimal places). Some of the reasons for this are well documented. In time-series, where the environment is quite stable, R^2s are routinely as high as .90. Whereas in changing environments, R^2s may fall to .20. A difference is also seen between models where the aim is to predict aggregate measures (sales, market shares, etc.) and those where it is to predict individual behaviour or attitudes - the latter being prone to low values of R^2.

We appear to face two rather different images: one image is of something quite fundamental to our modelling efforts, and another image is of something that is used in a rather "flexible" way. Indeed, there appears to be much ambiguity about the role and purpose of overall goodness-of-fit statistics like R^2 or adjusted-R^2 in academic marketing research.

Therefore, how reliant should we be on these statistics? Are there grounds for legitimate doubts about the appropriateness of these tests, or do we simply need to pay more attention to the modelling context (e.g. the effects of sample size)? If the modelling context is of such importance is this properly reflected in the way research is reported in the discipline?

THE CONVENTIONAL VIEW OF OVERALL GOODNESS-OF-FIT STATISTICS

There are several distinct viewpoints on the status of overall goodness-of-fit statistics (some of which appear to be incompatible).

Statistical Arguments

- There is more to modelling than maximising R^2 or adjusted- R^2. Moreover, there is no hard-and-fast statistical argument for deciding what level of R^2 is "high enough".

- A somewhat different angle on this topic is the suggestion that marketers have become so much more sophisticated in conceptualising and modelling marketing phenomena they are now far more prepared to rely on theory than

purely statistical conventions. In essence, theory (be it conceptual or empirically-grounded) is deemed to matter more than "mere statistics".

Those who put forward the preceding arguments are essentially asking for a balance to be kept between "the science" and "the art" of model building, but they do not question the use of regression as a tool of analysis. However, there are those who have more fundamental concerns about the appropriateness of regression and - by extension - the relevance of overall goodness-of-fit testing:

- For R^2 to be calculated there needs to be meaningful variation in the response variable, such that there is something to be accounted for, or "explained", by the independent variables. Yet, in many competitive marketing situations this may not be the case (e.g. one company's promotional activity might be neutralised by the activities of competitors).

- There is also more general concern with "the ritual of null hypothesis significance testing", such that the American Psychological Association Board of Scientific Affairs has been considering changes to the requirements for statistical significance testing.

- Another criticism is that values of R^2, and of statistical significance, are considered in isolation - one data set at a time. This relates to issues of replication and empirical generalisation to which we will return later.

In-use Arguments

- A positive aspect is that marketers increasingly have access to very large data sets (e.g. scanner panels containing records for thousands of households over extended periods of time). With such very large data sets the probability of sampling errors fades, and so too do the problems of statistical inference. Relationships - to the extent that they are observed - must really be there in the population. Unfortunately research and teaching practices in marketing continue to rely on statistics designed for small isolated studies. An historical perspective on this is of help. Student, and subsequently Fisher, were concerned with small experiments or small sampling schemes. They needed to extract maximum information from these limited data, however these men were fully aware of the role of their isolated experiments, whereas subsequent researchers appear to have been less aware of the scope of their work.

- A more questionable argument is based on the premise that the procedures recommended in standard texts bear no relationship to what is done in marketing research practice. Moreover, there are entirely appropriate and understandable reasons for us to accept the approach of practitioners, because of such considerations as cost constraints (which limit what can be achieved) and non-sampling errors (which out-weigh any statistical failings). This is a

slippery path - it lets academics "off the hook" by caricaturing market research practice, and in so doing may mislead both groups of analysts.

- A even less generous view is that academic marketers are disregarding the problem of overall fit out of confusion, a problem confounded by the fact that many analysts are unclear whether their goal is to predict or explain phenomena.

- A variation on this theme is the practice of listing the limitations, but then proceeding with the use of regression analysis and, for the sake of convention, reporting values of test statistics. This amounts to "hedging your bets" and probably arises as an unwitting outcome of the review process.

It is evident that some of the statistical and in-use arguments raise ambiguities, not all of which can be simultaneously upheld. Some point to greater sophistication as the reason for less reliance being placed on overall goodness-of-fit, while others suggest the issue has more to do with a relaxation, or disregard, of statistical conventions.

CURRENT PRACTICES IN ACADEMIC MARKETING RESEARCH

To undertake a systematic and empirical assessment of current practices the marketing literature was surveyed following similar procedures to those used in content surveys on such themes as null results in marketing, goodness-of-fit indexes in confirmatory factor analysis, and the assessment of causal models in marketing.

Surveyed were a range of A-grade refereed journals where quantitative work is reported (JM, JMR, IJRM, and Marketing Science), and one set of refereed A-grade conference proceedings (EMAC). Attention is confined to refereed A-grade publications on the grounds that any statistical problems ought to have been clarified or resolved through the review process.

Among these sources, the number of papers using statistical techniques was calculated. To qualify, descriptive statistics such as means and standard deviations were ignored, but the use of all higher-level statistics was noted (eg. regression coefficients, and the use of inferential tests such as z-tests, t-tests, Chi-square tests, etc.). Focus was then placed on conventional regression.

The classification can be further refined by distinguishing the following:

1. Cross-sectional regression models versus time-series models.
2. Ordinary least-squares, from weighted least-squares, from stepwise, etc.
3. Aggregate predictions versus individual predictions.
4. Fit statistics for the estimation sample versus results for the prediction sample.

Our initial assessment of the data did not consider these refinements, but in further work we are addressing these aspects. It is true to say that most of the models we looked at were based on cross-sectional data, using least-squares estimation, and the fit statistics typically referred to the estimation sample.

DISCUSSION OF CURRENT PRACTICES IN ACADEMIC MARKETING RESEARCH

Responses to the current practices can be grouped into two main categories: (1) those that attempt to deal with problems of overall fit within the framework of conventional inferential statistics, and (2) those that call for a completely different view of modelling and where "fit" takes on a new meaning.

Responses Within the Framework of Conventional Inferential Statistics

- If goodness-of-fit measures are to be reported, then at the very least the information to judge the fit should be reported clearly (including the number of observations, number of variables, value of R^2 or adjusted- R^2, F-statistic, and significance level). Empirical analysis shows that this very basic request is quite commonly violated. Further, even where the information is contained in the paper it is often hard to find.

- Another response is to fit more robust regression models and pay far more attention to regression diagnostics - outliers etc. While techniques of robust regression have been known for several decades and feature in textbooks used in the research training of marketing analysts, there are very few examples of their use in academic papers. Our survey revealed hardly a single example over a four year period in any of the leading quantitatively-oriented journals.

- Yet another response is to have an overall measure which describes the predictive power of the model in the face of new data. We note that the "new data" in this instance are typically hold-out samples. Often this is a very weak test - the check is being undertaken on largely the *same* data (exactly the same data if the hold-out sample has been randomly selected from the original data set). This "more of the same" testing differs from prediction to specially designed test samples or prediction samples (e.g. a model based on national data checked against regional figures) and differs again from tests on differentiated data set (we develop this next).

Responses Beyond the Framework of Conventional Inferential Statistics

- It has been suggested that the better overall measure is one that describes the predictive power of the model in the face of truly new data. This requires the analyst to shift away from isolated studies based on single sets of data, and

instead to work with many sets of data, or prepare research designs that will allow for differentiated replication (Lindsay and Ehrenberg 1993). An example would be the collection of data about phenomena of interest across different product markets. A systematic way to think about this is by sampling across the whole data cube. This is not the same as meta-analysis, which is a *post hoc* approach and which invariable means comparing "apples with pears".

- This approach raises some issues that go well beyond goodness-of-fit tests, and lie at the heart of modelling in marketing. In essence, we are asking what is meant by "good". Is it a "specifically-best fit" (such as we obtain from a statistical significance test) or a "generally-good fit" (where conventional tests would appear to be deficient). What we want is a measure of significant *sameness* across a range of conditions (thereby establishing the scope and boundary conditions of our model). The obstacle here is that we lack guidance from statistics. Moreover, most textbooks in marketing research show little recognition of the subject. It is not surprising therefore that academic marketing research practice largely ignores the topic. The pages of JM, JMR, etc. are virtually devoid of any examples.

CONCLUSIONS

Marketing is increasingly a models-based discipline. Empirical papers in leading journals rely heavily on statistical techniques - almost 60% of papers in leading journals rely on these techniques, and over 20% of all papers make use of regression. Marketing research practitioners are equally reliant on such methods. Even marketing managers are now making routine use of decision-support packages, which rely on underlying statistical and quantitative techniques. For these reasons it is all the more important for us to have confidence in the appraisal of our models.

What we have shown in this paper is the need to openly discuss and debate the question of overall goodness-of-fit statistics and significance tests. Also, more serious consideration should be given to alternative procedures. Options range from the use of robust regression techniques to checks across differentiated conditions. It is not as if these techniques are wholly unknown in marketing, but they are uncommon in academic practice and are remarkably absent from the research literature - despite very clear position papers (notably Sawyer and Peter 1983). Few of the papers in our survey even mentioned these alternatives, and virtually none made use of them.

ACKNOWLEDGEMENTS

Support from a Faculty of Commerce & Economics (UNSW) SRG Grant is acknowledged. A full-length version of this paper, with complete references, is available from the authors.

REFERENCES

Lindsay, R.M. & Ehrenberg, A.S.C. (1993) "The design of replicated studies" *The American Statistician*, 47 (3), 217-228

Sawyer, A.G. & Peter, J.P. (1983) "The significance of statistical significance tests in marketing research" *Journal of Marketing Research*, 20 (May), 122-133.