2012

Using Baseline Models to Improve Theories about Emerging Markets

Andreas Schwab
Iowa State University, aschwab@iastate.edu

William H. Starbuck
University of Oregon

USING BASELINE MODELS TO IMPROVE THEORIES ABOUT EMERGING MARKETS

Andreas Schwab
3315 Gerdin Business Building
Iowa State University
Ames, IA 50011-1350
Phone: (515) 294-8119
aschwab@iastate.edu

and

William H. Starbuck
420 Lillis Hall 1208
Lundquist College of Business
University of Oregon
Eugene, OR 97403-1208
Phone: (541) 346-0751
starbuck@uoregon.edu

Published in

Research Methodology in Strategy and Management

Reference: Schwab, A. & Starbuck, W. H. (2012). Using Baseline Models to Improve Theories About Emerging Markets. In C. Wang, D. Bergh, & D. Ketchen (eds.), *Research Methodology in Strategy and Management*, Vol. 7, 3-33. Bingley, UK: Emerald.

USING BASELINE MODELS TO IMPROVE THEORIES ABOUT EMERGING MARKETS

ABSTRACT

Purpose

This chapter reports on a rapidly growing trend in the analysis of data about emerging market economies – the use of baseline models as comparisons for explanatory models. Baseline models estimate expected values for the dependent variable in the absence of a hypothesized causal effect, but set higher standards than do traditional null hypotheses tests that expect no effect.

Design/methodology/approach

Although the use of baseline models research originated in the 1960s, it has not been widely discussed, or even acknowledged in the emerging markets (EM) literature. We surveyed published EM studies to determine trends in the use of baseline models.

Findings

We categorize and describe the different types of baseline models that scholars have used in emerging markets studies, and draw inferences about the differences between more effective and less effective uses of baseline models.

Value

We believe that comparisons with baseline models offer distinct methodological advantages for the iterative development of better explanatory models and a deeper understanding of empirical phenomena.

USING BASELINE MODELS TO IMPROVE THEORIES ABOUT EMERGING MARKETS

At the 1967 annual meeting of the American Agricultural Economics Association, agricultural economists debated methodological issues that remain troublesome in the emerging-markets research of today. Their debate concerned the usefulness of conventional null hypotheses and alternatives to them.

Wise and Yotopoulos (1968) argued that Greek farmers use labor and capital "rationally". Wise and Yotopoulos stated a hypothetical relationship between labor, capital, and output and inferred that the logarithms of labor, capital, and output should relate linearly to each other if farmers do indeed behave rationally. If farmers are not rational, they said, the regression calculation should yield coefficients for labor and capital that equal zero, which are two conventional null hypotheses about coefficients in multiple regressions. Based on their tests of statistical significance, they (1968: 396) estimated that "approximately two-thirds of the total variance in the inputs of capital and labor can be explained by variations in the profit-maximizing component of these inputs. . . . we find that the observed relationship between the inputs and the output is consistent with the hypothesis of profit maximization in somewhat imperfect factor markets."

When discussing Wise and Yotopoulos' study, Johnson (1968) protested that the null-hypothesis tests gave no support for their claim about rationality. Johnson observed that the null hypothesis of zero correlation would almost certainly be rejected when tested against "almost any other alternative". He (1968: 398) said: "In accepting this conclusion as empirical verification of their measure, one should note carefully the alternative that is rejected. By attributing the observed correlation to their systematic component, they thus

reject a model that considers entrepreneurial behavior to be entirely random. A model that makes entrepreneurial decisions unpredictable in any systematic fashion is an unlikely candidate for acceptance over almost any other alternative. “ Johnson (1968: 398-399) next proposed an alternative comparison model: “Consider a model that said all Greek farmers were taught as young men to apply so much capital and so much labor per unit of land. . . . A regression of capital on labor should show a high degree of correlation. If we ran such a regression in logs, we should get not only a high degree of correlation but also a coefficient that is an estimate of one. . . . The test, then, cannot distinguish between a purely traditional and a purely profit-maximizing behavioral model.”

In responding to Johnson's critique, Yotopoulos and Wise (1969a: 203) said, “Testing is a process of discriminating among competing hypotheses. A test by confirming a hypothesis rejects one or more alternative hypotheses -- be it the null hypothesis, a naive hypothesis or a serious alternative hypothesis. In our case, a conceivable null hypothesis is that entrepreneurs behave in an entirely random way. Rejecting this hypothesis is hardly an impressive record for our test, as Professor Johnson justly observed in discussing our paper [Johnson, 1968]. Moreover, Johnson suggested that the results of our test, the log-linear relationships between inputs and output, are consistent with the "traditional behavior" hypothesis in which "all Greek farmers were taught as young men to apply so much capital and so much labor per unit of land" [Johnson 1968, p. 398]. Our test, then, does not discriminate between the profit-maximization hypothesis and the "traditional" or fixed-proportions hypothesis." Yotopoulos and Wise (1969a) then proposed yet another comparison model, which provoked further disagreements from Johnson (1969) and Lianos (1969). In reply, Yotopoulos and Wise (1969b: 210) said: “Still, however, the sifting

and winnowing of competing hypotheses need not end with the results of any single test. More tests may further narrow the set of acceptable hypotheses. Also, as a last resort, the literature on economic methodology suggests auxiliary criteria for ranking hypotheses which are all consistent with the facts: the realism of the assumptions; the simplicity, fruitfulness, and elegance of the hypothesis; its consistency with other theory; etc."

Thus, the economists in this debate articulated two important ideas: Firstly, researchers should compare their preferred theories with alternative models, not with conventional null hypotheses. The debaters agreed that conventional null hypotheses offered no serious competition for a proposed theory, but they argued about what kinds of alternative models would provide a more useful comparison. Secondly, researchers should treat each analysis as a step in a sequence of analyses, not as a final judgment. As a component of a sequence, a comparison model should suggest useful successors to itself.

This chapter follows up these two ideas by reporting on a rapidly expanding trend in the analysis of data about emerging markets economies – the use of baseline models as comparisons for explanatory models. Although this trend began in the 1960s, it has not been widely discussed, or even acknowledged. We, the authors, discovered the trend somewhat accidently, as we did not know that it existed when we began our research for this chapter.

The trend involves comparisons between researchers' explanatory models and alternative models, which we call "baseline models". These alternative models are typically simpler than explanatory models and they typically make no assumptions about causation.

We believe that comparisons with baseline models offer a distinct methodological improvement over tests of conventional null hypotheses.

Because they are emerging innovations, baseline models are heterogeneous and their users have adopted diverse terminologies. The heterogeneity and diversity may continue to grow as researchers explore more kinds of baseline models. There is reason to conjecture that baseline models become more useful when researchers tailor them to specific studies. However, as we describe later, most of the baseline models used to date fall into a small number of general categories.

The next section of this chapter explains why we began to search for these baseline models. We believe that our reasons for searching probably shed light on the motivations of researchers who have been creating these models. The ensuing sections then describe the kinds of baseline models we found in studies of emerging markets economies and draw inferences about the differences between more effective and less effective uses of baseline models.

WHY WE SEARCHED FOR BASELINE MODELS

It is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail.

A. H. Maslow (1966: 15–16)

A series of experiences over many years has revealed various ways that statistical analyses can produce misleading results. We tell this story mainly from Starbuck's viewpoint, as he migrated gradually from teaching conventional statistics toward calling for methodological changes.

For Starbuck, the story began in 1966 or 1967, when he read a mathematical proof that a researcher who obtains a large-enough sample size can reject *any* point null hypothesis. As sample size increases, the confidence interval around the sample estimate shrinks and becomes less and less likely to include the hypothesized infinitesimal point. One consequence is that even trivial differences or differences based on measurement errors can lead to statistically significant findings. A second consequence is that statistical significance depends on researchers' motivation – their willingness to continue collecting data until they have enough. In general, researchers face a dual challenge of discovering substantively important relationships while also screening out relationships that have no importance. Conventional null-hypothesis significance tests set thresholds that are much too low. Consequently, hypothesis tests label as "statistically significant" many findings that have no substantive importance or that only describe idiosyncrasies of specific data.

Although Starbuck did not know this, researchers had discovered the sensitivity of statistical significance to sample sizes as early as the 1930s (Berkson, 1938). Since then, researchers have debated this and many other problems associated with the use of conventional null-hypothesis significance tests (Cohen, 1994, Edwards and Berry, 2010, Gigerenzer, Krauss and Vitouch, 2004, Greenwald, 1975, Rodgers, 2010, Schmidt and Hunter, 1997, Schwab et al., 2011, Seth et al., 2009, Thompson, 1999). Because contemporary computer-based data management and data analysis facilitate large samples, modern technology has amplified these problems by helping researchers to convert trivial and meaningless differences into statistically significant findings. In the absence of rigorous replication of research findings, it is incredibly difficult to assess the frequency of false positives in research studies.

Some Reasons Why False Positives Are Prevalent in Empirical Studies

Conventional null hypotheses are especially problematic for analyses of time series. Two of Starbuck's colleagues, Ames and Reiter (1961) showed that autocorrelations in economic time series make it extremely likely to find statistically significant correlations among variables even in the absence of any direct causal relationships between them. Ames and Reiter found that five-sixths of the series in the *Historical Abstract of the United States* have autocorrelations above .8 for a one-year lag. These autocorrelations cause series to correlate with each other. Ames and Reiter observed that starting with randomly chosen target time series and then choosing a second series at random, an economist on average would need only three trials to discover a correlation of .71 between the two series.

Two decades later, Peach and Webb (1983) demonstrated a practical implication of spuriously significant correlations among economic time series. They created hypothetical 'models' by selecting random combinations of a dependent variable and three independent variables, which they analyzed as if these were macroeconomic models. They found that 64 percent to 71 percent of the independent variables had 'statistically significant' coefficients. With a simple linear model, 64 percent of the combinations of variables had R^2s over 0.99 and 94 percent had R^2s over 0.95. Even when they converted all variables to first differences, which greatly reduces autocorrelation, 32 percent of the combinations of variables had R^2s over 0.60 and 10 percent had R^2s over 0.80.

The foregoing studies induced Webster and Starbuck (1988) to investigate the possibility of spurious correlations in cross-sectional data. They compiled nearly 15,000

correlations published in three prominent management journals. These data included all of the correlations among studied variables, not only the correlations in hypotheses. The correlations had very similar distributions in all three journals, and the distributions for small samples were quite similar to those for large samples. Both the mean and the median correlations were close to 0.09 and 69 percent of the correlations were positive. As well, researchers had tended to obtain larger samples when they found correlations with smaller absolute values. These weak correlations form a background of meaningless or substantively unimportant correlations that researchers may mistake for significant relationships, especially when they have large samples. Webster and Starbuck simulated what would happen if researchers would search randomly for statistically significant correlations in such data. Researchers' would have 2 to 1 odds of coming up with a statistically significant correlation on the very first try, and 24 to 1 odds of discovering a statistically significant correlation within three tries.

Webster and Starbuck pointed to several reasons for the existence of confounding background correlations. Two of these reasons have relevance to studies of emerging markets economies. Firstly, researchers are intelligent, observant people who exhibit foresight about their findings. They are likely to have intuitive understanding of people and of social systems; they formulate hypotheses that are consistent with their intuitive understanding; they are quite likely to investigate correlations and differences that deviate from zero; and they are less likely than chance would imply to observe correlations and differences near zero. Secondly, a few broad characteristics of people and social systems link social-science data. For example, intelligence correlates with many other characteristics and behaviors, such as leadership, job satisfaction, job performance, social

class, income, education and geographic location during childhood. These correlates of intelligence tend to correlate with each other, even if there are no direct causal relations among them, because of their common relation to intelligence. Other broad characteristics that correlate with many variables include sex, age, social class, education, group or organizational size, geographic location, culture, affluence, and political system.

A Brief Digression: False Positives in Emerging Markets Research

We replicated part of Ames and Reiter's (1961) investigation of autocorrelations with data about emerging markets. We randomly selected 40 time series from the World Bank's world-development indicators; the series included 20 countries that are in the MSCI emerging-market index (World Bank, 2011). Figure 1 shows the estimated absolute values of the autocorrelations for various lags, and Figure 2 shows the autocorrelations for absolute values, natural logs of values, and changes in values.

These findings leave no doubt that the studies by Ames and Reiter (1961) and Peach and Webb (1983) apply to emerging-markets research. If a researcher randomly selects two time series for the same country, then regresses the current values of one on the lagged values of the other, there is a 43 percent probability that lagged variable will appear to explain more than 50 percent of the variance in the current variable. If the researcher repeats this process three times, the chance of finding such a highly correlated pair increases to 82 percent.

We also replicated the study by Webster and Starbuck (1988) with data about emerging markets. We identified all published research studies on ISI Web of Knowledge for the years 2007 to 2011 in the fields of business and management research that had

titles containing 'emerging market', 'emerging country,' 'developing country' or derivations of these terms. We identified 43 quantitative empirical studies that had reported a total of 3132 simple correlations. The mean correlation was +0.08 and the median +0.05 (see Figure 3).

Assuming that these correlations approximate the population of correlations that emerging-markets researchers typically encounter, we can estimate the probabilities for finding statistically significant results. Using the median sample size in the actual studies (n = 232), researchers have a 45 percent chance of finding a statistically significant correlation if they randomly draw one correlation from the population of reported correlation coefficients. For two or three random draws, the probability of finding a statistically significant correlation increases to 70 percent and 83 percent respectively. With a sample size of 896, which is an average that omits the two largest samples, the probabilities of finding a statistically significant correlation rise to 60 percent for a single draw, and to 99 percent for three draws. These simulations confirm that conventional null hypotheses set a low threshold for the identification of important findings, and these simulations support concerns about a high proportion of false positive results in emerging-markets research.

Two Proposals for Raising the Threshold

The single most important limit of null-hypothesis testing is that there is only one statistical hypothesis -- the null, which does not allow for comparative hypotheses testing. . . . Only when one knows extremely little about a topic (so that one cannot even specify the predictions of competing hypotheses) might a null-hypothesis test be appropriate.

Gigerenzer et al. (2004: 397)

Bioecologists began to question the usefulness of conventional null-hypothesis statistical tests in the early 1980s. In particular, Connor and Simberloff (1983, 1986) argued that conventional null-hypothesis tests have no value because interactions within ecological communities make simple null hypotheses very unrealistic. They proposed that bioecologists should replace null hypotheses with "null models", which they (1986: 160) defined as non-causal distributions for the studied variables. These discussions in bioecology supported the usefulness of comparing explanatory hypotheses with models that are more complex than conventional null hypotheses.

For example, the Galapagos Islands have different numbers of species of land birds. These numbers have fed debates about whether they reflect competition between species, physical differences among the islands, or vegetation differences. Connor and Simberloff (1983) used a "null model" to estimate the co-occurrence patterns of species of birds on each island. They assumed that each island had the number of species observed on it and that each species inhabited as many islands as observed, but that species were otherwise distributed randomly and independently. All the numbers of species pairs actually observed lay within two standard deviation confidence limits, and most observations lay close to the expected values. From this comparison, Connor and Simberloff concluded that competition between species has had little effect on the actual numbers of species pairs.

Although Starbuck found the notion of null models interesting but his personal experience led him to prefer a different alternative. He had been teaching a course in forecasting, which made him aware that forecasting researchers normally compare their

forecasts with "naïve hypotheses" instead of null hypotheses. One of these naïve hypotheses — no-change in value — says that the next value in a series will be the same as the current value. Another naïve hypothesis — no change in trend — says the next value will differ from the current value by the same amount that the current value differs from the previous value. These hypotheses are "naïve" insofar as either of them could come from a young child who has no understanding of the causal processes that generate a series. Forecasting researchers argue that a complicated forecasting technique should be able to make more accurate predictions than these naïve hypotheses. Thus, Starbuck (1994) proposed that researchers ought to compare their preferred theories with naïve hypotheses instead of with null hypotheses. Although naïve hypotheses set very modest thresholds for performance, they set higher thresholds than do conventional null-hypotheses.

However, the bioecologists and Starbuck were actually latecomers to these discussions. As we described in the introduction, agricultural economists had already been discussing these issues in the 1960s. Indeed, we have found examples of researchers using baseline models throughout the social sciences, economists being among the earliest experimenters with these methods. However, this paper focuses on studies related to emerging markets and developing countries.

HOW WE SEARCHED FOR BASELINE MODELS IN EMERGING MARKETS RESEARCH AND WHAT WE FOUND

The experiences described above suggested two kinds of baseline models: null models and naïve hypotheses, so these terms became the starting points for our

investigation of the use of baseline models. We had no idea that we would find a trend. We were merely curious whether emerging-markets researchers had used these terms, or variants of them, and if so, in what ways.

According to the data in Google Scholar, emerging-markets researchers have been using the terms null models and naïve hypotheses with escalating frequency. Figure 4 graphs the frequencies of the phrases "naïve hypothesis," “naïve model,” or “null model” in papers that also mentioned "developing countries" or "emerging markets". Researchers who study emerging markets economies made little use of baseline models before 1995 but usage has been growing rapidly since then.

Of course, searches of Google Scholar based on a few specific terms are crude instruments. Some researchers, for example, have used the term “naïve" to denote hypotheses or models that (1) the public at large might use, (2) might arise from a lack of information, (3) prior studies disconfirmed, or (4) misguided competing researchers proposed. Some researchers have used the phrases "naïve hypothesis," “naïve model,” or “null model” without actually making comparisons with their explanatory models. A few researchers have even characterized conventional null hypotheses as "naïve hypotheses" or “null models”. However, we followed up the Google searches by examining hundreds of specific papers, and did not find a time-linked pattern that would cast doubt on the qualitative profile in Figure 4. If anything, Figure 4 likely understates the acceleration in the use of baseline models, as various terms have emerged over time to supplement the older ones. For instance, in longitudinal or time-series studies, some researchers have described random walks as null models, but other researchers have compared their

explanatory models with random walks without characterizing the random walks as null models, and this latter practice appears to have become more prevalent over time.

This paper applies the label "baseline models" broadly to all of the alternatives to conventional null hypotheses that researchers use as bases for comparison with explanatory models. Baseline model seems to be a better label than others because it highlights the idea of a platform for comparisons with an explanatory hypothesis.

For clarity, we propose that a baseline model predicts an effect on a dependent variable in the absence of the effect hypothesized by researchers' explanatory model. The effect predicted by a baseline should also differ from the no-effect hypothesized by a conventional null hypothesis. For example, some studies have used the terms null model, naïve model, or naïve hypothesis to describe assumptions of homogeneity across countries or periods. However, the assumptions used to create such models are equivalent to conventional null hypotheses of no difference or zero effect, and they suffer from the same liabilities. Sufficiently large samples will inevitably demonstrate that differences exist, but this demonstration will say little about the substantive relevance of these differences between countries or periods. Thus, we do not regard these simple homogeneity models as baseline models.

Conventional null-hypothesis tests propose no effect for all phenomena in all empirical contexts (Gigerenzer et al., 2004). In contrast, users of baseline models adjust assumptions and evaluation approaches to specific phenomena and empirical contexts. These adjustments can take many forms, such as accounting for random processes other than random sampling, effects of third variables, or non-linear underlying effects.

Consequently, baseline models are diverse. The next section categorizes them, but these categories only describe what researchers have done to date, and future research will likely introduce additional kinds of baseline models and create new categories. Diversity and inconsistency in terminology and usage are natural and desirable during the creation of a methodological innovation. Researchers are groping for better ways to develop understanding. Some of their methodological innovations will work well and some poorly.

Distinctions between baseline models and explanatory models are not always clear. As an abstract principle, baseline models can serve as benchmarks to help researchers evaluate the effectiveness of explanatory models and to suggest ways to improve explanatory models. Some researchers, however, blur these distinctions and leave it unclear whether they intend models to be baselines or proposed explanations. As well, some researchers do not actually propose explanatory models but use baseline models solely to highlight properties of data. We argue later that a grey zone between baseline models and explanatory models is not surprising and that it can support meaningful model improvement.

Based on the broad definition of baseline models introduced above, we investigated the actual use of baseline models in specific emerging markets studies, and developed categories for alternative types of baseline models. Because they only describe what researchers have done and we have found, the categories are a bit untidy. Some of them overlap. The next section introduces these categories and describes some exemplar studies. The ensuing section concludes this chapter with some thoughts on the value and promise of baseline models for emerging markets research.

EXAMPLES OF BASELINE MODELS

Table 1 categorizes the baseline models that we have found in studies about emerging markets and developing countries. Some of these categories offer rather rigorous challenges to theorists. The following subsections give illustrative examples of these categories. However, studies do not always fall cleanly into a single category, mainly because researchers' assumptions about probability distributions are sometimes ambiguous and overlapping.

Log-Linear Scaling

So, how should researchers compare small countries with large ones? How does a small country change when it grows larger? Many researchers deal with such issues multiplicatively: if the inputs double, the outputs should double . . . approximately.

One widely used version of this idea is the Cobb-Douglas production function, which takes the form:

$$Q = AL^{\alpha}C^{(1-\alpha)}$$

Where Q is the total output (or consumption) of an economy, L is the input of labor (or employment), C is the input of capital, and A and α are constants. The constant A is usually called "total factor productivity" and interpreted as an indicator of technological effectiveness. One useful property of this function is that logarithmic transformations yield a linear function:

$$\log Q = \log A + \alpha \log L + (1- \alpha)\log C$$

First introduced in the mid 1800s, the Cobb-Douglas function has been used in many, many theoretical and empirical analyses. For example, Mastromarco (2008) looked at the effects of foreign direct investment, imported capital goods, and human capital on the productivity of 57 developing countries over 40 years. During this analysis, she used the Cobb-Douglas function as a baseline to compare with the more complex transcendental-logarithmic production function. She compared the baseline model

$$\log Q = \beta_1 \log L + \beta_2 \log C$$

with the transcendental-logarithmic model by testing the point null hypotheses that β_3, β_4, and β_5 all equal zero

$$\log Q = \beta_1 \log L + \beta_2 \log C + \beta_3 (\log L)^2 + \beta_4 (\log C)^2 + \beta_5 \log L * \log C + \text{Error}$$

Of course, no one should have been surprised that she could reject these null hypotheses. Thus, by the logic of conventional null-hypothesis tests, the complex function was more effective than the simpler baseline model.

Traditional Behavior

Many researchers have used concepts about "traditional behavior" as baseline models and they have used traditional behavior with both cross-sectional and longitudinal data. Some studies have obtained observational data about traditional behavior, but many studies have not, they have merely ascribed various characteristics to traditional behavior. It has also been popular to contrast "traditional behavior" with "rationality." The notion that traditional behavior differs from rational behavior dates back at least to 1914 when Weber distinguished among four "ideal types" of action. Weber defined these ideal types as (a) emotional action, when people act on their emotions without consciously thinking

through the results, (b) traditional action that conforms to customs and habits, (c) instrumental action, when people make careful calculations about goals to pursue and means to attain these goals, and (d) value-rational action, when people pursue highly valued goals, possibly adopting undesirable means while doing so (Zafirovski, 2004).

Barrett (1993) provided a good example of the use of "traditional behavior" as a baseline model. He gathered data about the frequencies of traditional healthcare practices, which he defined as herbal remedies, reliance on spiritual healers, and childbirth at home. He also gathered data about the frequencies of more modern "biomedical practices" that included vaccination and reliance on physicians and hospitals. His data had high-quality: he surveyed a randomly selected ten percent of the households in a Nicaraguan city and virtually all of the households in four Nicaraguan villages. His goal was not to demonstrate that biomedical practices had better outcomes than traditional practices, but to discover factors that influence people's use of different types of healthcare practices. Although he found statistical correlates of traditional practices, he concluded that these correlations are mainly side-effects of the availability of biomedical care. Biomedical care was not equally available to everyone. Thus, the comparison between traditional practices and biomedical practices went beyond mere statistics and illuminated the economic, ethnic, and social processes that generated the statistics.

No-Change and No-Change-in-Trend

Researchers often propose baseline models about the stability of variables over time. Social trends rarely change very rapidly; even fads rise and wane gradually. Consequently, social trends have inertia in both their magnitudes and their rates of change.

This inertia allows simple baseline models to fit most time-series data very well, and it produces causal processes in which random perturbations have lasting effects. The latter property makes it easier to forecast future values of series, but also makes it more difficult to infer the causal processes that produced past values of series.

A conventional null hypothesis tests whether a variable has any effect at all. In contrast, the simplest baseline model for time series assumes that the variables of interest have some kind of effect on the dependent variable, but predicts that the size of this effect will not change over time.

Darrat and Zhong (2000) described how they applied a no-change baseline model, which they called a 'naive model,' in their study of stock prices on Chinese stock exchanges:

> A NAIVE model maintains that the best forecast for next week's stock price is simply this week's price. . . . Given their prominence in the literature, we employ the NAIVE, ARIMA, and GARCH models to generate ex post weekly forecasts of Chinese stock prices. In addition, we also utilize another, potentially powerful, forecasting technique known as an Artificial Neural Network (ANN) model. (p. 109)

> The two central findings then are that the [baseline] random-walk model does not receive support from the forecasting results, and that the ANN dominates other models in forecasting Chinese stock prices in both Shanghai and Shenzhen markets. The evidence is underscored by the considerable percent improvement in the sum of RMSE from the ANN over the NAIVE model, amounting to 57% improvement for the Shanghai market and 43% for the Shenzhen market. Such significant gains in forecasting Chinese stock prices attest to the departure of the Chinese markets from the random walk

hypothesis [baseline model] and also further support the superiority of the ANN as a powerful forecasting device. (p. 113)

Other researchers who have used no-change models include Gonzales (2000) in a study of real growth, inflation, and international trade in Mexico, Norhayati et al. (2006) in a study of correlates of sudden changes in dividends, and Ozsoz et al. (2010) in a study of interventions into foreign currency markets by the central banks of Croatia, Czech Republic, and Slovakia.

The next simplest baseline model for time series predicts that variables will continue to change at the same rates as they have been changing recently. Chen and Leung (2003) used such a 'no-change-in-trend' baseline model, which they called a 'naive forecast' or 'random walk model,' to evaluate several Bayesian vector models to predict changes in currency exchange rates in Australia, Japan, and Korea.

> . . . the random walk model forecast [baseline model] of next month's change in exchange rate is just this month's change in exchange rate." (p. 892)
>
> Both the [Bayesian vector error correction model] and the [Bayesian vector autoregression model] . . . are able to forecast the 1 month ahead changes in exchange rates better than the naive model. (p. 898)

Other researchers who have used this no-change-in-trend model include Arora and Smyth (1990) who evaluated the accuracy of the economic forecasts made by the International Monetary Fund, Coën and Desfleurs (2004) who examined the accuracy of earnings forecasts by financial analysts in Hong Kong, Korea, Indonesia, Malaysia, the

Philippines, Singapore, Taiwan and Thailand, and Sensarma and Jayadev (2009) who examined the effects of risk-management practices on the stock prices of Indian banks.

Although several of the foregoing researchers used the term "random walk", we have not categorized their baseline models as stochastic processes because their baseline models were actually deterministic rather than probabilistic. The researchers used "random walk" only as a verbal justification for the assumptions made by their baseline models. There have been researchers who used "random walks" as baseline models (Levinthal, 1991), but we have not found examples of this practice in studies of developing countries or emerging markets.

Statistical Independence

Every statistics course introduces the idea of statistically independent variables and illustrates this idea with tables in which the marginal probabilities of variables on the axes determine the probabilities of combinations of events. Tables 2 and 3 illustrate two variations on this idea. Table 2 imagines that two variables can each take three values that have differing probabilities of occurrence; the nine cells at the lower-right show the implied probabilities of combinations of these values if the two variables are statistically independent – that is, if the value of Variable 1 is uncorrelated with the value of Variable 2. Researchers can use a table of this sort as a baseline for assessing the degrees to which observed events look like those implied by statistical independence, and the Chi-square statistic is a familiar metric to evaluate the possible effects of randomness.

Table 3 imagines imports and exports among three countries; the cells on the major diagonal are empty because countries do not export to or import from themselves. In their

analysis of favored-nation biases for trade among 14 nations and regions, Schmidt and Vandenborre (1970) described how they used a table of this sort as a baseline:

> It develops a set of expected data from assumptions of complete indifference among the trading partners and thus allows one to measure the plus or minus differences between these base values and the actual amounts of transactions in each direction for every pair of countries or regions. The method removes gross size effects by taking into account the actual volumes of trade as registered by every country (exports as well as imports) and locates departures from the null-model which could then be examined in a subsequent investigation. The causes for the departures from the null-model could be prices, transportation costs, formally established preference policies, etc. . . . The no-preference assumption is made without regard for reality, insofar as expected data deviate from the actual data will a system of preferences be revealed. (p. 8-9)

Unfortunately, such tables lure researchers into making Chi-square tests, which portray the assessment as a binary one -- independent or not? Baseline models can be more valuable when researchers use them as diagnostic tools to evaluate patterns in deviations from statistical independence. For example, Morgan et al. (2010) in their study of education in a community in northern Nigeria used their baseline model to highlight correlates of educational attainment that included gender, wealth, and social networks. Likewise, Schmidt and Vandenborre (1970) used their baseline model to spot deviations from statistical independence that corresponded to trade agreements, cultural similarities, and price differentials.

Markov Chains

Many baseline models for longitudinal theories start with assumptions or data about period-to-period changes. The simplest such models assume that changes in every period have constant probabilities. Researchers have made assumptions that are more complex in two ways: Markov chains and bootstrapping.

In a Markov chain, each possible current state of a stochastic variable defines a distinct probability distribution for the next change, and as originally formulated, the different probability distributions in a Markov chain are unvarying. With modern statistical software, Markov chains have often been reformulated as conditional correlations.

Some researchers have fitted longitudinal data to Markov chain models to see whether the data fit the assumption of unvarying probabilities. For example, Diamandis (2008) analyzed interactions among financial markets in Argentina, Brazil, Chile, Mexico, and the US by estimating conditional correlations among stock prices. He observed that the conditional probabilities changed several times, and he attributed these shifts to financial crises that occurred around these times. Masson and Ruge-Murcia (2005) loosened the assumption of unvarying probabilities by assuming that the probabilities varied as functions of four explanatory variables. Using data about 2,430 changes in exchange rate policies by 168 countries, they drew inferences about differences between developed and developing countries.

Bootstrapping

Bootstrapping uses the available data as a foundation for simulating possible alternative data. Bootstrapped models of longitudinal data generate new changes in

variables by making random draws from observed distributions of past changes. Several researchers have used bootstrapping in retrospective efforts to evaluate the effectiveness of technical rules for trading stocks. There seems to be agreement that technical rules are not effective with US stocks but they might be more useful with stocks in emerging markets economies. The bootstrapped analyses indicate that the usefulness of technical rules in emerging markets either does not exist (Marshall and Cahan, 2005) or has been decreasing (Cai et al., 2005). However, Muga and Santamaría (2007) used bootstrapping to support the idea that momentum trading is useful in the stock markets of Argentina, Brazil, Chile, and Mexico.

Stochastic Processes with Controls

In field studies, changes in an independent variable of interest frequently coincide with changes in other variables that might also have affected the dependent variable. In such cases, baseline models that account for the effects of other factors help to correctly identify effects of the independent variable of interest. For example, Bekaert, Harvey and Lundblad (2001) investigated the annual economic growth of emerging economies following liberalization of financial equity markets. Market liberalization, however, often coincided with other major political reforms and economic restructuring that might have had direct effects on economic growth. Using Monte-Carlo simulations and the generalized method of moments (GMM), the researchers estimated baseline models that took account of macroeconomic conditions, banking sector development, and equity market development. Then they made hierarchical comparisons of these baseline models with models that also included the explanatory variable "equity market liberalization". The models that controlled for other influences estimated that "liberalization" had had an

average growth effect that was much smaller than the estimated growth effect without the controls.

Topic-Specific Baseline Models

Some baseline models make sense within a specific topic area but have limited meaning elsewhere. Specialists in Finance are very familiar with the capital asset pricing model (CAPM) formula for estimating the cost of capital. Economists know the hypothetical properties of optimal resource allocations. Epidemiologists know the development curve for the spread and decline of epidemics.

Traditional CAPM. Donovan and Nuñez (2012) estimated the costs of capital for investments in renewable energy in Brazil, China, and India. The authors conjectured that a so-called "downside Beta CAPM" would give more reliable estimates because it allows for the possibility that investors have strong aversion to losses. The authors also conjectured that the estimation formulas should be different for firms in different countries. Donovan and Nuñez compared their conjectures with the estimates produced by a "global CAPM" baseline model, which they computed across 60 companies distributed equally among all three countries and which gives no special emphasis on losses. They inferred that their conjectures were correct: Investors do pay more attention to losses and costs of capital differ across countries.

Optimal resource allocation. Pickett et al. (1974) investigated production methods used in the sugar and footwear industries in Ethiopia and Ghana. As a baseline model, the authors (1974: 47) adopted a conventional microeconomic inference: "In profit-

maximizing equilibrium in a two-factor model, the ratio of marginal products of the two factors is equal to the factor-price ratio." However, the authors (1974: 51) found:

> The research . . . (a) records the fact that sugar and footwear production tends to be much more capital intensive than would be expected in Ethiopian and Ghanaian conditions; (b) establishes that there is considerable scope for choice of techniques in the two industries considered; and (c) suggests that in sugar production (in Ghana) and footwear production (in Ethiopia and Ghana) labour-intensive as compared to capital-intensive processes would economize in the use of capital, yield higher net present values for given projects and provide more employment. . . . Consequently, in its strongest and most challenging form, the puzzle which has to be resolved – as a prelude to policy action – is why, given this technological flexibility and demonstrated scope for improvement even in the face of factor-price distortions, appraisal of developing country projects does not produce more rational results than those now obtained.

The researchers attributed these findings, in part, to the use of engineers trained in developed countries to design plants for developing countries and, in part, to economists' willingness to defer to engineers.

Epidemic curve. Hallett et al. (2007) used a complex baseline model in their discussion of HIV trends in Kenya, Thailand, Uganda, and Zimbabwe. The curve asserts that the number of cases first increases slowly, then increases rapidly to a peak, whereupon the number begins to decline but does not drop to zero. The authors argued,

> In the past years, HIV prevalence has declined from very high levels in a few countries, including Zimbabwe and parts of urban Kenya. . . . Care needs to be taken in analysing declines in prevalence because the natural pattern of incidence (ie without any behaviour change) in an epidemic is to increase until the at-risk population is saturated and then to decline to a new lower level matching the supply of new susceptible individuals. . . . Such a decline could be expected approximately a decade after the period of most rapid HIV spread. These natural dynamics need to be excluded as an explanation of decreases in HIV prevalence if we are to detect success in HIV control. . . . Observational studies exploring national reductions in HIV prevalence require mathematical models of the transmission dynamics of HIV to create a null model describing the trajectories of incidence and prevalence in the absence of the intervention. It is then possible to compare the observed data with this null model to estimate the impact of the intervention.

WHERE LIES THE PATH TOWARD PROGRESS?

Comparisons of explanatory models with baseline models can be better than conventional null-hypothesis tests. However, researchers can use baseline models much more effectively than has been the case to date. Any methodological innovation poses challenges as well as opportunities. The studies we have reviewed raise five issues that warrant future thought.

Firstly, baseline models raise the standard for identifying substantively important findings. However, the height of this standard depends on the appropriateness and complexity of the baseline models. Most of the baseline models we found are very simple

ones, and some of them differ only slightly from conventional null-hypothesis tests. As a result, they leave the standard rather low. For example, Mastromarco (2008) compared a baseline model that included just two independent variables

$$\log Q = \beta_1 \log L + \beta_2 \log C$$

with an explanatory model that included three more quadratic terms

$$\log Q = \beta_1 \log L + \beta_2 \log C + \beta_3 (\log L)^2 + \beta_4 (\log C)^2 + \beta_5 \log L * \log C + \text{Error}$$

Her comparison tested the point null hypotheses that β_3, β_4, and β_5 equal zero. She inferred that her explanatory model is better than the baseline model because she could reject the three null hypotheses. However, those three null hypotheses were *point* hypotheses that insisted the data must be consistent with β_3, β_4, and β_5 being exactly zero – not very close to zero, but exactly on the infinitesimal point at zero. As a result, Mastromarco could have made certain of rejecting the three null hypotheses by gathering a large-enough sample.

Secondly, when baseline models become ritualistic, researchers devote less care to puzzling over the best ways to adjust them to the immediate empirical contexts and to find defects in their explanatory theories, and consequently, they derive less value from their analyses. For example, many studies of time series used one of two very conventional baseline models – either no-change or no-change-in-trend. Although these baseline models do set higher standards than has often been the case with conventional null-hypothesis tests, they are still very elementary starting points: They incorporate no assumptions about the causal processes that generate the time series or about the potential uses for explanatory models. The most interesting and useful examples of baseline models

incorporate assumptions that reflect the idiosyncrasies of specific research topics and contexts. For instance, studies of time series could consider baseline models that take account of specific exogenous influences or specific types of feedback.

Thirdly, comparison of an explanatory model with a baseline model is quite different from using the baseline model to "test" the explanatory theory. The word "test" implies a dichotomous judgment about the adequacy of an explanatory model, whereas the word "compare" implies an analysis of differences between models. How does the explanation provided by the baseline model differ from the explanation provided by the explanatory model? For example, how do the residuals around a baseline regression function differ from the residuals around an explanatory regression function? By analyzing the different patterns in the residuals, a researcher can identify observations that each function explains well or poorly, and possibly, the researcher can develop better understanding of what makes the explanatory model better. Wimsatt (1987, 2002) pointed out that, when researchers are trying to analyze unexplained residuals rather than trying to fit data more closely, baseline models built on false assumptions can be more useful than baseline models built on realistic assumptions.

Fourthly, comparisons of explanatory models with baseline models have usually differed from competitions between alternative explanatory models. Since researchers design or select their baseline models, they have full control over the properties of these models, and they can decide how closely their baseline models resemble their explanatory theories.

Leavitt, Mitchell, and Peterson (2010) and Rodgers (2010) have advocated that researchers should stage competitions between alternative theories. For competitions to be meaningful, of course, the competitors have to be more than artificial constructions controlled by the researchers. New models can compete with preexisting models, or the models proposed by one researcher can compete with the model proposed by a different researcher. Researchers have used baseline models in such competitions. For example, Connor and Simberloff (1983) regarded their model of randomized species pairs as a competitor to existing models. However, most baseline models have been diagnostic devices rather than genuine alternative models.

Fifthly, use of baseline models implies that researchers are trying to progress from a less understanding toward greater understanding. Such researchers may not be sure they understand the phenomena they are investigating, or they may have little confidence in some tentative conjectures, or they may be very aware that their theories are incomplete. They want to improve their theories rather than to demonstrate the excellence of their theories. They are aware that they could have designed or managed their studies better, or that their data are unreliable, or that they may not have noticed significant influences or patterns. Thus, use of one baseline model implies that the ensuing step is likely to be development of subsequent and different baseline models.

One way to progress in understanding is to use series of analyses to develop from simpler baseline models toward more complex baseline models. For example, Mastromarco could have examined deviations of her data from the baseline model

$$\log Q = \beta_1 \log L + \beta_2 \log C$$

to develop ideas about a better explanatory model. These deviations might have led her to consider quadratic effects of the sort she actually considered. Or, they might have led her to add another variable such as technological development (Solow, 1957):

$$\log Q = \beta_1 \log L + \beta_2 \log C + \beta_3 \log T$$

An alternative way to progress in understanding is to move from less abstracted baseline models toward more abstracted baseline models. Gotelli and Graves (1996) reported that bioecologists have often debated about the inclusion of data properties in baseline models. For example, in the study mentioned earlier in this chapter, Connor and Simberloff (1983) assumed that each of the Galapagos Islands had the actual number of bird species observed on it and that each species inhabited as many islands as actually observed. These assumptions incorporate information about some physical and vegetation differences among the islands. A baseline model that incorporates no data properties is easy to reject, but rejecting it yields no information because the model is so unrealistic. On the other hand, a baseline model that incorporates many data properties is very difficult to reject because it differs little from the data, but rejecting such a complex model also yields no information because it confounds too many assumptions. Although the two extremes are not useful, bioecologists have argued about the desirability of options between the extremes.

There are other ways to progress from lesser understanding toward greater, but optimal paths cannot exist. In order to find the optimal path to a destination, one has to know one's destination. The essence of research is that exact destinations are unknown, and in the social and behavioral sciences, researchers may discover more interesting and

exciting destinations as they progress. Researchers need to reflect again and again about what they have learned so far and what would be a productive next step. What would be some important questions to ask next? What would be an even tougher challenge to attempt?

Archibald's world

Possibly the most surprising outcome of our search for baseline models was a renewed awareness of and respect for the difference between testing a theory and analyzing data. Many users of baseline models have not seen themselves as "testing" their explanatory theories.

More generally, data analyses provide very weak foundations for refuting inadequate theories. Archibald (1967) is one of several philosophers who have criticized the idea that data provide grounds for refuting theories, even very bad ones. He argued that knowledge is never complete and never entirely correct. All hypotheses have defects and limitations, many of which are unknown to those who formulate the hypotheses. In the social and behavioral sciences, knowledge is probabilistic, if only because attempts to measure variables always entail errors. Thus, said Archibald, researchers should not be trying to prove their explanatory theories or trying to disprove alternative hypotheses because every hypothesis and every theory must inevitably fail rigorous tests. He (1967: 295) said:

> It remains to point out that comparison is multi-dimensional. In the first place, we compare rival theories by reference to such criteria as scope, generality, elegance, etc.: we ask, e.g. if one is 'about more things' than another, or carries excess baggage

> in the form of unnecessary elements, or connects more satisfactorily with other parts of our theoretical structures. We should not be surprised if a theory which is superior on some counts is inferior on others! In the second place, when we compare theories with observation, we commonly find more than one criterion. Thus we may ask which better accounts for, e. g. total variance, or for turning points, or for amplitude of fluctuations. Once again, we should not be surprised if the theory which does better by some criteria does worse by others. The case of vector-dominance is the rare and lucky one in which one theory at last wins all down the line, so that we may reject its rival without waiting for the refutation that never occurs.

A perspective similar to Archibald's has often motivated the use of baseline models. Researchers have not striven to refute baseline models; they have used baseline models to learn more about the phenomena they are studying or to learn how to create better theories.

The foolishness of refutation does not arise solely from the character of social and behavioral data. Researchers need to beware of their own personal limitations and deficiencies. For instance, Rothman pointed out that theories in physics incorporate logical impossibilities, and theories that purport to mesh with each other, do not do so. He (2011: 186) said: “Vanishing few problems in physics have exact solutions and a physicist’s career is one of finding approximations and hopefully not being too embarrassed by them.”

Rothman's observations point to deficiencies in human logic (Faust, 1984; Meehl, 1991). Theories of all sorts incorporate assumptions that breakdown when data fall

outside limits or even assumptions that can never be true. Gray and Cooper (2010) have urged researchers to devote more effort to identifying such logical boundaries. However, logical limitations are only some among many challenges. Researchers also have to deal with their attributional biases, their propensities to search for confirmatory data, and blind spots generated by their hypotheses and theories (Beach, 1966; Calhoun and Starbuck, 2003; Erev and Barron, 2005; Hansen, 1980; Kahneman and Tversky, 1973; Lichtenstein et al., 1982; Phillips et al., 1966; Slovic, 1991). Because researchers write and revise their papers after they have completed data analyses, hypotheses and theories incorporate retrospective sensemaking, which amplifies the influence of idiosyncratic random events. And, because realistic researchers can predict that later research will reveal issues that they have not yet seen, they ought to be cautious about making claims about the generality, completeness, or accuracy of their theories.

The biases inherent in personal commitments and the social contexts of research imply that the content and characteristics of data are at least as important as the content and characteristics of theories. Even superb logic and the most careful methods cannot overcome inadequate, deceptive, or misunderstood data. In 1543, Copernicus created a cosmological theory that was far better than the theories that had existed, but his proposal did not win rapid acceptance. A major reason was the ambiguity of extant data. In 1651, Riccioli spelled out 77 reasons why Copernicus's theory could not possibly be correct, and some of Riccioli's objections remained unanswerable for two centuries. Scientists needed time to develop observational techniques that could produce appropriate data. Similarly, it may be that today's behavioral and social theories will appear either foolish or profound as researchers develop better measures of economic and social phenomena or better

techniques for examining cognitive processes inside human bodies. What is crucial is not the volume of data but its quality and relevance. For instance, Davis (2010) has pointed out that theories about organizations have remained rather stagnant for over three decades even though researchers have been able to subject masses of data to statistical analyses.

What does the future hold?

Comparisons with baseline models are more useful analytic tools than conventional null-hypothesis tests. Baseline models can provide standards that can range from easy-to-meet, if researchers are timid, to very-difficult-to-surpass, if researchers dare to make them so. Series of baseline models can support increasingly realistic assumptions about the context under study, or increasingly abstract assumptions. Careful and systematic comparisons between models support iterative theory development (Archibald, 1967, Rodgers, 2010). Table 4 offers a few suggestions for iterative development that derive from the activities of emerging markets studies.

The flexibility of baseline models highlights decisions made by researchers. Instead of applying standards set by the traditions and rituals associated with conventional null hypotheses, researchers decide what standards to use, and they can select or design baseline models that are appropriate for their specific research contexts. Researchers ought to explain and defend these decisions in terms that relate to their specific studies. One resulting challenge for researchers is to become more comfortable with developing and explicating the baseline models they use.

Although some emerging-markets studies have been using baseline models for over half a century, these studies still constitute only small fractions of the published work. Baseline models deserve much wider discussion, experimentation, and understanding.

REFERENCES

Ames, E., & Reiter, S. (1961). Distributions of correlation coefficients in economic time series. *Journal of the American Statistical Association*, 56, 637-656.

Archibald, G.C. (1967). Refutation or comparison? *The British Journal for the Philosophy of Science,* 17(4), 279-296.

Arora, H.K., & Smyth, D.J. (1990). Forecasting the developing world: An accuracy analysis of the IMF's forecasts. *International Journal of Forecasting*, 6, 393-400.

Barrett, B. (1993). Health care behavior on Nicaragua's Atlantic coast. *Social Science & Medicine*, 37(3), 355-368.

Beach, L.R. (1966). Accuracy and consistency in the revision of subjective probabilities. *IEEE Transactions on Human Factors in Electronics*, HFE-7, 29-37.

Bekaert, G., Harvey, C. R., & Lundblad, C. (2001). Emerging equity markets and economic development. *Journal of Development Economics*, 66, 465-504.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association,* 33, 526-536.

Cai, B.M., Cai, C.X., & Keasey, K. (2005). Market efficiency and returns to simple technical trading rules: Further evidence from U.S., U.K., Asian and Chinese stock markets. *Asia-Pacific Financial Markets*, 12, 45–60.

Calhoun, M.A., & Starbuck, W.H. (2003). Barriers to creating knowledge. In Easterby-Smith, M., & Lyles, M.A. (Eds), *Handbook of organizational learning and knowledge management*. Malden, MA: Blackwell, 473-492.

Chen, A.S., & Leung, M.T. (2003). A Bayesian vector error correction model for forecasting exchange rates. *Computers & Operations Research*, 30, 887–900.

Coën, A., & Desfleurs, A. (2004). The evolution of financial analysts' forecasts on Asian emerging markets. *Journal of Multinational Financial Management*, 14, 335–352.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.

Connor, E. F., & Simberloff, D. (1983). Interspecific competition and species co-occurrence patterns on islands: Null models and the evaluation of evidence. *Oikos*, 41, 455-465.

Connor, E.F., & Simberloff, D. (1986). Competition, scientific method, and null models in ecology. *American Scientist*, 74(2), 155-162.

Copernicus, N. (1543). *De revolutionibus orbium caelestium*. Nuremberg: Johannes Petreius.

Darrat, A.F., & Zhong, M. (2000). On testing the random-walk hypothesis: A model-comparison approach. *The Financial Review*, 35, 105-124.

Davis, G.F. (2010). Do theories of organizations progress? *Organizational Research Methods*, 13(4), 690-709.

Diamandis, P.F. (2008). Financial liberalization and changes in the dynamic behaviour of emerging market volatility: Evidence from four Latin American equity markets. *Research in International Business and Finance*, 22, 362–377.

Donovan, C., & Nuñez, L. (2012). Figuring what's fair: The cost of equity capital for renewable energy in emerging markets. *Energy Policy*, 40, 49-58.

Edwards, J. R., & Berry, J. W. (2010). The presence of something or the absence of nothing: Increasing theoretical precision in management research. *Organization Research Methods*, 13(4), 668-689.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, 112(4), 912-931.

Faust, D. (1984). *The limits of scientific reasoning*. Minneapolis: University of Minnesota Press.

Gigerenzer, G., Krauss, S. & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In Kaplan, D. (Ed), *The Sage handbook of quantitative methodology for the social sciences*. Sage, Thousand Oaks, CA, 391-408.

Gotelli, N.J., & Graves, G.R. (1996). *Null models in ecology*. Washington: Smithsonian Institution Press.

Gray, P.H., & Cooper, W.H. (2010). Pursuing failure. *Organizational Research Methods*, 13, 620-643.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82, 1-20.

Hallett, T.B., White, P.J., & Garnett, G.P. (2007). Appropriate evaluation of HIV prevention interventions: from experiment to full-scale implementation. *Sexually Transmitted Infections*, 83(Supplement I), i55–i60.

Hansen, R.D. (1980). Commonsense attribution. *Journal of Personality and Social Psychology*, 39(6), 996-1009.

Johnson, P.R. (1968). Discussion: A test of the hypothesis of economic rationality in a less-developed economy. *American Journal of Agricultural Economics*, 50(2), 398-399.

Johnson, P.R. (1969). On testing competing hypotheses: Economic rationality versus traditional behavior: Reply. *American Journal of Agricultural Economics*, 51(1), 208-209.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237-251.

Leavitt, K., Mitchell, T.R., & Peterson, J. (2010). Theory pruning: Strategies to reduce our dense theoretical landscape. *Organizational Research Methods*, 13, 644-667.

Levinthal, D.A. (1991). Random walks and organizational mortality. *Administrative Science Quarterly*, 36: 397-420.

Lianos, T.P. (1969). A comment on a traditional behavior model. *American Journal of Agricultural Economics*, 51(4), 937.

Lichtenstein, S., Fischhoff, B. & Phillips, L. (1982). Calibration probabilities: The state of the art to 1980. In Kahneman, D., Slovic, P., & Tversky, A. (Eds), *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press, 306-333.

Marshall, B.R., & Cahan, R.H. (2005). Is technical analysis profitable on a stock market which has characteristics that suggest it may be inefficient? *Research in International Business and Finance*, 19, 384–398.

Maslow, A.H. (1966). *The psychology of science*. Harper & Row, New York.

Masson, P., & Ruge-Murcia, F.J. (2005). Explaining the transition between exchange rate regimes. *Scandinavian Journal of Economics*, 107(2), 261–278.

Mastromarco, C. (2008). Foreign capital and efficiency in developing countries. *Bulletin of Economic Research*, 60(4), 0307-3378.

Meehl, P.E. (1991). *Selected philosophical and methodological papers*. Minneapolis: University of Minnesota Press.

Morgan, S.L., Mohammed, I.Z., & Abdullahi, S. (2010). Patron–client relationships and low education among youth in Kano, Nigeria. *African Studies Review*, 53(1), 79–103.

Muga, L., & Santamaría R. (2007). The momentum effect in Latin American emerging markets. *Emerging Markets Finance and Trade*, 43(4), 24–45.

Norhayati, M. Hamid, M.A.A., Nassir, A.M., & Mohamed, S. (2006). Information content of dividend changes: Cash flow signalling, dividend clientele and free cash flow hypotheses. *Malaysian Accounting Review*, 5(1), 65-84.

Ozsoz, E. Rengifo, E.W., & Salvatore, D. (2010). Deposit dollarization as an investment signal in transition economies: The cases of Croatia, the Czech Republic, and Slovakia. *Emerging Markets Finance and Trade*, 46(4), 5–22.

Peach, J.T., & Webb, J.L. (1983). Randomly specified macroeconomic models: Some implications for model selection. *Journal of Economic Issues*, 17, 697-720.

Phillips, L.D., Hays, W.L. & Edwards, W. (1966). Conservatism in complex probabilistic inference. *IEEE Transactions on Human Factors in Electronics*, HFE-7, 7-18.

Pickett, J., Forsyth, D.J.C., & McBain, N.S. (1974). The economic choice of technology, economic efficiency and employment in developing countries. *World Development*, 2(3), 47-54.

Riccioli, G.B. (1651). *Almagestum novum*. Bologna: Victorij Benatij.

Rodgers, J.L. 2010. The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *American Psychologist*, 65, 1-12.

Rothman, T. (2011). The man behind the curtain. *American Scientist*, 99, 186-189.

Schmidt, S. G., & Vandenborre, R J. (1970). Preference patterns in the world coarse grain trade. *Canadian Journal of Agricultural Economics*, 18(1), 6–19.

Schmidt, F.L., & Hunter, J.E. (1997). Eight common but false objections to the discontinuation of significance testing in analysis of research data. In Harlow, L., Mulaik, S., & Steiger, J. (Eds), *What if there were no significance tests?* Erlbaum, Mahwah, NJ, 37–63.

Schwab, A., Abrahamson, E., Starbuck, W.H. & Fidler, F. (2011). Researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science*, 22(4), 1105-1120.

Sensarma, R., & Jayadev, M. (2009). Are bank stocks sensitive to risk management? *The Journal of Risk Finance*, 10(1), 7-22.

Seth, A., Carlson, K.D. Hatfield, D.E. & Lan, H.W. (2009). So what? Beyond statistical significance to substantive significance in strategy research. In Bergh, D. & Ketchen, D. (Eds), *Research Methodology in Strategy and Management*, Volume 5. Elsevier JAI, New York, 3-28.

Slovic, P. (1991). Perceptions of risk: Paradox and challenge. In *Hazmat '91: Proceedings*. Evanston, IL: Northwestern University, Transportation Center, 4-3—4-22.

Solow, R.M. (1957). Technical change and the aggregate production function. *Review of Economics and Statistics*, 39(3), 312-320.

Starbuck, W.H. (1994). On behalf of naïveté. In Baum, J.A.C. & Singh, J.V. (Eds), *Evolutionary dynamics of organizations*. Oxford University Press, New York, 205-220.

Thompson, B. (1999). Journal editorial policies regarding statistical significance tests: Heat is to fire as p is to importance. *Educational Psychology Review*, 11, 157-169.

Webster, E.J., & Starbuck, W.H. (1988). Theory building in industrial and organizational psychology. In Cooper, C.L. & Robertson, I. (Eds), *International Review of Industrial and Organizational Psychology, 1988*. Wiley, London, 93-138.

Wimsatt, W.C. (1987). False models as means to truer theories. In Nitecki, M.H. & Hoffman, A. (Eds), *Neutral Models in Biology* (pp. 23-55). Oxford, UK: Oxford University Press.

Wimsatt, W.C. (2002). Using false models to elaborate constraints on processes: Blending inheritance in organic and cultural evolution. *Philosophy of Science*, 69(S3, September), S12-S24.

Wise, J., & Yotopoulos, P.A. (1968). A test of the hypothesis of economic rationality in a less-developed economy: An abstract. *American Journal of Agricultural Economics*, 50(2), 395-397.

World Bank (2011, November 5). World development indicators 2011. Retrieved from http://data.worldbank.org/data-catalog/world-development-indicators

Yotopoulos, P.A., & Wise, J. (1969a). On testing competing hypotheses: Economic rationality versus traditional behavior: A further development. *American Journal of Agricultural Economics*, 51(1), 203-208.

Yotopoulos, P.A., & Wise, J. (1969b). On testing competing hypotheses: Economic rationality versus traditional behavior: Rejoinder. *American Journal of Agricultural Economics*, 51(1), 209-210.

Zafirovski, M. (2004). Sociologics of the economy: The social logic, composition and structuration of economic behavior. *Social Science Information*, 43(4), 691–743.

Figure 1. Frequency distributions of autocorrelation coefficients in emerging-markets time series, 1989-2009

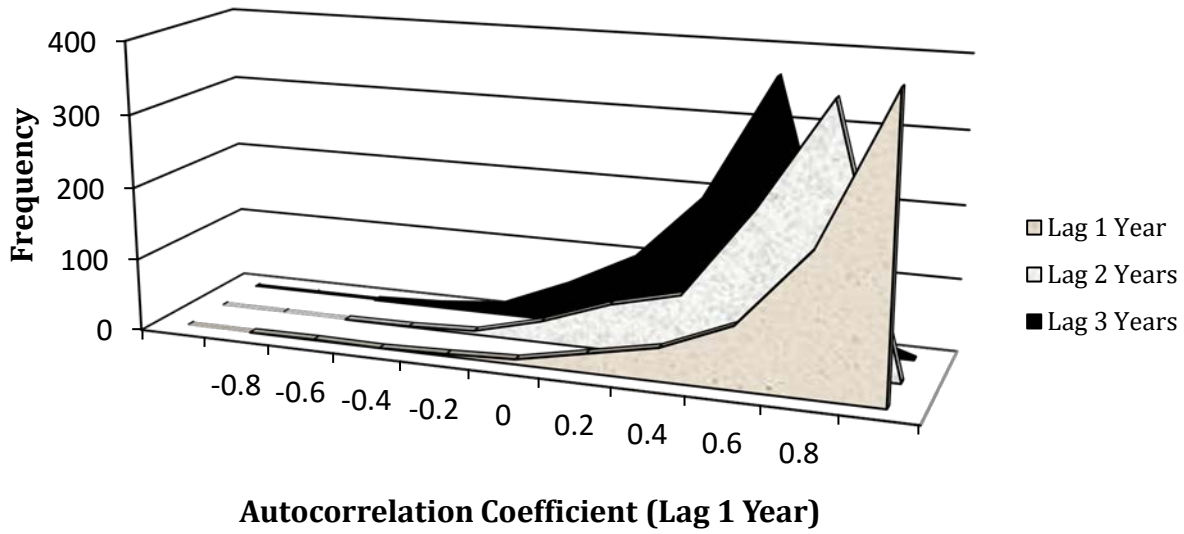

Figure 2. Frequency distributions of autocorrelation coefficients in emerging-markets time series, 1989-2009

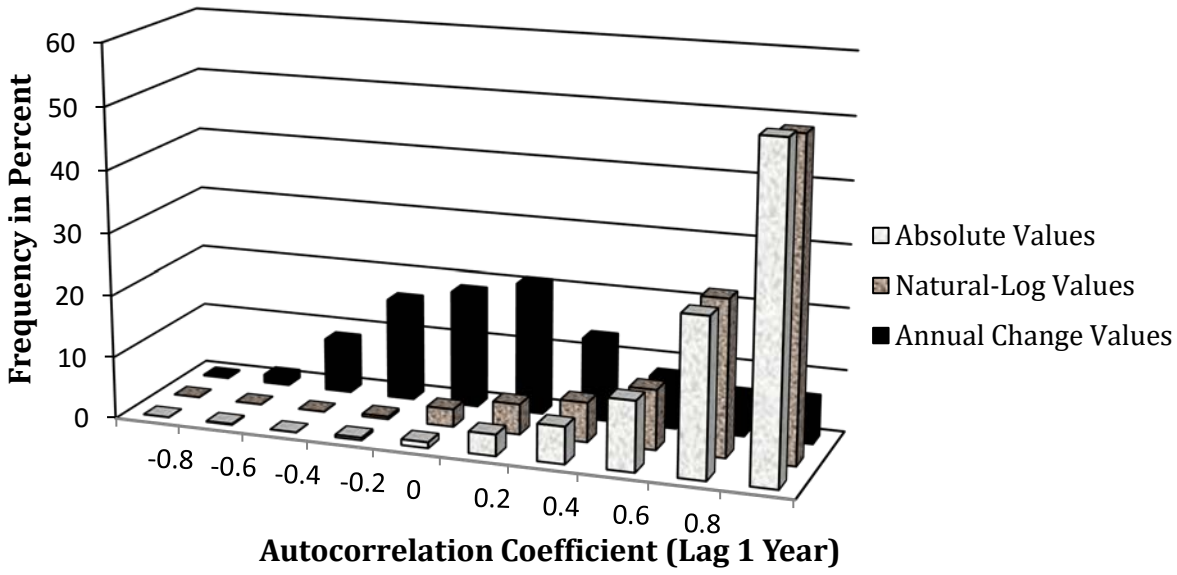

Figure 3. Histogram of simple correlation coefficients reported in published emerging-markets studies, 2007-2011

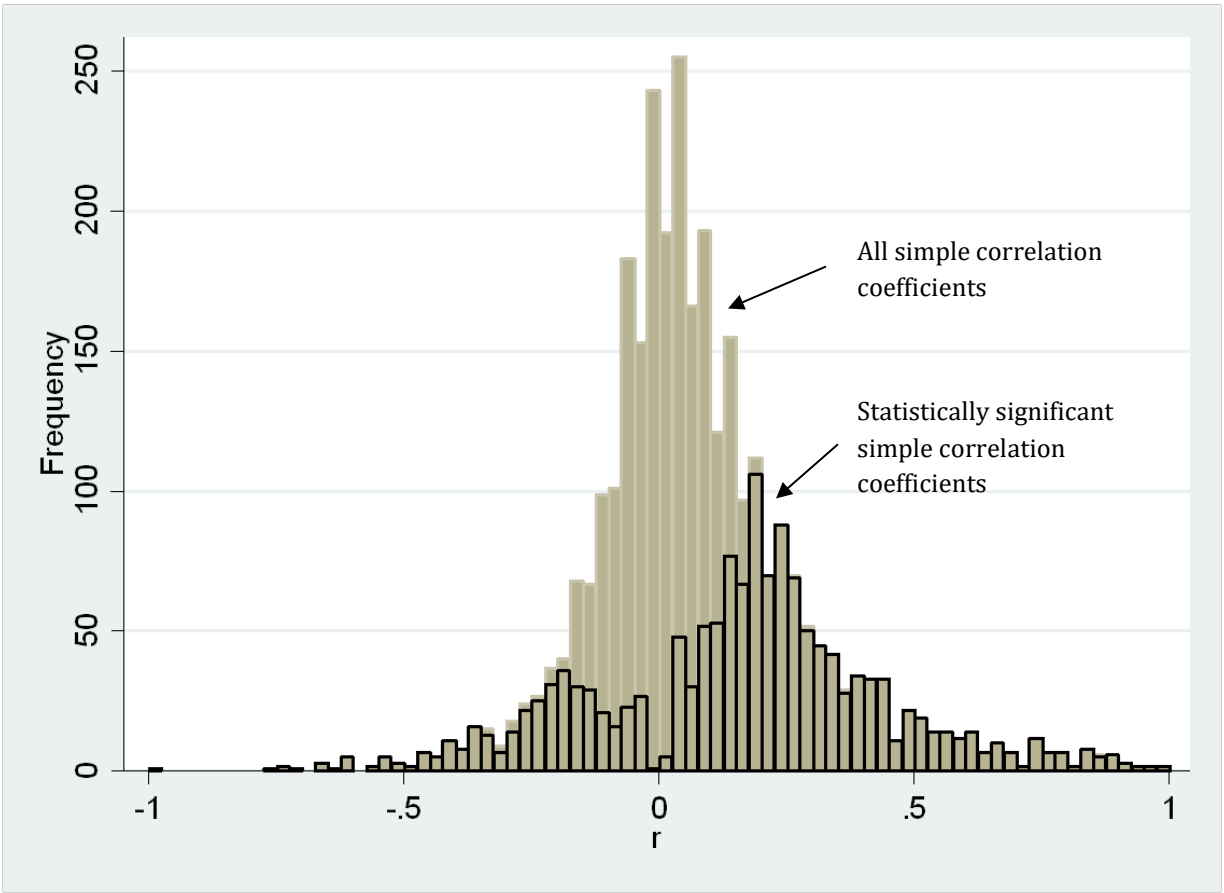

Figure 4. Percentage of academic papers that mention emerging markets or developing countries that also mention baseline models (3-year moving averages)

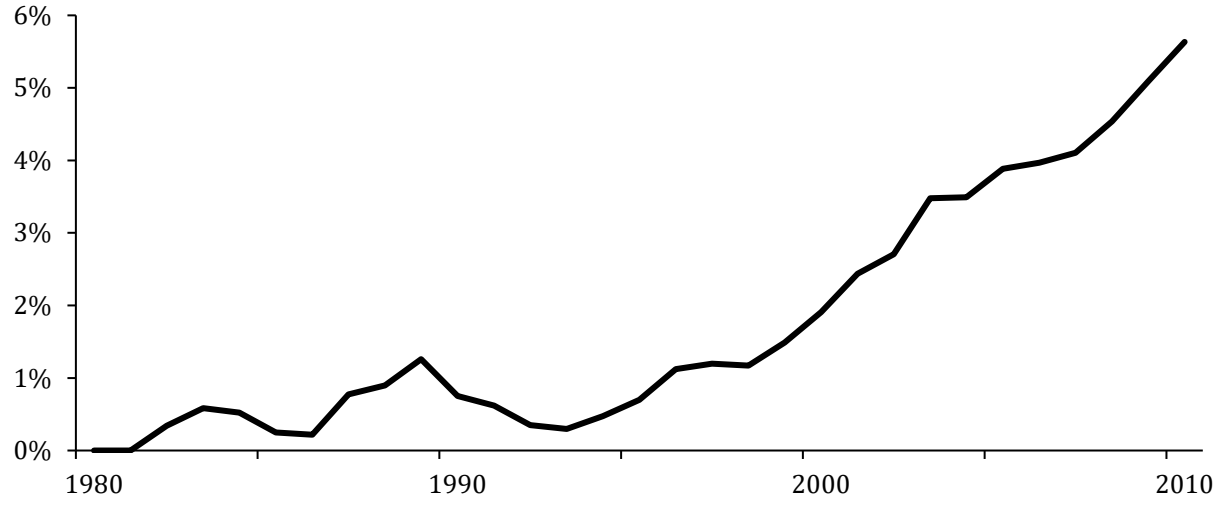

Table 1. Types of theories and examples of baseline models		
	Status at one time	**Changes over time**
Stability and change	Log-linear scaling Traditional behavior	Log-linear scaling Traditional behavior No-change No-change-in-trend
Random processes	Statistical independence	Markov chains Bootstrapped distributions of period-to-period changes Stochastic processes with control variables
Topic-specific baseline models	Traditional CAPM Optimal resource allocation	Epidemic curves

Table 2 **Independent probabilities**		**Variable 1**		
		Event 1A Probability = 0.2	Event 1B Probability = 0.3	Event 1C Probability = 0.5
Variable 2	Event 2A Probability = 0.2	0.04	0.06	0.10
	Event 2B Probability = 0.2	0.04	0.06	0.10
	Event 2C Probability = 0.6	0.12	0.18	0.30

Table 3 Independent flows		Exports		
		From Country A Percent = 20	From Country B Percent = 30	From Country C Percent = 50
Imports	To Country A Percent = 20		7.5	12.5
	To Country B Percent = 20	5.6		14.4
	To Country C Percent = 60	24.0	36.0	

Table 4. Suggestions for developing and revising baseline models		
Initial formulation	What kind of very simple process could generate data similar to the observed distribution of the dependent variable?	Examples include the types of models in Table 1.
Comparisons with explanatory model	Are there large differences between predictions of baseline model and explanatory model?	
	Are there large differences in the fits to data of baseline model and explanatory model?	
Possible revisions	Would a baseline model that incorporates additional properties of the data fit the data very much better?	
	Would a baseline model that incorporates fewer properties of the data fit the data equally well?	
	Would a more complex baseline model fit the data very much better?	
	Would a simpler baseline model fit the data equally well?	
	What changes in the baseline model would reduce the differences in fit between the baseline model and the explanatory model?	