JOIM
www.joim.com

JOURNAL OF INVESTMENT MANAGEMENT, Vol. 14, No. 3, (2016), pp. 62–73


# SURVEYS AND CROSSOVER

This section provides surveys of the literature in investment management or short papers exemplifying advances in finance that arise from the confluence with other fields. This section acknowledges current trends in technology, and the cross-disciplinary nature of the investment management business, while directing the reader to interesting and important recent work.

## CORRELATION OR CAUSATION?: THE SORRY STATE OF INFERENCE IN EMPIRICAL MODELING

*Xiaojing Dong and John Heineke*

*For decades, statistical methods, many based upon the "general linear model," have been used to do estimation and test hypotheses in the social and natural sciences, in medicine, and in the private sector. These tools have become increasingly sophisticated and are often paired with powerful open source data analytic software. We now regularly see mathematical/statistical output combined with data visualizations that are truly mind-boggling and, once in a while, thought provoking. But an increasing number of papers and studies appear to have little statistical validity, in which the line between causality and correlation is often non-existent. This is a danger sign not only in science and medicine but also to companies who unwittingly rely on such results for forecasting and business strategy. Could it be true that researchers and analysts who learn ever more powerful analytical methods lack even a basic understanding of the limitations of these methods? The purpose of this short, non-technical paper, which relies heavily upon examples, is to shed some light on the underlying statistical issues. The ideas here are certainly not original with us and have been raised for a number of years across multiple disciplines.*

For decades, statistical methods, many based upon the "general linear model," have been used to do estimation and test hypotheses in the social and natural sciences, in medicine, and in the private sector.[1] These tools have become increasingly sophisticated and are often paired with powerful open source data analytic software. We now regularly see mathematical/statistical output combined with data visualizations that are truly mind-boggling and, once in a while, thought provoking. But an increasing number of papers and studies appear to have little statistical validity, in

which the line between causality[2] and correlation is often non-existent.[3] This is a danger sign not only in science and medicine but also to companies who unwittingly rely on such results for forecasting and business strategy. Could it be true that researchers and analysts who learn ever more powerful analytical methods lack even a basic understanding of the limitations of these methods? The purpose of this short, non-technical paper, which relies heavily upon examples, is to shed some light on the underlying statistical issues. The ideas here are certainly not original with us and have been raised for a number of years across multiple disciplines.

We proceed by enumerating empirical modeling problems in a number of papers appearing in multiple disciplines and professions, including finance as discussed in our first example. Our purpose is to demonstrate the scope of the problem. In a sentence, the problem we discuss might be best described by the question: "Causation or Correlation?" We separate problems into those that probably are best categorized as statistical (see the section, "Overfitting, p-Hacking. . .") and those that are best thought of as econometric ("Endogeneity and Related Issues"). We conclude by offering a few recommendations for the practitioner which may help to alleviate some of the problems which arise when researchers use data sets that are not generated from well-thought-out, well-controlled experiments.

*Overfitting*, *p-Hacking and Related Issues*: A persistent problem for those who rely on the general linear model and its variants, in science, in academia and in the workplace, is the countless attempts to extract more information from a data set than is inherent in the data itself—often called "overfitting." This can take many forms, from using the available data to identify the variables which will be included in the model,[4] to using automated procedures which run a model in all possible variants to select the "best" model,[5] to running regressions repeatedly until "something works". The point here is that there are only a given number of degrees of freedom in a set of data, and as the number of degrees of freedom falls (as the model is repeatedly re-estimated), one ultimately ends up fitting the model to the noise in the sample, i.e., to the idiosyncrasies that are specific to the sample.[6] The result is higher $R^2$s, more significant $t$-statistics but results which will not replicate outside the particular sample,[7] because they exist only in the sample and not in the population that the sample represents. Several papers point out just how poorly automated procedures perform (see, for example, Derksen and Keselman, 1992; Mundry and Nunn, 2009). One simulation study shows that stepwise regression, a widely used automated procedure, produced final models in which 30–70% of the selected explanatory variables were not at all related to the response variable in the population (Derksen and Keselman, 1992). In brief, the problem is that more degrees of freedom are used in estimation than are available in the given sample. In cases of this type, not only is causality missing but even the correlations observed do not exist in the underlying population being sampled. In science and in most workplace modeling tasks, one needs results which are replicable outside of the sample.[8]

**Example 1.** Empirical finance is beset with issues of this type according to the president-elect of the American Finance Association who recently stated that "most claimed research findings in financial economics are likely false" (Harvey *et al.*, 2015). The paper's startling revelation was that at least 313 "factors" driving investment returns have been discovered through multiple regressions, which have been published in hundreds of articles and working papers. That phenomenon had already been characterized in 2011 by finance professor John H. Cochrane who termed it a "zoo" of factors. The problem here is

closely related to overfitting and in the statistical literature is often termed "multiple testing." Suppose one picks the often used confidence level of 95%, so the probability of type-I error is 5% for the first time one tests. But this error gradually increases as one tweaks the model and tests it a second time and then again and again... Eventually, the probability of type-I error converges to one. This fact was known in the first half of the twentieth century, but appears to have been seldom (never?) acted upon. Of course it is understandable that it was ignored, since few journals or superiors in companies would be interested in results where the actual level of significance in models tested multiple times was known and very high. This is also known as p-hacking or more colloquially as "running the data until it gives up". See López de Prado (2015) for more details in a finance setting and some ideas for a remedy.

**Example 2.** Harvey *et al.*'s paper sounded the same alarm for empirical finance research as had been sounded in 2005 for medical research in a widely referenced article by John P. A. Ioannidis, "Why Most Published Research Findings Are False" (Ioannidis, 2005). In the words of the author, "There is increasing concern that in modern (medical) research, false findings may be the majority or even the vast majority of published research claims. However, this should not be surprising. It can be proven that most claimed research findings are false". Ioannidis provides more than 30 references in many of the best medical journals, to buttress his claim. The fields cited include epidemiology, molecular research, genetics, and cancer research. As in empirical finance, the problem came to light because of the non-replicability of published research findings. Ioannidis summarizes the issues into three categories: (1) low power tests, (2) low prior likelihood of the truth being tested, and (3) bias. Ioannidis makes many suggestions for improving this state of affairs and concludes that "statistical significance testing in the report of a single study gives only a partial picture, without knowing how much testing has been done outside the report and in the relevant field at large. Despite a large statistical literature for multiple testing corrections, usually it is impossible to decipher how much data dredging by the reporting authors or other research teams has preceded a reported research finding."

As should be evident, any "results" generated in the manner described in Examples 1 and 2 will severely compromise forecasts and any plans or policies based upon such results. The costs are high not only in academic studies leading to reduced understanding of complex processes and impeding progress in the field, but also in the private sector where forecasts fail as do policies and strategy based upon them. Arguably the cost is highest in medical science where medical advice is dispensed based upon invalid studies.[9] In these cases, there is not only no causality but the correlations are "faux" correlations which do not exist in the underlying populations being sampled.[10] Researchers and practitioners need to be cognizant of these issues whether they are doing academic, medical, financial, marketing, or human resource based statistical studies.

One observation which arises from our discussion above, is to recognize that many quantitative researchers rely far too heavily on the power of control variables. This is true whenever regressions are run repeatedly with all conceivable sets of controls until published results are not replicable and any element of causality that might have been present early on, has long since vanished.

*Endogeneity and Related Issues*: Next we discuss a different type of problem. We now assume that there are only well-thought-out, well-designed research plans, where all variables are measured correctly and all estimations and analyses are executed by highly ethical practitioners who do not

overfit or p-hack. The problems discussed above arise from the fact that often there are no randomized experiments and the data available are observational data that are not the outcome of a natural experiment. Much of the data in finance, marketing, the social sciences, many areas of business and medicine fall into this category. There are many instances where no matter how good are the data or how large is the sample, there is no possibility of obtaining an accurate estimate of the parameters of a model, even if the sample size increases without bound. Some of the social sciences, especially those ensconced in business schools and economics departments, have been aware of this problem for some time. The most often used word, in the past decade or two, to describe the problem, which also results in the absence of causality, is "endogeneity", meaning that the values of one or more explanatory variables in a model are not exogenous to the model,[11] but rather are determined by the model (i.e., are endogenous to the model) in a specific sense to be described below.[12] In these cases, parameter estimates are biased and inconsistent.[13] Technically, endogeneity arises because the error term in the statistical model is correlated with one or more of the factors ($X$ variables) driving outcomes (the $Y$ variable), where the "error term" in a model includes all influences on $Y$ that are not included in the right-hand-side $X$ variables. Roughly then, there is an influence on the $Y$ variable that also influences one or more $X$ variables and that has not been included in model. Typical usage would say that endogeneity arises from either: omitted variables, measurement error, simultaneity, or sample selection, although these factors are often not truly distinct. We will speak more about this below. Nonetheless these distinctions are probably useful. Older econometric textbooks used "the identification problem" to describe the same issues[14] which occurs when the $X$ variables of interest are not independently determined and therefore no causal relationship can be established using standard regression analysis. To motivate the discussion of one source of endogeneity we provide a recent example.

**Example 3 (Missing Variable(s)).** Citations are a key currency in assessing the relevance of scientific reports. The number of times other researchers cite a scientist's work is often an important metric in hiring and in workplace evaluations in academia, government and the private sector. Citations also play a role in determining a scientific journal's place in the scholarly pecking order, with journals that publish more heavily cited papers earning a higher "impact factor." In the abstract of a paper bearing the title "The advantage of short paper titles,[15]" the authors, Letchford *et al.* (2015) report that: "Vast numbers of scientific articles are published each year, some of which attract considerable attention, and some of which go almost unnoticed. Here, we investigate whether any of this variance can be explained by a simple metric of *one aspect* (our emphasis) of the paper's presentation: the length of its title. Our analysis provides evidence that journals which publish papers with shorter titles receive more citations per paper.[16]" Researchers whose income, raises and promotion depend upon citation counts typically believe and, romantically perhaps, hope that citation counts depend critically upon a far different set of variables; including a paper's contribution to knowledge, its role in clarifying long-standing problems, the number of people working in the targeted area, the exposition style itself, all factors that are not even remotely related to the length of the title. The basic point to be made here is that one can learn *nothing* about any single factor driving citation counts by running a regression whose focus is on that single factor. The coefficient of that factor cannot be identified since the impacts of all other relevant, but missing, variables driving the citation count are allocated to the coefficient(s) of the variable(s) included in the regression. The

parameter values of all missing drivers of citation counts are determined endogenously by the structure of the error term[17] and the decision on which variables were included in the model. It is easy to find examples where the significance level of an included variable (e.g., title length) goes from "highly significant" to "statistically irrelevant" as important missing variables are added to the model.[18]

In cases where important determinants of the dependent variable ($Y$) are missing and correlated with one or more of the $X$'s, there is almost always an endogeneity problem and without re-thinking the model, the coefficients of the included variables are not identified—are biased and inconsistent.[19] Once again, there is not only no causality but any correlations found are "faux" correlations which do not exist in the underlying populations being sampled. The solution to this type of endogeneity problem is straightforward: re-think the model, collect data on missing variables, and re-estimate the model.

Other types of endogeneity like simultaneity or measurement error in right-hand-side variables are not so simple to deal with.[20] Simultaneity arises when "$X$ causes $Y$" and "$Y$ causes $X$". The problem with simultaneity, and most endogeneity problem for that matter, is that neither adding more control variables addresses the problem nor increasing the size of the sample helps. Estimates will remain biased and inconsistent even as the sample size tends to be infinite. A famous and very simple case of simultaneity was widely discussed in economics over 70 years ago.

**Example 4 (Simultaneity).** At the time John M. Keynes published his "The General Theory of Employment, Interest and Money" (Keynes, 1938), econometricians began to estimate the equations within his revolutionary model of the macro-economy. One important equation in Keynes's model of the macro-economy is the consumption function which relates national income to consumption. It answers the question, if national income increases by one unit, what proportion would be spent on consumption? In its simplest form the consumption function is $Y = a + bX + e$, where $Y$ is consumption, $X$ is national income and $e$ is the error term. The coefficient of interest is the marginal propensity to consume from income at the economy-wide level, $b$. Unfortunately, no sample of $Y$ and $X$, no matter how large and no matter how accurately $X$ and $Y$ are measured, can accurately estimate the marginal propensity to consume by estimating the parameters of the single equation, $Y = a+bX+e$. Estimates of "$b$" (and of "$a$") will be biased and inconsistent. The reason is that there is a second, but missing, equation that links consumption, $Y$, to income, $X$. That equation is the identity $Y + Z^* \equiv X$, where $Z^*$ is autonomous investment in this simple model. So the consumption function determines $Y$, and $Y$ in turn in the second equation along with $Z^*$ determines $X$, i.e., $X$ and $Y$ are simultaneously determined by the two equations.[21] An estimate of the consumption function equation will overestimate "$b$" and lead policymakers to error in policy prescriptions. Few reading the current paper will be interested in this example from the history of economics but the identical statistical problem appears in all fields, simultaneous equation bias, and is a major reason for the failure of simple, one equation regression models.

We now provide several more examples of simultaneity bias all in seemingly different situations, but all leading to the same problem.

**Example 5 (Simultaneity).** A long-standing question in marketing research is to ascertain the impact firm advertising expenditures have on sales of a company's products. The problem is that advertising expenditures increase product sales and at the same time increases in sales typically

lead to increased advertising expenditures. In other words, sales and advertising are simultaneously determined in any realistic model of firm marketing decisions. Any attempt to use simple OLS regressions to determine the marginal impact on sales of a unit change in advertising outlays will cause this impact to be erroneously estimated and is doomed to failure, and depending on the circumstances, sometimes spectacularly so.[22]

Our discussion of simultaneity does not imply that there is no way of consistently estimating the desired parameters—there often is—but rather our discussion is meant only to warn practitioners of the futility of using simple, single equation methods in such circumstances. Our goal in this paper is to raise these issues in the hope that researchers will perhaps think more carefully about the processes that underpin the phenomena of interest and exercise more care early on at the level of model specification. The simultaneity problem is now widely recognized in the social sciences, at least at a theoretical level, but many practitioners seem to be unaware of the problem and its consequences. But the problem is ubiquitous as the next example should make clear.[23]

**Example 6 (Simultaneity).** There is considerable evidence that the quality of a country's institutions determines much of the variation in economic development across countries. For example, countries with secure property rights and the rule of law will be more highly developed, all else the same, than countries which are corrupt and rife with nepotism. One would likely find a strong correlation between the stage of a country's development and the quality of its institutions. Is it causal? Perhaps, but the data generally will not be able to provide an answer, since high levels of economic development are likely to drive institutional quality and high-quality institutions are very likely determinants of economic development. The two variables are simultaneously determined in a much more complicated model of economic development.[24] No simple, single regression equation will ever be able to determine the answer to the question. For interested readers, please refer to Acemoglu *et al.* (2001), where they provide an approach to identify the causal relationships using instrumental variable. In addition, Raghuram and Zingales (1998) adopt difference-in-difference approach to study a similar problem which examines the causal influence of a country's financial sector's development level on the level and the growth of its per capita income.

Another major cause of endogeneity arises when right-hand-side ($X$) variables are measured with error. In statistical models the left-hand-side variable ($Y$) is by definition measured with error. Unfortunately meaningful results from the usual approach (OLS) require the $X$ variables be accurately measured. It is easy to see why this is so.

**Example 7 (Measurement Error).** There are a number of studies claiming that "college is worth it even at high and rising tuition levels".[25] These studies claim to show higher wages and lower unemployment rates for college graduates. "On virtually every measure of economic well-being and career attainment—from personal earnings to job satisfaction to the share employed full time—young college graduates are outperforming their peers with less education".[26] But these studies appear to suffer from a measurement error problem. The population of students sampled—students who graduated from college—are a subset of all students who enrolled in college, some of whom dropped out and some of whom graduated. Those who graduate are a very special subset of those who enrolled in college—students who are on average both smarter and harder-working than the average student, which means they will make

more money in life whether they go to college or not. One way of characterizing this problem is as measurement error, arising because the sample does not measure the pertinent population—all students who enrolled in college as freshmen—rather it includes only those who graduated, i.e., only those who succeeded in obtaining a college degree are measured.[27,28] Unless studies control for differences across admitted college students in work ethic and cognitive and non-cognitive abilities, estimates of the wage premium will be overestimated and we have another instance of correlation, not causation.[29]

*Practitioner's Guide to Empirical Modeling*: *Suggestions*: In each of the situations we discussed above, and the accompanying examples used to illustrate the various problems the same conclusion holds: If the problems in question are not addressed, no conclusions can be drawn about causal connections between variables in the model. In some cases there may be correlations among variables and in others the correlations are false correlations because they do not exist in the underlying populations being sampled. The three fundamental causes of endogeneity problems mentioned in the various literatures are missing variables, simultaneity and measurement error.[30] For this reason we focused primarily on those mentioned, but we acknowledge that it is at times difficult to distinguish between them. For example, assume one wants to forecast the impact that education has on wages. Say education is measured as simply the number of years someone spent in certified educational institutions. But if there are differences in the impact on wages by type of education, one might describe the problem as one of measurement error in the education variable. Or the problem could be described as an omitted variables problem (variables which specify the types of education are absent). Alternatively, another reasonable hypothesis is that wages may well impact an individual's education decisions. If this is indeed the case, we might describe the problem as one of simultaneity.[31]

In the next sections, we focus our discussion on two types of issues: (1) those which are primarily statistical in nature, including over-fitting and p-hacking which lead to correlations that exist only in the data, but not in the population they represent; and (2) those which are econometrics related, which lead to biased and inconsistent inferences. In this section we explore procedures which working practitioners can do on a day-to-day basis to avoid these problems and get better, more reliable statistical results. Are there procedures that researchers in the field can employ that would mitigate the damage done from inappropriate estimation? That is the question.

*Statistical Issues*: Problems that are primarily statistical in nature can often be resolved or mitigated to some extent using largely statistical tools. A common solution which avoids "fitting" the data to the noise is to validate the model with a test data set. The test data set is often a subset of the data set in hand.

If the data set is large enough, one can split it into a "training/estimation subset of the data" and a "test/hold out subset of the data". One then, estimates the model using the first and validates the model using the second data set. If a model suffers from these issues, it won't perform well using the validation data set.

If the data set is insufficiently large to split, "cross-validation" could be employed in which the whole data set is split into, say, five parts. One then estimates the model using four parts of the data, and validates the model using the fifth part. Then returns the fifth part to the estimation set, and uses the fourth part for model validation. This is repeated five times, each time leaving one part of the sample for validation, and four parts for estimation.

These approaches should avoid ending up with a model that is fit largely to the noise in the sample. Unfortunately, if the issues are econometric in nature, this approach will not help.

*Econometric Issues*: Our endogeneity examples of simultaneity, errors in measurement and missing variables suggest that it is impossible to include enough control variables in a simple regression model to insure meaningful inference. We believe that in any field where reliable statistical estimation is an essential part of the work effort, researchers often misallocate their time. This holds in academia, in government where the duty of the analyst is to generate reliable input for policymakers and in the private sector where reliable estimates are essential for accurate forecasting and timely decision making. Sufficient time is not spent pondering an appropriate research design for projects. Rather, far too much time and effort go into collecting an abundance of control variables, when in retrospect many of those variables may be only remotely related to the phenomenon being investigated.

Our reading of the literature is that much more time and effort at the beginning of a project needs to be directed toward research designs which consider the various forms of endogeneity and how they might manifest themselves in the project at hand.[32] "More thoughtful research designs" involve thinking long and hard about the phenomenon being studied and the factors that likely are responsible for driving it. Theoretical models are often helpful to begin the search for potential $X$'s, and this implies thinking hard about the processes that generate not only the values of $Y$, but also those processes that determine the values of the $X$'s. Thinking deeply, not only about control variables, the $X$'s, but also about the processes that generate these data, is a first step toward identifying potential endogeneity problems early on—as the research design is being formulated.

Data are not merely numbers, since each number has a story behind it and needs to have a reason to be present in the sample. After all, simultaneity, missing variables, poorly measured variables and selection bias all originate processes that help determine the values of the $X$'s. Thoughtful reflection on the processes responsible for both the inclusion of right-hand-side variables and their values might help anticipate and aid in identifying sources of endogeneity bias at the outset.

One of the key reasons that "endogeneity" issues have been overlooked is that the data have been taken for granted and the underlying data generating processes have been ignored. In Example 5 above we have a clear instance where, if company processes which generate advertising expenditure data were known by the analyst, it would be apparent that running a regression of advertising expenditures on sales levels would not isolate the effect of advertising on sales.

Any attempt to address endogeneity problems must consider the following:

1. Understand the data generating process (DGP)

It is important to know how data are generated and any behavioral implications which may be tied to the values of each data point. Even more important, the analyst must understand what information is embedded in the data what information is missing, and the reason why it is missing. As the above examples make clear, mistakes often arise because of ignorance about what information is *missing* from the data.

2. If the researcher has no control over the data generation process, appropriate models need to be developed.

In many cases, especially in academics, data that have been generated earlier are used, and the analyst has no control over the DGP. Economists

have been dealing with this problem for decades and have developed methodologies in econometrics to help come to grips with it, including two and three-stage least squares, simultaneous equation methods, instrumental variables, Heckman selection models, natural/quasi-experiments, etc. Interested readers are encouraged to check out any *Econometrics* textbook to learn more. The fundamental idea in all of these techniques is to account for the DGP using more sophisticated models (involving more parameters) which account for more complex underlying processes.

This is one of the drawbacks to these methods: As the number of estimated parameters rise, coefficients are estimated less precisely, implying less accurate predictions. So if prediction and forecasting are paramount, and sample sizes are limited, there are situations where it may be preferable to estimate a model which ignores endogeneity completely. This is especially the case when: (1) predictions are short term[33]; (2) the DGP remains unchanged in the prediction period; and (3) the range of values of the independent variables are expected to change little. If any of these conditions are expected to be violated, a model which attends to endogeneity is necessary.[34]

It is worth repeating that, to the extent possible, researchers should rely upon domain knowledge and theoretical frameworks from which their empirical quest arises to guide them in their search for relevant explanatory variables.

3. If the researcher can control the data generating process, either via controlled experiments or, if possible, by designing A/B tests most endogeneity problems will vanish.

As we have suggested, the cause of endogeneity is the DGP and information missing in the available data. The best way of avoiding these issues is, if it is possible, to take over the DGP by running well-designed controlled experiments or A/B testing or natural/quasi-experiments (see Chang *et al.*, 2014 or Balakrishnan *et al.*, 2015). A rising number of big-data companies have recognized the importance of understanding causality and the value of running large-scale controlled experiments and A/B/C/…/ testing. The basic idea of these experiments is to use groups of subjects (such as users or customers), with no distinguishable differences between the groups, and to provide a different treatment to each group. As a result, any difference in outcome across groups can be attributed to treatment differences, and hence causality is established.[35]

Perhaps the reason that so little attention is paid to the processes which actually generate values of independent variables is that in many statistics courses there is scant attention, if any, to the difference between a regression model estimated from observational data and one estimated from a randomized experiment—the gold standard in academia, in government and in industry. The beauty of randomized experiments, and the reason that researchers often assume causality between right and left-hand-side variables, is that people or groups are assigned to different values of the $X$ by chance.[36]

It is plausible to argue that the primary benefit from the explosion in the use of the word "endogeneity" is that it may force researchers to internalize at least aspects of the problems of, simultaneity, missing variables, measurement error and related problems, and compel researchers to make their assumptions explicit. The single most important lesson for researchers who desire reliable, useable statistical analyses is to avoid overfitting and p-hacking and to embed their concerns with the various types of endogeneity early on during the design stage of a research project—including a full knowledge of the consequences for OLS regressions. The techniques

used for identification and for dealing with endogeneity are mathematical and statistical, but are often needed only because of problems arising from limitations in the design of the project.

## Notes

[1] Early private sector adopters of these methods include, so-called "legacy" users whose original business plans were based on statistics and data analytics, like Amazon, Google, Facebook and LinkedIn, and now include "followers" like Target, Macy's, Whole Foods and a myriad of other retailers, professional sports teams and health care delivery systems, and has spread to the "newbies" in not-for-profits, arts organizations and local governments.

[2] In this paper, we refer to "causality" as the fundamental connection that links "effect" to "cause". Temporally, cause precedes effect and should not to be confused with "Granger causality" which is focused on prediction accuracy and is neither necessary nor sufficient for causality as used here.

[3] To illustrate with an extreme case, one of us was presented with a model that "established" that recently in North Carolina, every increase of 10 attorneys led to 8 suicides from strangulation, suffocation and hanging. The data were from unimpeachable sources, the ABA and the CDC, from 1999 to 2009. The model had an $R^2$ of 0.93 and the $t$-statistic on the number of attorneys was 11.0.

[4] If there is an underlying theoretical model for the process being modeled, this problem is mitigated to some extent but can still occur. See *Practitioner's Guide to Empirical Modeling: Suggestions* below for further discussion.

[5] For example, this **occurs** when researchers use stepwise regression procedures to choose the model. Unfortunately, this is a surprisingly common occurrence.

[6] See *Practitioner's Guide to Empirical Modeling: Suggestions* section below for a discussion of statistical learning in which a subsample is held out for validation after the model is "trained," which in principle would obviate this problem.

[7] See Silver (2012) for an excellent non-technical discussion of this problem. A couple of Silver's comments are relevant here: Paraphrasing: Big data is increasingly blurring the line between correlation and causation, much to our detriment. And, big data has caused the ratio of signal to noise to fall (see page 163).

[8] Much more non-technical detail on overfitting can be found in Babyak (2004). This is not to say that there are situations, for example in very short-run forecasting problems, where high correlations may be sufficient. More on this later.

[9] A recent egregious example in medicine is discussed by Nina Teicholz (2014). She documents, in excruciating detail, that the cholesterol-saturated fat hypothesis, which ties heart disease to cholesterol and saturated fats has little statistical support. See f.n. 38 for more detail.

[10] Another example of the causality/correlation confusion appeared in the early medical research literature which linked lung cancer and tobacco use and pitted the eminent statistician R. A. Fisher against medical researchers. For an interesting account of these events and a discussion of the "battle" between Fisher and medical researchers on whether one could establish a causal link between tobacco use and lung cancer, *given the studies which had been doneat the time,* see Ellenberg (2014). Eventually it became clear that there was indeed a causal link. But R. A. Fisher argued strenuously that, *at the time*, there was not a causal link. Ellenberg's discussion of causation and correlation and other statistical topics are informative and wonderful reading.

[11] As is required for unbiased parameter estimation using simple, single equation methods.

[12] Econometric textbooks printed more than 50 years ago discussed this problem (For example, see Valavanis, 1959). A very early paper on an aspect of the problem was written by Haavelmo (1944).

[13] "Biased" means that estimates are systematically wrong on average and "inconsistent" indicates that even as the sample increases without bound the estimate does not converge to the correct value. This is a fundamental problem and one that "big data" cannot help.

[14] This usage refers to the problem of "accurately" estimating (identifying) underlying population parameters. As implied above, there are situations where this is impossible and only a well-designed experiment can accomplish the task.

[15] The apparent reason for this choice of explanatory variable is the total absence of any underlying theory which connects title length to paper quality. See *Practitioner's Guide to Empirical Modeling: Suggestions* below.

[16] The problem mentioned in f.n. 5, lack of an underlying model, is certainly at play here, where the exercise appears to be a statistical fishing expedition.

[17] Coefficient values will depend upon the signs and importance of the omitted variables.

[18] This is undoubtedly responsible for the "result" reported in f.n. 3.

[19] There is one important exception to this claim: The case when important determinants of the dependent variable ($Y$) are missing from the model and are **not** correlated with one or more of the $X$'s—this tends to be unlikely in practice.

[20] Still other endogeneity problems like unobserved heterogeneity and selection bias are cited in the various literatures but most can be viewed as missing variable problems. More below.

[21] Haavelmo (1943) was, to our knowledge, the first to discuss this type of estimation error. Valvanis (1959) made the Haavelmo results known to large audiences of graduate students.

[22] Still another example comes from the intersection of sociology, economics, marketing and psychology. In sociology there is something called the "Matthew's Effect." It posits that individual choices are determined directly by the characteristics and preferences of the individual which in turn are indirectly influenced by the social networks the individual is embedded in—the neighborhood, schools attended, the church, etc. In this case an individual's choices are simultaneously determined by the individual's preferences and the preferences of others in the individual's network of acquaintances. This is an instance of the "reflection problem" discussed by Manski (1993) in which he shows the rather dire effect on inference when average behavior in a group influences the behavior of individuals in the group. In the social sciences this phenomenon has many names "contagion", "bandwagons", "neighborhood effects", "imitation" …

[23] We have included a number of diverse simultaneity examples, since after surveying several literatures it appears that in most cases where "endogeneity" is mentioned, people seem not to be talking about omitted variables (which seems to be fairly well understood), rather they are worried about simultaneity and sometimes selection bias.

[24] This ignores many other issues, e.g., there are likely numerous factors both observed and unobserved that determine institutional quality across countries which would also need to be modeled.

[25] For example, see Georgetown Center on Education and the Workforce (2013) and Pew Research Center (2014).

[26] Georgetown Center on Education and the Workforce (2013) and Pew Research Center (2014).

[27] The studies calculate the earnings of those with a college degree and compare their earnings to those with only a high school degree. This is not the relevant population, if one desires measuring the earnings gap.

[28] This could also be labeled "selection bias" as the appropriate population was not sampled. Another possibility is to call it a "missing variables" problem since there are few if any measures of cognitive and non-cognitive abilities.

[29] Or suppose we are interested in the factors that determine wages and the working hypothesis is that education, experience and ability are prime movers. We may be able to gather good measures of education and experience, but will likely never find a good measure of ability and any proxy we use will measure ability with error.

[30] Selection bias (e.g., inferences made from polls based upon listener or user responses) and unobserved heterogeneity (variation among observations which are not measured) are also often discussed. Each of these can typically be thought of as measurement error problem or missing variables problem. (See discussion and example below.) A deep understanding of endogeneity problems means understanding the implications of unobserved heterogeneity whether one views it as a missing variable problem or a problem of measurement error.

[31] A compendium of many of the problems we have discussed in layman's language fills a recent best seller devoted to epidemiological studies which linked saturated fats and cholesterol to heart disease. The book, The *Big Fat Surprise: Why Butter, Meat and Cheese Belong in a Healthy Diet*, by N. Teicholz won the "one of best books of 2014" from *WSJ* and the*Economist* and *The American Journal of Clinical Nutrition* among others. A careful read shows that Teicholz lays out numerous examples of each of the issues we have explored, from p-hacking and overfitting to missing variables, selection bias, measurement error—all in easy to understand, non-statistical sentences. It is these epidemiological studies, accepted as "science," which led to the low-fat/high carbohydrate diet promulgated in the US for the past 50+ years.

[32] In all cases, endogeneity results in biased and inconsistent parameters if simple OLS procedures are used.

[33] Short-term or long-term are defined relative to the data observation period. Roughly, if the data are collected over a one-year period, a prediction for the next month is probably short-term, while a prediction for the next two years is long-term?

[34] Here may be situations, the Lucas Critique, where prediction depends upon identification.

[35] Of course, external validity concerns remain.

[36] Some authors seem to say that since they have controlled for a vast number of factors, the values of independent variables are automatically randomly assigned. But this is precisely the problem: people don't think about how these variables are generated.

## Acknowledgments

Dong and Heineke founded the Business Analytics programs at the Leavey School of Business, Santa Clara University, Santa Clara, CA. The authors are professors of Marketing, and Economics, respectively. We would like to thank Ke Wang, Ye Cai, George Chacko, Sanjiv Das, Carrie Pan and Bill Sundstrom for their valuable comments. Any remaining errors are their responsibility.

## References

Acemoglu, D., Johnson, S., and Robinson, J. (2001). *The American Economic Review* **91**(5), 139–141.

Babyak, M. A. (2004). "What you see may not be what you get: A brief non-technical introduction to overfitting in regression type models," *Psychosomatic Medicine* **66**(3), 411–421.

Balakrishnan, K., Billings, M. B., Keley, B., and Ljungqvist, A. (2014). "Shaping liquidity: On the causal effects of voluntary disclosure," *Journal of Finance* **69**(5), 2237–2278.

Chang, Y.-C., Hong, H., and Liskovich, I. (2015). "Regression discontinuity and the price effects of stock market indexing," *Review of Financial Studies* **28**(1), 212–246.

Derksen, S. and Keselman, H. J. (1992). "Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic noise variables," *British Journal of Mathematical and Statistical Psychology* **45**(2), 265–282.

Ellenberg, J. (2014). *How Not to Be Wrong: The Power of Mathematical Thinking*. Penguin Press.

Georgetown Center on Education and the Workforce (2013). *The College Payoff*.

Haavelmo, T. (1943). "The statistical implications of a system of simultaneous equations," *Econometrica* **11**(1), 1–12.

Harvey, C. R., Liu, Y., and Zhu, H. (2015). ". . . and the cross-section of expected returns," *Review of Financial Studies* **29**(1), 5–68.

Ioannidis, J. P. A. (2005). "Why most published research findings are false," *PLoS Medicine* **2**(8), e124.

Keynes, J. M. (1938). *The General Theory of Employment, Interest and Money*. Palgrave Macmillan.

Letchford, A., Moat, H., and Preis, T. (2015). "The advantage of short paper titles," *Royal Society Open Science* **10**(1), 1–8.

López de Prado, M. (2015). "The future of empirical finance," *Journal of Portfolio Management* **41**(4), 140–144.

Manski, C. (1993). "Identification of endogenous social effects: The reflection problem," *Review of Economic Studies* **60**(3), 531–542.

Mundry, R., Roger, R., and Nunn, C. (2009). "Stepwise model fitting and statistical inference: Turning noise into signal pollution," *American Naturalist* **173**(1), 119–123.

Pew Research Center (2014). *The Rising Cost of Not Going to College*.

Raghuram, R. and Zingales, L. (1998). "Financial dependence and growth," *The American Economic Review* **88**(3), 559–586.

Silver, N. (2012). *The Signal and the Noise*. Penguin Press.

Teicholz, N. (2014). *The Big Fat Surprise: Why Butter, Meat and Cheese belong in a Healthy Diet*. Simon and Schuster.

Valavanis, S. (1959). *Econometrics*. McGraw-Hill Book Company.

*Keywords*: Causation; correlation; overfitting; p-hacking; endogeneity; identification; simultaneity; measurement error; data generation processes; inconsistency; bias