

Research Note

Three Common Myths in Quantitative Social Research

Tony Tam *

致謝：

This research was supported by grants from the Research Grants Council of Hong Kong (CUHK 3/93H) and from Academia Sinica through the Organization-Centered Society Project, the Learning 2000 Project, and the Institute of European and American Studies. I thank Ly-yun Chang, Min-hsiung Huang, and the anonymous referees for comments. Direct correspondence to Tony Tam (譚康榮), Academia Sinica, Institute of European and American Studies, Nankang, Taipei, 11529, TAIWAN (email: tam@sinica.edu.tw; tel: (02)2789-9390x249).

* Tony Tam Associate Research Fellow, Academia Sinica

收稿日期 2001/6/1 • 接受刊登 2001/11/27



Three Common Myths in Quantitative Social Research

Abstract

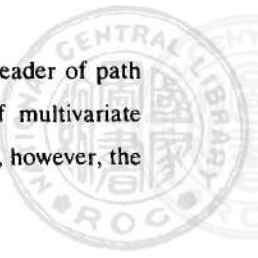
Significance testing and multivariate models are revolutionary tools for quantitative social research. However, even great tools have the potential for misuse. This paper discusses three common ways in which these fundamental tools have been misused: (1) misinterpreting significance tests, (2) a misplaced focus on net effects, and (3) overconfidence in the control variables. For each misuse, I identify a conceptual myth, specify the source of the misconception, and suggest a remedy. In addition, numerical data on college graduates and a recent debate over gender inequality in the American labor market illustrate these problems and remedies. The problems and remedies are applicable to all kinds of multivariate analysis based on statistical modeling (e.g., OLS regression, logit, tobit, other limited dependent variable regressions, simultaneous equations models, and so forth).

Key Words: Significance Test, Multivariate Model, Statistical Control, Measurement Error



The history of modern sociology is in part a revolution driven by the use of statistics in quantitative research. Statistics has demonstrated how it is logically possible for social scientists to infer population reality from limited sample data. Statistics has also developed multivariate methods with two fundamental objectives and contributions: (1) The first contribution of multivariate methods (multiple regression and all of its extensions) is to enable social scientists to do what case studies and ethnographic methods are by design unable to accomplish — *conduct statistically controlled comparisons of outcomes across a complex range of conditions and scenarios*. (2) The second contribution is to *specify the structure of causal relationships among multiple variables* — in many ways the ultimate objective of social science inquiry. Before the invention of multivariate modeling, this kind of specification proved to be an impossible challenge even for a couple of hundred cases and half a dozen variables. For whatever their worth is, qualitative methodologies (such as ethnography) cannot handle the multivariate specification problem. The relationships among multiple variables often involve questions of how each variable may be dependent on (i.e., endogenous to) some of the other variables directly and indirectly. The complex controlled comparisons required for specifying multivariate relationships would be impractical without the aid of multivariate statistical methods and a computer.¹

¹ This objective of multivariate analysis would probably remind the reader of path analysis (Hanushek and Jackson 1977:217-243) — the epitome of multivariate methods. To fully utilize the elegant features of classical path analysis, however, the



For better or worse, a lot of the published articles in top sociology journals employ the tools of statistical methodology. They are indispensable for quantitative research in sociology, ranging from education to criminology, from stratification to ethnic relations, from organizational behavior to health behavior, and from small group analysis to network analysis. Sound methodology to empirical social scientists is akin to proper tools and operating procedures to surgeons. Methodology, both qualitative and quantitative, is about matters of life and death in empirical research. Innovative methods can turn an impasse into a promising research frontier. But lousy procedures can destroy the validity of an empirical study. It is therefore dangerous to let bad practices perpetuate themselves. Indeed, some fundamentally bad practices of quantitative research have escaped the attention of gatekeepers, even those of elite sociology journals. The first step toward eliminating any bad practice is the public recognition of what the bad ones are.

Although numbers do not lie, numbers may fail to tell a story that many researchers think they do. The problem usually is not with numbers *per se*. Numbers may be easily misinterpreted when researchers are blind-sighted by unacknowledged myths. This paper calls attention to three pervasive myths that have had adverse impacts on the quality of quantitative social research.² The perpetuation of any of the myths could

user must be prepared to accept many strong assumptions. But one can do multivariate analysis without adopting many of the assumptions and decomposition techniques of classical path analysis (see, e.g., Agresti and Finlay 1986:292-307).

² It is unfortunate that the myths are pervasive in the sociological literature to date,

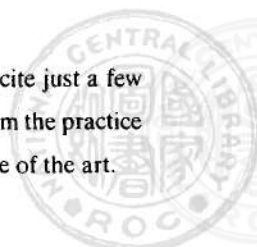
(1) undermine the validity of substantive conclusions and (2) prevent the effective utilization of available data to fully address the questions posed in a study. The purpose of this paper is to help realize the full potential of, rather than discourage, quantitative research. I will deconstruct the common myths by explicitly identifying three problematic practices in quantitative research, specifying the underlying sources of misconception, and suggesting ways to eliminate them.

For simplicity and familiarity, I will illustrate the problems and solutions with OLS regression for continuous dependent and independent variables. However, it should be emphasized that the issues raised and the remedies proposed are relevant to virtually *all kinds* of multivariate modeling concerned with the structural effects of any variable, irrespective of the mix of measurement scales, sampling design, and statistical techniques used (e.g., OLS regression, binary probit, multinomial logit, tobit and other limited dependent variable models, simultaneous equations models, and so forth).

Numerical Example

A numerical example will help fix ideas. For the purpose of this paper, it is especially useful to use simulated data. The primary advantage of simulated data is that the true generating mechanism for the data and thus all the relevant structural parameters are known without any

even among the most selective journals. Indeed, it would be unfair to cite just a few papers and not to mention so many others. I would therefore deviate from the practice of citing examples or sources to support negative remarks about the state of the art.



uncertainty. Knowing the true model parameters is essential for the purpose of the numerical example. In real data, it would be difficult to unambiguously demonstrate the undesirable consequences of the conventional practices to be discussed in this paper.

To be concrete, consider a comparative study of two countries that appear to follow different labor market regimes. The substantive interest is in the role of personal ability and educational attributes in the determination of hourly wages in two very different countries: A and B.³ Employers in country A tend to emphasize pre-hiring testing and screening of job candidates, and wage levels reflect on-the-job performance of individual workers. By contrast, wages in country B are largely attached to jobs and initial job assignment determines a relatively predictable wage growth path. Job candidates appear to form a queue that is stratified according to the quality of the schools from which they last graduated. Employers of good jobs (high-paying and good wage growth) prefer graduates from high quality schools.

Now assume that there are two exactly parallel samples of educational attainment. For each country, a random sample of 1,000 students is selected from all students who graduated from college in 1990. The dependent variable (Y) is the logarithm of hourly wage. The first independent variable (X) is a composite measure of indicators and proxies

³ The myths to be discussed in this paper go well beyond this substantive interest. The issues are not confined to school effects, cross-national comparisons, or wage determination. Virtually any quantitative study of education, stratification, and other areas of sociology would have to come to grip with the same set of traps, problems, and remedies.

for personal ability (such as family background, cognitive skills measured by national college entrance exams, and academic achievements). The second independent variable (Z) is a composite measure of the commonly assumed quality attributes of a college (such as faculty-student ratio, average salary of professors, average prestige of graduate departments, average test scores of entering cohorts). By virtue of conceiving X and Z as composite variables without much constraint on their internal structures, the model for Y is actually much richer and more general than it seems. The two variables may be conceived as representing two blocks of independent variables, each block playing a different role in the generation of Y depending on what the true model is.

For the purposes of illustration, I have generated simulated data for each country: e, u, X, Z, Y .⁴ In fact, I have taken full advantage of simulated data in order to facilitate the comparison and interpretation of any divergent numerical results between countries. First, all but the dependent variable are set to be identical for the two countries. So the basis of comparison is exactly matched for the two countries. Second, the wage determination mechanisms are different across countries. Specifically,

$$\text{Country A: } Y_A = 5 + X + e \quad [1]$$

$$\text{Country B: } Y_B = 5 + X + Z + e \quad [2]$$

where

⁴ Simulation was conducted with the random number generators of STATA 6.0 for normal and uniform distributions.

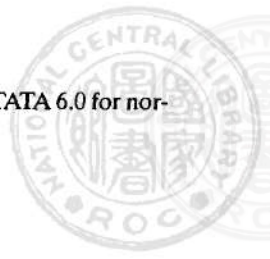


Table 1. Two Conventional Formats for Presenting Regression Results:
Simulated Data on Log-wage, Ability, and College Quality (N=1,000)

Independent Variables	Country A (1)	Country B (2)
Personal Ability (X)	1.08	1.08 (1.34)
College Quality (Z)	-.01	.99* (4.93)
Constant	4.84*	4.84* (29.32)
R-squared	.03	.39

The dependent variables for columns (1) and (2) are Y_A and Y_B , respectively.
T-ratios are in parentheses.

* Significant at .05 level.

X ranges between 0 and 1, with mean = .51 and std. dev. = .29

$Z = 4X + u$ with mean = 2.55 and std. dev. = 1.16

X and Z are strongly correlated with Pearson $r = .97$, $r^2 = .94$.

Table 1 presents the estimation results of OLS regression for the two countries. To present this kind of estimation results, many quantitative studies adopt the format illustrated by column (1) of table 1, the model for country A (Y_A). The column lists the parameter estimates and attaches a flag to signal if an estimate is significantly different from zero according to a statistical criterion (e.g., $p < .05$). Another common practice is to also present the test statistic associated with the parameter estimates, as in column (2) of table 1, the model for country B (Y_B). Both practices reflect an unwarranted interpretation of significance test — the absence of statistical significance means the absence of an effect. This interpretation

is based on the first myth that plagues quantitative social research and will be discussed in the next section.

Although the model in column (1) appears to be overly simplistic, it covers a general class of scenarios where X denotes all truly relevant predictors for Y_A , and Z represents all potentially relevant predictors in the data but actually irrelevant for Y_A . Table 1 illustrates an emphasis on the estimates for what is presumably the full model given the available data, ignoring and not presenting any of its submodels. The underlying belief is that accurate substantive interpretation of evidence entirely hinges on estimated net effects from a full model, i.e., effects that have taken into account the confounding influences of other independent variables. This belief is central to the second myth.

Finally, the third myth is related to the belief that the model specification in column (1) is safer than one that omits Z , even though the true model is given by equation [1] and does not contain Z . The assumption behind the belief is that a bigger model is safer than a smaller one. Put differently, a model is safer than its submodel. A perfect substantive example is exemplified by the recent debate over the sources of occupational sex composition effects on wages (Tam 1997, 2000; England, Hermsen, and Cotter 2000). This is the focus of the third myth.

Myth 1: Safety of Significance Tests

Background. Classical significance tests are designed for simple hypotheses in the form of whether a parameter of interest is zero. This kind of hypothesis does not in itself tell us much about most kinds of

sociological research questions, even though it is the necessary building block of practically all quantitative studies.⁵ The subtlety of this distinction seems to have escaped the attention of many researchers (both authors and referees). Many quantitative studies have chosen to report and emphasize the significance tests (the sign and p-value of each test statistic) as the main results of statistical modeling. The implicit assumption is that the test results are the most clear-cut findings after fitting a statistical model for a dependent variable.⁶

Many researchers apparently presume that *significance tests are the most straightforward part of quantitative results, with little room for disagreement or ambiguity*. The reason appears to be the myth that the significance test of an estimated effect tells us whether the effect is large enough to warrant attention. In fact, it is dangerous to interpret an insignificant result without also attending to the size and standard error of an estimated effect. Neither the test statistic nor the p-value associated with it would tell us what exactly is going on.

⁵ Significance tests are fundamental to the entire discipline of statistics and integral to the generalization process in empirical knowledge building. Virtually all empirical sciences have to use significance tests. From medical to psychology laboratories, from quality control departments of factories to census bureaus, from engineering research to sociological studies, significance tests are part of the nuts and bolts of quantitative inquiry. Significance tests enable the cost-effective use of relatively small samples to study large populations.

⁶ After all, the size of an estimated effect may be difficult to interpret, especially when the dependent variable does not have a natural scale of measurement as does income or years of education.

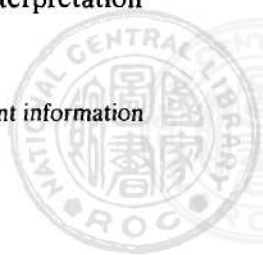
Source of Misconception. To understand the problem with the myth, we must recognize the fact that all commonly used significance tests for estimated effects are based on two components: (1) the size of an estimated effect and (2) the standard error of the estimator. The size of an estimated effect indicates how large the true effect may be. Its standard error quantifies the degree of cross-sample variability in the estimated effect if we were to infinitely replicate the sampling and estimation process. Cross-sample variability reflects the amount of statistical information in the data useful for the estimation (large standard error means lack of useful information), which is partly driven by sample size and often driven by the structure of the model estimated.⁷

When a test rejects the usual null hypothesis of nil effect, we know that the estimated effect must be large relative to the standard error. But the story is no longer clear when a test fails to reject the null hypothesis. In fact, there are two nonexclusive reasons for failing to reject the null:

- (1) the *size* of an estimated effect may be absolutely small and so statistically indistinguishable from zero; and
- (2) the estimated effect may be absolutely large, but the *standard error* (sampling uncertainty) is so large that the data do not provide enough information to precisely distinguish the estimated effect from zero.

The two reasons entail drastically different substantive interpretations. The first and the standard reason implies the interpretation

⁷ Even a large sample of millions of cases may contain very little relevant information to identify a model of interest.



that the corresponding independent variable has no direct effect in the model. By contrast, the second and the often overlooked reason implies that even if the independent variable has a strong effect, the data are not rich enough to precisely estimate the effect in this particular model.

Both the data and the model specification may be the real reason for getting an insignificant test result even when the null hypothesis is wrong. The lack of precision may be due to, for instance, the collinearity between this particular independent variable and some related indicators of the same underlying concept in the model — a consequence of model specification rather than any inherent weakness of the data. *The tricky part is that simply looking at the p-value would not tell us which reason is behind a test result.* One must dig into the details and even try estimating some alternative models to find out which reason is driving the test result. Unfortunately, the sociological literature is not alert enough to problems of this kind because too many sociologists seem to have taken the first reason for granted.

Part of the blame should go to statistical textbooks. The best caution textbooks usually provide a student of significance tests is that statistical *significance* should not be confused with substantive *importance*. An estimated effect can be small but statistically significant or large but statistically insignificant. This is undoubtedly a sound and crucial distinction for all users and consumers of significance tests to take heed. Nonetheless, some researchers continue to ignore the distinction, continuing to misinterpret a *statistically* significant effect as evidence for a *substantively* large and important effect. Most important, the distinction does not go

deep enough. It does not guard users of significance tests against the danger of misinterpreting an insignificant test result as most likely the outcome of a small and unimportant effect. Nor does it suggest practical ways to sort out the divergent reasons for an insignificant result.

Remedy. According to textbooks, the statistical significance of a test is summarized by how small the p-value is. Many quantitative studies in Taiwan tend to emphasize the level of statistical significance in reporting findings. To cautiously interpret an insignificant result, however, we must attend to both the *size* and *standard error* of an estimated effect.⁸

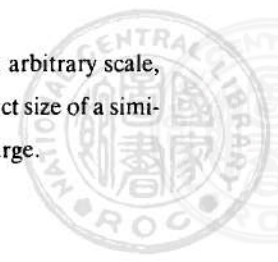
First, paying attention to the size of an estimated effect is essential for determining its substantive importance. Three factors should always

⁸ This approach stands in stark contrast with standardized regression in which an estimated effect is expressed and evaluated in terms of the standard deviation unit of the dependent variable and that of the independent variable. Standardization makes sense when the dependent variable and all independent variables are measured with arbitrary metrics. Otherwise, the standardization approach actually makes interpretation more indirect and far removed from the substantive units of measurement, hence making it more difficult to gauge the substantive importance of an effect size. In addition, since standardization is based on the sample variance of each variable, the size of standardized regression coefficients is a function of the sample variance of each variable. Sample variance is a tricky matter that can be the source of much confusion. Even during the heyday of classical path analysis that helps popularized the use of standardized regression coefficients, Hanushek and Jackson (1977:78) have rightly pointed out the severe problems of standardized regression. To date, very few articles in the very top sociology journals would use standardized regression, not to mention only present standardized regression coefficients. Most referees would have demanded the presentation of raw regression coefficients together with their standard errors.

be taken into account when assessing the size of an effect: (1) the scale of the dependent and independent variables (e.g., respondent's annual earnings and respondent's years of schooling), (2) the functional form of a model (e.g., linear versus loglinear), and (3) the substantive context (e.g., 7% per year of schooling is close to most estimated earnings return to schooling, implying that an average college graduate would earn $7 \times 4 = 28\%$ more than an average high school graduate).⁹ Much of the distinctive contribution of quantitative research rests on its ability to quantify effects that case studies and ethnographic approaches are ill equipped to do. Not taking the size of an estimated effect seriously would risk giving up a big portion of the unique contribution of quantitative research.

Second, careful assessment of the standard error of an estimated effect would usually tell us the precision with which the effect is estimated, and signal if the model is ill conditioned or improperly specified. One of the easiest diagnostics is to compare the standard error for an independent variable across models. Consider, for instance, the scenario where the standard error for Z of one model is large when compared to its standard error when Z is the only independent variable in another model. While the standard error in the first model is usually complicated by the presence of a host of other independent variables that are often correlated with Z , the standard error in the second model is driven solely by the true effect and sample size. More generally, if the standard error of Z in a

⁹ If the scale of either the dependent or independent variable has an arbitrary scale, such as a five-point attitudinal scale of agreement, one may use the effect size of a similar scale in the same model to judge if a particular effect is relatively large.



more complicated model is much larger (say, 50-100 percent higher) than the one in a simpler model, it is a signal that the data are unable to provide much information for identifying the effect of Z. The reason must be due to its correlation with some of the other independent variables. If the size of effect has substantially increased while the standard error also becomes much larger to render the test insignificant, then this is likely a paradigmatic example in which the conventional reading of significance test results would be misleading. We should closely examine the possibility of both an ill-conditioned model and a really substantial effect of Z.

In light of the principles just described, quantitative studies published in the *American Sociological Review* and *American Journal of Sociology* almost always report both the raw estimated effects and their standard errors, instead of reporting test statistics (usually the t-statistic, sometimes the z-statistic) or p-values. Statistical significance is usually flagged with an asterisk and few readers would really care about the exact p-value. Reporting standard errors is a direct way to quantify the extent of statistical uncertainty, its validity is independent of the validity of any test statistic used. Standard errors tell us the same message in all situations, no matter what the appropriate test statistic should be and no matter what its sampling distribution looks like. This is a sound practice that should be followed by all sociologists. Moreover, before concluding that any key independent variable has no effect on the dependent variable, a researcher should *routinely ascertain whether the insignificant effect is due to an absolutely small coefficient with a small standard error*. If not, the conclusion should be qualified accordingly, or the result should be deemed as inconclusive.

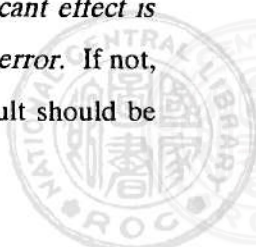


Table 2. Recommended Multivariate Analysis and Presentation of Results:
Simulated Data on Log-wage, Ability, and College Quality (N=1,000)

Independent Variables	Country A		Country B		
	(1a)	(1b)	(2a)	(2b)	(2c)
Personal Ability (X)	1.08 (.80)	1.06* (.20)	1.08 (.80)	4.92* (.20)	
College Quality (Z)	-.01 (.20)		.99* (.20)		1.26* (.05)
Constant	4.84* (.16)	4.83* (.12)	4.84* (.16)	5.42* (.12)	4.72* (.14)
R-Squared	.03	.03	.39	.38	.39

The dependent variables for columns (1) and (2) are Y_A and Y_B , respectively.

Standard errors are in parentheses.

*Significant at .05 level

Example. It is instructive to look at the simulated data. We know that equation [1] represents the true model for Y_A . But if one looks only at the estimated results in table 1, the conclusion for Y_A based on a naive reading of significance test would be that neither X nor Z has any significant effect on Y_A . However, this conclusion is premature. Table 2 takes a different approach to presenting the results. Column (1a) of table 2 replicates model (1) of table 1, but also presents standard errors for the estimated effects. Since the data are simulated, we know the true parameters underlying the data. This information is useful here. Notice that the estimated coefficients of X and Z (1.08 and -.01) are remarkably close to the true values (1 and 0). However, the standard error of the estimated effect of X is very large — .80, which is almost as large as the true parameter and

nearly three times the standard deviation of X.¹⁰ The same does not apply to the effect of Z and the intercept term. The data are *not* rich enough to simultaneously estimate the effect parameters of X and Z. An alternative look at the same thing is to compare the standard errors of the effect of X in columns (1a) and (1b) and they are strikingly different — .80 versus .20. The standard error for X in model (1a) is four times its standard error in model (1b), strongly suggesting that the imprecision of estimate is not due to insufficient number of cases but due to the structure of the model being estimated. Indeed, the strong correlation between X and Z ($r = .97$) is the reason why there is a substantial loss of the statistical information available for the estimation of the net effect of X.

Consequently, despite the fact that the OLS regression model of column (1a) provides very accurate point estimates of the true effects, the significance test indicates that both the estimated effects are insignificantly different from zero. While the conventional conclusion appears to be fine for the effect of Z,¹¹ the conclusion would be misleading if mechanically applied to the effect of X. The insignificant result for the effect of X should not be taken as evidence for the null hypothesis (nil effect). The proper reading of the estimation results would emphasize that the data contain insufficient information to precisely estimate the

¹⁰ To get a sense of the magnitude of these estimated effects, it is worth noting that the standard deviations of the independent variables X and Z are .29 and 1.16, respectively.

¹¹ It is proper to interpret the insignificant result for Z as reflecting a nil effect because the standard error of the estimate (.20) is only a fraction of the standard deviation of Z (1.16). There is no indication of a particularly inflated standard error, hence no lack of statistical information to come up with a reliable test.

effect of X in country A.

A similar case can be made about the effect of X in country B. What have we learnt? The significance tests for the effects of X are simply far from conclusive. A naive reading of the test results can indeed be highly misleading. Substantively, a naive reading would have suggested that only college quality does not matter in country A but does in country B when, in fact, personal ability (X) plays a crucial role in both countries. However, only when the reader is presented with the standard errors of the estimated coefficients does s/he have what it takes to arrive at this proper conclusion. This is why the remedy adopted by the leading sociological journals is safe and well-justified.

Myth 2: Safety of Net Effects

Background. Statistical control is one of the revolutionary achievements of social science methodology. Ever since the invention of multiple regression, quantitative analysis has made a huge contribution to the accumulation of social science knowledge. To date, we can hardly find a quantitative sociological study that does not rest its case on multivariate models. There are, however, dangerous traps for users of multivariate models.

Published quantitative studies here and abroad often *emphasize and focus on "full" models to draw substantive conclusions — models that have already included all explanatory and control variables available in the data.*¹² Submodels are often considered inferior and irrelevant except for the purpose of showing the impact of adding or dropping some of the key

independent variables. Underlying this focus on full models is the apparent belief that the net effects estimated from a full model are the safest and most informative numbers to look at, because these are the best or most reliable estimates of the true effects of an independent variable of interest. The belief is a myth: the net effects alone are not the most informative numbers for understanding multivariate relationships and the absence of net effect does not imply the absence of causal effect.

Source of Misconception. Why are the net effects from a full model thought to be the safest estimates? The standard answer is that the net effects are the least likely contaminated by spurious correlation. Under normal assumptions, the net effects are indeed most likely free from spurious correlation. However, no matter how many controls a model includes, a single-model analysis cannot substitute for a multi-model analysis of multivariate relationships. To understand this, one must accurately understand what a single multivariate model can and cannot do for us.

Two facts are key to a proper understanding. First, the true total effect of a variable is the sum of direct and all indirect effects.¹² The more variables are included in the model, the further a net effect may be from the total effect of an explanatory variable because more of its total effect

¹² Fitting a full model is never a well-defined goal of social science modeling because the idea of a full model is ill defined. No model is complete in any absolute sense. As a practical matter, a full model can only be defined relative to a specific question and the available data, i.e., a model that has included all the relevant explanatory variables available in a given data set.

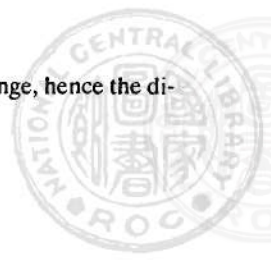
¹³ Unlike total effect, direct and indirect effects are relative concepts. The direct effect of X on Y, as it is normally used, refers to the net effect of X in a specific model for Y.

may be accounted for as indirect effects. Second and most important, a single-model analysis can never specify both the direct and indirect components of a total effect.

A single model can only tell us the net effect of each independent variable controlling for the presence of all other independent variables in this specific model. A net effect is just part of a total effect. If there is no net effect of X in a model, it does not imply that X does not have a causal effect on the dependent variable. The causal effect of X may be fully mediated by some of the other independent variables in the model. Thus a single multivariate model cannot tell us the net effect of X after adding or dropping any other variable from the current model, not to mention decomposing for us the total and indirect effects of X .

Remedy. Multivariate methods enable us to address the relationships among multiple endogenous variables using multiple specifications of single-equation models. Multiple regression and logistic regression, for instance, are the most common examples of single-equation models that can be adapted to do the kind of multivariate analysis called for by modern social science. The trick is simply to estimate and present not only the most complete model feasible with the available data, but also submodels that would indicate the impacts of control variables on the coefficient of each key independent variable, and how the key variables may mediate the effects of each other on the dependent variable. Such a multi-model

If the model is expanded or reduced, the net effect will generally change, hence the direct and indirect components of the total effect of X on Y .



analysis is anything but routine, especially when many explanatory variables are involved. It challenges a researcher to have a well-defined conceptual map of the theoretical relationships among all the variables before the researcher can decide which submodels would be useful to estimate. *But the basic point is simple: we must estimate multiple models to specify multivariate relationships.*

Unfortunately, statistical textbooks have not emphasized the fact that the crowning accomplishment of multivariate methods is not just in enabling statistically controlled comparisons but also in the specification of multivariate relationships. All textbook discussions of multivariate models should emphasize the methodology of how to use multi-model analysis to map out multivariate relationships. This methodology is one of the most powerful components of the basic toolkit for quantitative research.

Example. In table 2, the models estimated for Y_B provide an illustration of the use of multi-model analysis involving two independent variables X and Z . Suppose we know from the theoretical and substantive context that X may be mediated by Z but not vice versa. Column (2a) of table 2 is a replicate of column (2) of table 1, columns (2b) and (2c) are submodels motivated by the theoretical assumption that X may be mediated by Z but not vice versa. With the model in (2a), the point estimates can closely approximate the true coefficients of X and Z (both are exactly 1). But from the results there, we can only be confident that Z has a net effect. The effect of X is ambiguous. Nevertheless, if we look at (2b) that examines the total effect of X , the picture becomes much clearer.

First, the total effect is about five times as large as the net effect and

is absolutely large. Second, the total effect is statistically significant at any conventional level of significance. Third, the standard error is just about one-fourth of that for the net effect in model (2a). The direct effect X is difficult to estimate precisely because of the correlation between X and Z , but the finding does not mean that the direct effect is absent. When we also look at model (2c), the effect of Z is 25 percent larger than the net effect in (2a) and the standard error is very small. Some of the Z effect in (2c) is spurious of the omitted variable X . In short, X has a strong total effect but four-fifth of which is indirect via Z .

In this example, I have demonstrated the usefulness of examining both standard errors and multiple models in quantitative analysis. Sidestepping either one would severely restrict the ability of a researcher from getting the most out of the data. It is high time that quantitative researchers and referees routinely attend to the need for both kinds of information. Quantitative research will be significantly richer and more informative than under current norms.

Myth 3: Safety of More Control Variables

Background. In the introduction I have emphasized two major contributions of multivariate modeling: (1) enabling statistical control and (2) specifying multivariate relationships. The myth of net effect is an unfortunate side effect of over-emphasizing the first contribution while downplaying the second contribution. The main consequence of the second myth is to mislead many researchers to focus on net effects in "full" models, even at the expense of specifying multivariate relationships. But

there is another side effect: many quantitative researchers believe that it is best to add as many control variables as possible to a model in order to get a safe estimate of net effects that are least likely contaminated by spurious correlation. This is what Tam (2000) dubbed as the bigger-is-better (BIB) assumption.

The BIB assumption is the third myth that all producers and consumers of quantitative research should understand. This myth can lead to a reckless preference for larger models in the name of minimizing the influence of spurious correlation. Lieberman (1985) has devoted a whole book to debunk the common preference for adding control variables, and illustrated with many substantive examples the pitfalls of mechanical application of the control variable approach to multivariate analysis. Yet the BIB assumption keeps recurring in the quantitative social science literature and big models with a long series of control variables continue to be well represented even in the leading sociological journals.

The problem with the BIB assumption is never about the number of variables that enter a model. Instead, it is the discrepancy between what a researcher wishes the controls would do and what the controls actually do in a model. A bigger model may be highly sensible and a smaller model may be nonsensical depending on the structure of a model. The important thing is to watch out for the traps of control variables that textbook treatments of statistical control simply assumes away from the outset.

Source of Misconception. Paradoxically, the main source of the myth is that the BIB assumption is perfectly safe and sound provided that some ideal conditions are satisfied. Indeed, statistical textbooks always make these ideal assumptions in discussing multivariate control. The trouble is

that most researchers do not recognize what the ideal conditions are or what serious consequences would ensue if any of the ideal conditions fails. This ignorance may well be one of the major reasons for the pervasiveness of the BIB assumption.

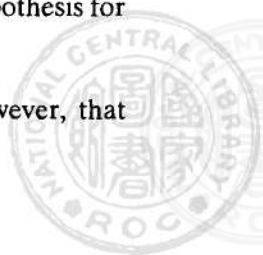
Fortunately, Tam (1997:1660-1662, 1685) offers a non-technical discussion of this issue in an attempt to alert sociologists to the damaging consequences of the BIB assumption. More recently, Tam (2000: 1756-1757) provides a formal explanation of the ideas necessary to understand the pitfalls of the BIB assumption. Without attempting to reproduce the exposition of Tam (1997, 2000), suffice to say that there are two crucial ideal conditions for the BIB assumption to hold. First, sample size is sufficiently large so that there is enough statistical variation to identify the true effects of correlated independent variables. Second, there is no measurement error in those independent variables correlated with any the explanatory variables of interest (e.g., occupational sex composition in the case of Tam [1997]).

In the absence of measurement errors and with sufficiently large sample size, adding any control variables would not distort our ability to test a true model. In a finite sample, however, even in the absence of measurement error, adding control variables can take its toll on model estimation. Adding control variables that are correlated with an explanatory variable would always increase the standard error of the estimated effect by reducing the amount of effective statistical variance available for analysis. A comparison of columns (1a) and (1b) of table 2 would demonstrate the point. Without the inclusion of Z, the estimated

effect of X is sharp and precise. Even though Z is by design an irrelevant control variable, adding Z to the model has greatly diminished our ability to estimate the coefficient of X — there is an increase of the standard error from .20 to .80. The moral is that it is ill advised to add control variables without good justification. Even with good justification, it is important to also attend to the implications of the diminished ability to estimate the parameters of interest. What appears to be a precautionary move (i.e., adding as many control variables as possible) could end up distorting the results and make things worse.

In the presence of measurement error, the problem can be devastating. The recent literature on the nature of the wage penalty against female-dominated occupations provides a dramatic example. Using a combination of methodological critiques, powerful new findings, and a formalization of competing interpretations, Tam (1997) shows that noisy correlated independent variables have produced critically misleading findings in the prior literature. Moreover, Tam (2000) is able to turn even apparent counter-evidence into supportive evidence. Tam (2000) achieves this by simultaneously reproducing four key parameter estimates of Tam (1997) after adding a modest amount of measurement error to the BIB-motivated model estimated by England et al. (2000). The simulation results, together with all the available evidence, strongly confirm that noisy correlated independent variables have distorted previous findings that are purportedly supportive of the devaluation hypothesis for the wage penalty.

Remedy. The preceding argument does not imply, however, that



adding control variables is necessarily bad or harmful. To the contrary, the addition of controls is no doubt a necessary practice in quantitative research. The desirability of adding control is central to the multivariate revolution in quantitative social research. Future progress in sociology continues to hinge on a proper application of multivariate modeling. The real problem is that *not all* controls are helpful or desirable.

We must be vigilant of the facts that (1) all meaningful control variables included in a model are *by design* correlated with some of the key explanatory variables of interest,¹⁴ and (2) most variables have a certain level of measurement error. Hence bigger models may be messier rather than safer than smaller models. The choice of control variables should be done with care. In addition, bigger models do not necessarily produce more valid estimates. The noisier is a control variable, the greater caution should be exercised in ascertaining that measurement error does not play havoc with a multivariate analysis. Consequently, the inclusion of any control variables should be clearly justified and the risk of negative consequences carefully evaluated.

Example. The recent debate over the wage penalty of occupational

¹⁴ The general purpose of including a control variable Z in a model is always that Z is thought to have direct effect on the dependent variable and have partial correlation with one or more key variables of interest. If Z is conditionally correlated with both the dependent variable and with some of the independent variables, omitting Z from the model will introduce a bias on at least some of the model parameters. This is the classic result on omitted variable bias. If Z is, in the case of OLS regression, uncorrelated with all included variables, omitting Z will not result in any statistical bias for any parameter of interest.

sex composition calls attention to an important scenario (the presence of noisy and correlated control variables in models for panel data) in which bigger models are not better (Tam 1997). The debate illustrates how we should stay vigilant against the ruses of control variables (Tam 1997, 2000).

It is well known that panel data are often essential for addressing many social science questions of causality. Because of the availability of measurements on the same subjects over multiple time points, panel data offer one of the most powerful survey design for investigating questions of complex causation. However, panel data models have to address the problem of unobserved individual differences that may fundamentally bias estimation results if not properly accounted for. The usual statistical method to address this problem is a fixed-effects model that in effect requires a dummy variable for every person in the data. But this solution creates yet another problem. Specifying fixed-effects usually removes a very large portion of the statistical variance in all variables - variance that would otherwise be useful for statistical estimation (Tam 2000).¹⁵ That is why the R-squared of fixed-effects models can often reach 90 percent or more. In non-technical terms, this is tantamount to drastically reducing the effective sample size in the data. Although this loss of statistical information is not directly reflected in the reduction of sample size, the loss is

¹⁵ Since the typical fixed-effects model in the wage penalty debate is equivalent to adding a dummy variable for each person, fitting fixed-effects parameters would remove all between-person variance in both the dependent and independent variables. Only *within-person* variances are available for the statistical estimation of other parameters. More often than not, within-person variance is only a small fraction of total variance available for statistical estimation.

very common and often overlooked. Furthermore, when two independent variables are correlated, the amount of statistical variance available for estimating their effects further declines. In sum, both fixed-effects and correlated controls would bring about a reduction in statistical variance for model estimation. A critical impact of this reduction is to aggravate measurement error bias. This is what happened with the prior literature on the apparent wage penalty of female occupations. Tam (1997) demonstrates how a cautious reading of the prior literature should have produced skepticism over the old evidence, suggested useful new analysis, and prompted the discovery of dramatic new findings that are contrary to the old consensus in the literature. *Even though there is no simple and general guideline for the evaluation of control variables, substantive knowledge and a healthy dose of skepticism toward all hypotheses would certainly be helpful.*

While the preceding discussion focuses attention on the statistical reasons for advocating a cautious approach to multivariate model building, Lieberman's (1985) insightful analysis raises the same concern from a different perspective. Being a superb empirical sociologist, he grounds his methodological critique of modern quantitative research on a strong substantive understanding of many social processes and how conventional statistical analysis creates misleading and counter-intuitive results. A recurrent theme of his substantive examples is the presence of social selection processes that would have to be taken seriously in any statistical model building. The remedy necessary is often much more than staying with clean models. It often would require solving difficult data and modeling problems that have earned James J. Heckman a recent Nobel

Prize in Economics. Thus both statistically and substantively, the BIB assumption is dangerously misleading. The safety of adding control variables is a myth.

Conclusion

Significance test is a cost-saving invention that empowers empirical scientists of all fields to do inference based on random samples. Multivariate modeling is a mighty innovation that greatly extends the types of researchable questions and kinds of information deducible from data, and it empowers social scientists to do statistical control across millions of cases and to sort out the relationships among many concepts. All these have proven to be revolutionary tools in the accumulation of social science knowledge. But every tool has a potential for misuse. These wonderful tools are no exception.

This paper has identified three common ways in which these fundamental tools have been misused in sociology. For each misuse, I have identified a conceptual myth (misconception), specified the source of the misconception, and proposed a remedy. All three misuses are fundamental and pervasive. They have severely undermined the quality of quantitative research by proliferating misguided interpretation of results, premature inferences, ill-conceived presentation of findings, and under-utilization of available data. The misuses are certainly *not* confined to the context of OLS regression. Practically any multivariate analysis of structural effects are vulnerable to the same set of potential problems.

This paper has also illustrated the problems mostly with simulated

data, and with a notable debate in the study of gender inequality. It is high time that the practitioners, teachers, and gatekeepers of the discipline work to help deconstruct these myths at the conceptual level and eliminate them before a quantitative study reaches the publication phase.

If successful, the quality of quantitative research in sociology will be significantly and permanently upgraded. The exposition and the examples of this paper are meant to be a constructive step toward this upgrading effort.

作者簡介

Tony Tam（譚康榮），中央研究院歐美所副研究員。主要研究興趣為社會階層、教育、組織與經濟社會學及計量方法。目前主持一項大型合作計畫，研究教育制度對學生學習之影響。



References

Agresti, Alan and Barbara Finlay

- 1986 *Statistical Methods of the Social Sciences*. San Francisco: Dellen Publishing Co.

England, Paula, Joan M. Hermesen, and David A. Cotter

- 2000 "The Devaluation of Women's Work: A Comment on Tam." *American Journal of Sociology* 105:1741-1751.

Hanushek, Eric A. and John E. Jackson

- 1977 *Statistical Methods for Social Scientists*. San Diego: Academic Press.

Lieberson, Stanley

- 1985 *Making It Count: The Improvement of Social Research and Theory*. Berkeley and Los Angeles: University of California Press.

Tam, Tony

- 1997 "Sex Segregation and Occupational Gender Inequality in the United States: Devaluation or Specialized Training?" *American Journal of Sociology* 102:1652-1692.
- 2000 "Occupational Wage Inequality and Devaluation: A Cautionary Tale of Measurement Error." *American Journal of Sociology* 105: 1752-1760.



中文摘要

統計顯著檢定和多變項模型對量化的社會研究具有革命性的貢獻。然而，即使是偉大的工具也有可能被誤用。本篇研究紀要討論量化研究中對於這兩個基本工具，最為常見的三類誤用：（1）對統計顯著檢定的誤解；（2）錯誤地依賴淨效果做為推論的基礎；以及（3）對於控制變項的過度信任。對於每一種誤用，我指出其觀念上的迷思和誤解的來源，並提出補救之道。我會以大學畢業生的虛擬數據資料和近年來有關美國勞力市場性別不平等的辯論為例，闡明其間的問題以及補救方式。本文所指出的問題與補救方法適用於任何以統計模型為基礎的多變項分析（例如 OLS 迴歸、對數迴歸、tobit、其他受限依變項迴歸分析、以及相聯方程式模型等）。

關鍵詞：統計顯著檢定、多變項模型、統計控制、測量誤差

