Feminist Economics

ISSN: 1354-5701 (Print) 1466-4372 (Online) Journal homepage: https://www.tandfonline.com/loi/rfec20

Economic importance and statistical significance: Guidelines for communicating empirical research

Jane E. Miller & Yana van der Meulen Rodgers

To cite this article: Jane E. Miller & Yana van der Meulen Rodgers (2008) Economic importance and statistical significance: Guidelines for communicating empirical research, Feminist Economics, 14:2, 117-149, DOI: 10.1080/13545700701881096

To link to this article: https://doi.org/10.1080/13545700701881096

Feminist Economics **14(2)**, April 2008, 117–149

ECONOMIC IMPORTANCE AND STATISTICAL SIGNIFICANCE: GUIDELINES FOR COMMUNICATING EMPIRICAL RESEARCH

Jane E. Miller and Yana van der Meulen Rodgers

ABSTRACT

A critical objective for many empirical studies is a thorough evaluation of both substantive importance and statistical significance. Feminist economists have critiqued neoclassical economics studies for an excessive focus on statistical machinery at the expense of substantive issues. Drawing from the ongoing debate about the rhetoric of economic inquiry and significance tests, this paper examines approaches for presenting empirical results effectively to ensure that the analysis is accurate, meaningful, and relevant for the conceptual and empirical context. To that end, it demonstrates several measurement issues that affect the interpretation of economic significance and are commonly overlooked in empirical studies. This paper provides guidelines for clearly communicating two distinct aspects of ''significance'' in empirical research, using prose, tables, and charts based on OLS, logit, and probit regression results. These guidelines are illustrated with samples of ineffective writing annotated to show weaknesses, followed by concrete examples and explanations of improved presentation.

KEYWORDS

Economic significance, regression analysis, statistical significance, writing, feminist economics

JEL Codes: Y1, A29, C10

INTRODUCTION

In recent decades, economists have engaged in an ongoing debate about the rhetoric of economic inquiry and the meaning of inferential tests of statistical significance. Feminist dialogue on these issues has critiqued neoclassical economics studies, arguing that too many authors focus on the statistical machinery at the expense of emphasizing the issues that really matter – the substantive research question at hand (Diana Strassmann and Livia Polanyi 1995; Deirdre McCloskey 1998). This dialogue is an important concern for feminist economists, as it touches on a critical element of scholarship by feminist economists on relationships and issues that are of

Feminist Economics ISSN 1354-5701 print/ISSN 1466-4372 online © 2008 IAFFE
http://www.tandf.co.uk/journals
DOI: 10.1080/13545700701881096

economic importance. That work has improved economists' understanding of previously ignored topics that are of consequence to social and economic well-being, including the valuation of women's unpaid work, intra-household allocation of resources and tasks, and gendered processes in the paid labor market (for example, Julie A. Nelson [1995], Martha MacDonald [1995], and Nancy Folbre [1995]).

This dialogue has particular relevance for regression analysis, which is by far the dominant empirical tool used by economists, as indicated by any casual search of empirical studies. In formal support of this claim, Joyce P. Jacobsen and Andrew E. Newman (1997) find that 88 percent of articles published by economists in the top labor journals between 1981 and 1995 used regression analysis. Given the heavy use of regression analysis in the scholarly and policy arenas, guidelines that specify how to go beyond narrow, technical reporting of regression results to include clear discussion of the substantive meaning of those results in broader social and economic context are potentially an important element of journal editorial policies. As of early 2007, just two journals among the top twenty-five ranked journals in economics had implemented editorial policies requiring manuscripts to report specific indicators of statistical significance.[1] The journals *Econometrica* and *Feminist Economics* both require authors to report standard errors rather than *t*-statistics.[2] Motivated by the debate about the rhetoric of statistical reporting, *Feminist Economics* goes one step further by specifying that authors should address the economic importance of their regression results; it is the only one of the top twenty-five economics journals to explicitly require such a discussion.

Although these concerns about the rhetoric of statistical reporting have not had widespread impact on editorial policy, over the last decade, there has been considerable discussion of these issues in the literature. The exaggerated prominence given to reporting statistical significance and relative lack of attention paid to issues of substantive significance form the key arguments in Stephen T. Ziliak and Deirdre N. McCloskey's (2004a) study, which finds that over 80 percent of articles published in the *American Economic Review* (*AER*) during the 1990s failed to distinguish between statistical and economic significance. The percentage of journal articles that used statistical significance to make claims about economic significance actually increased compared with the previous decade based on a similar tally reported by the authors (Deirdre N. McCloskey and Stephen T. Ziliak 1996).

Ziliak's and McCloskey's (2004a) critique appears in a special issue of the *Journal of Socio-Economics* that focuses on the meaning of statistical significance. Other papers in that issue, such as those by Arnold Zellner (2004) and Stephen T. Ziliak and Deirdre N. McCloskey (2004b), go further to argue that economic significance has little to do with statistical significance, that economists use unsatisfactory testing procedures, and that

they place too much emphasis on statistical significance. As noted by special issue contributors such as Bruce Thompson (2004), these problems have a long history and are also found in other disciplines such as psychology, medicine, public health, sociology, and education.

These critiques have seen plenty of counter-arguments, most recently by Kevin Hoover and Mark Siegler (2008), who argue that economists do not confuse statistical and economic significance and that related criticisms of economists' procedures are inaccurate. They re-evaluate Ziliak and McCloskey's reviews of *AER* articles in the 1980s and 1990s and find that they excluded some articles containing regression analysis, and that they used a ''hodge podge'' of questions that failed to produce a clear, objective basis for identifying when authors conflated economic and statistical significance. While Hoover and Siegler agree that the economic significance of a result does not hinge on the coefficient's statistical significance, they disagree that confusion between the two is pervasive and systematic. The debate continues with a rebuttal of these counter-arguments in Stephen T. Ziliak and Deirdre N. McCloskey (2008).

In this paper, we use *Feminist Economics*' editorial policy on communicating significance and the ongoing debate about the meaning of statistical significance as launching points to develop a set of guidelines for distinguishing between statistical and substantive significance when presenting results of empirical research. In our discussion about assessing substantive significance, we review several often-overlooked measurement and specification issues. These issues include considering types of variables, examining the range and distribution of values, matching numeric contrasts to the context of the specific research question, and avoiding decimal system biases. Addressing these issues helps strengthen the research design and model specification and helps ensure that the presentation of results is accurate, meaningful, and relevant for the conceptual and empirical context. To enhance the effectiveness of the discussion and ground it in international feminist scholarship, we use original examples from regression results for female earnings and employment determinants based on survey data from Taiwan. Examples of pitfalls are modeled after those found in published articles in peer-reviewed journals. Finally, we provide detailed guidelines and examples of how to present coefficients and statistical test results in tables, charts, and prose to yield a comprehensive view of both substantive and statistical significance.

MEASUREMENT ISSUES RELATED TO SUBSTANTIVE SIGNIFICANCE

Substantive or economic significance of an association is assessed by asking, ''So what?'' or ''How much does it matter?'' Researchers in other

disciplines have also written about this problem in terms of ''clinical,'' ''practical,'' or ''meaningful'' variation (Thompson 2004; Jane E. Miller 2005). Typically, the underlying models are intended to identify factors that could be used to influence outcomes such as employment, wages, economic growth, or health. Multivariate regression or other related methods of controlling for potential confounding factors are used to simulate ''quasi-experimental'' conditions for situations in which random assignment is not feasible or ethical, or to adjust for possible differences in confounding factors that remained uncorrected in the process of random assignment under true experimental conditions (Paul D. Allison 1999). Coefficients from multivariate models, therefore, provide estimates of the net effects of each independent variable, taking into account the other variables in the model.

Often neglected in the explication of multivariate regression results is the substantive significance of the association between an independent variable X_1 and the dependent variable Y. Ideally, such a discussion should consider whether that association is causal, follows theoretical expectations in terms of the direction (sign) and size of the association, and is large enough to matter in its real world context. Also of importance is the extent to which the sign or magnitude of the effect changes when other variables are included in the model.

Statistical significance alone is not adequate for assessing the ''importance'' of one variable in affecting another: with a large enough sample size such as that provided in many national data sets (for example, the United States Survey of Income and Program Participation, the German Socio-Economic Panel Study, and India's National Sample Survey), even truly microscopic differences can be statistically significant, yet tiny differences are unlikely to be meaningful in a practical sense. Conversely, in a small sample or with large sampling uncertainty for some other reason, a result that is statistically insignificant might be economically important.

Authors can also be sloppy in their use of the term ''significant,'' using it as an adjective to describe a large relationship in contexts where readers might interpret it to refer to statistical significance. For instance, the estimated effect of some policy intervention X_1 might be small relative to the effects of other potential interventions X_2 or X_3. Or it might be unrealistic to induce a large enough change in X_1 to produce an economically meaningful change in Y. In such cases, the causal nature and statistical significance of the association between X_1 and Y are not sufficient to make the case that X_1 is an ''important'' enough cause to be the basis for explanations or interventions to affect Y. For example, if every elementary school student in Brazil were included in a regression analysis comparing a new math curriculum to an existing one, an improvement of even half a percentage point in average test scores might be statistically

significant because the sample size was so large. An assessment of substantive significance would involve considering whether it is worth incurring the cost of producing and distributing new materials and training all Brazilian elementary school teachers in the new curriculum for such a small gain.

To evaluate the substantive importance of research findings, there are several measurement issues to bear in mind. These issues can be classified into two broad categories: first, recognizing and explaining the difference between coefficients for different types of variables, and second, choosing appropriate numeric contrasts for continuous variables based on knowledge of their distributions and real-world context. We illustrate these points with examples based on data for female employees from Taiwan's *Manpower Utilization Survey* ($N = 7{,}944$), a household survey that provides detailed information on individual workers' earnings, hours worked, educational attainment, tenure, job descriptors, and personal characteristics (Directorate-General of Budget, Accounting, and Statistics, Executive Yuan [DGBAS] 1992; Joseph E. Zveglich, Yana V. Rodgers, and William M. Rodgers 1997).

Considering types of variables

A surprisingly common mistake is to directly compare the effect sizes of categorical and continuous variables, when in fact such comparisons make little conceptual sense (Daniel Powers and Yu Xie 2000; Miller 2005).[3] To illustrate both correct labeling and pertinent types of descriptive statistical information for such variables, Tables 1a and 1b report summary statistics for several categorical and continuous variables used in the analysis of women's earnings.

Table 1a Descriptive statistics for categorical variables, composition (percent) of earnings sample, women aged 15–65 in Taiwan, 1992 ($N = 7{,}944$)

	Percentage of sample
Highest level of school attended	
Primary school or less	24.2
Middle school	16.4
High school	9.4
Vocational school	29.0
Junior college	12.9
College or higher	8.1
All levels of schooling	100.0
Manager or supervisor	5.7
Live in urban area	48.3
Married	51.1

Table 1b Descriptive statistics for continuous variables, earnings sample, women aged 15–65 in Taiwan, 1992 ($N = 7,944$)

	Minimum	*Maximum*	*Mean*	*Median*	*Std. dev.*
Monthly earnings at primary occupation (New Taiwan $ = NT$)	500	132,000	18,837	17,000	8,727
Ln(monthly earnings)	6.21	11.79	9.74	9.68	0.47
Monthly hours worked	13	537	202	208	31
Ln(monthly hrs worked)	2.56	6.29	5.29	5.34	0.19
Years potential post-school experience	0	59	14.8	12	12.4
Years enterprise-specific tenure	0.1	42.0	4.5	2.8	5.1
Proportion women in occupation	0.01	0.95	0.58	0.56	0.22
Number of children < 15 years	0	6	0.64	0	1.03

Notes: The data are from Taiwan's 1992 Manpower Utilization Survey, and we restrict the sample to all civilian women of working age who are non-farm, paid employees.
The variable ''Proportion women in occupation'' is the proportion of workers in an occupation who are women.
Source: DGBAS 1992.

One reason these distinctions are important is that the coefficients on continuous and categorical variables are interpreted differently. To illustrate this difference, we turn to Model I in Table 2, which contains coefficients from an ordinary least squares (OLS) regression of women's monthly earnings in New Taiwan dollars (NT$) as a function of their observed productivity characteristics (education, experience, tenure, and hours worked), job characteristics (managerial status and proportion of workers in the occupation who are women), and personal characteristics (marital status, urban residence, and number of children under 15 years old).

For a continuous independent variable, such as the number of children under age 15, the unstandardized coefficient from an OLS regression is an estimate of the slope of the relationship between the independent and dependent variables. The coefficient estimates the marginal effect of a one-unit increase (an additional child) in that independent variable on the dependent variable (women's earnings), holding constant all other variables in the model. In Model I, the coefficient on the variable for the number of children is therefore interpreted as: ''For each additional child under age 15 years, a woman's monthly earnings decreased by NT$476.''

For categorical independent variables, such as the place of residence (urban versus rural), per-unit changes are not relevant. Consequently, the coefficient on a dummy or binary variable such as ''urban'' compares values of the dependent variable for the category of interest (urban) to the reference category (rural). In Model I (Table 2), the coefficient on

Table 2 Ordinary least squares regressions of monthly earnings, linear, and log specifications, women aged 15–65 in Taiwan, 1992

	Model I			*Model II*	
	Dep. var. = mean monthly earnings (NT$)			*Dep. var. = ln(mean monthly earnings)*	
	Unstdized. coeff. (β)	*Std. error of β*	*Stdized. coeff.*	*Coeff.*	*Std. error*
Intercept	-27,576.8**	2,038.9	NA	5.890**	0.114
Productivity characteristics					
Highest education level attended (Primary school or less)					
Middle school	2,366.7**	277.5	0.100	0.098**	0.015
High school	5,301.2**	332.2	0.177	0.273**	0.019
Vocational school	5,605.5**	298.5	0.291	0.282**	0.017
Junior college	11,069.8**	322.1	0.425	0.545**	0.018
College or higher	17,231.8**	367.9	0.539	0.788**	0.020
Potential years post-school experience					
Experience	414.1**	23.8	0.589	0.022**	0.001
Experience^2/100	-756.0**	48.1	-0.463	-0.045**	0.003
Enterprise-specific tenure (years)					
Tenure	521.8**	36.7	0.304	0.036**	0.002
Tenure^2/100	-401.2**	147.1	-0.055	-0.060**	0.008
Ln(hours worked per month)	6,729.9**	376.2	0.149	0.628**	0.021

(*continued*)

Table 2 (*Continued*)

	Model I			*Model II*	
	Dep. var. = mean monthly earnings (NT$)			*Dep. var. = ln(mean monthly earnings)*	
	Unstdized. coeff. (β)	*Std. error of β*	*Stdized. coeff.*	*Coeff.*	*Std. error*
Job characteristics					
Manager or supervisor	4,273.0**	343.9	0.114	0.183**	0.019
Proportion women in occupation	-770.7*	357.3	-0.020	-0.037	0.020
Personal characteristics					
Live in urban area	1,008.4**	148.2	0.058	0.055**	0.008
Married	24.4	219.7	0.001	-0.009	0.012
Number of children <15 years	-475.5**	96.7	-0.056	-0.036**	0.005
Number of observations (N)		7,944		7,944	
F statistic (df)		476.67 (15)**		412.31 (15)**	
Adjusted R^2		0.47		0.44	

Notes: The data are from Taiwan's 1992 Manpower Utilization Survey, and we restrict the sample to all civilian women of working age who are non-farm, paid employees.
The variable ''Proportion women in occupation'' is the proportion of workers in an occupation who are women.
Reference category in parenthesis. *$p < 0.05$; **$p < 0.01$.
Source: DGBAS 1992.

''urban'' is interpreted as: ''Women in urban areas earn on average NT$1,008 more per month than their rural counterparts.''

Although the coefficient in Model I on ''urban'' ($\beta_{\text{urban}} = 1{,}008$) is larger than the coefficient on number of children ($\beta_{\#\text{kids}<15} = -476$), it does not make sense to compare those coefficients directly. For urban/rural, the contrast is one category versus the other, whereas for number of children, the contrast can vary more than one unit (child) across cases; in the Taiwan sample, the range is zero to six children (Table 1b). A woman with two children is predicted to earn approximately NT$950 less than a woman with no children – an effect of nearly equivalent size to the urban/rural difference.

A researcher aware of these distinctions among variable types can then set up comparisons that make sense for each variable, taking into account the following issues that affect plausibility and relevance of numeric contrasts for continuous independent variables. Below, we describe three steps that can help in the choice of numeric contrasts: (1) examine the distributions of variables, (2) match examples to the substantive context, and (3) avoid falling into decimal system biases.

Examining the distribution of variables

Overlooking the distributions of variables can lead to some poor choices of numeric examples and contrasts. Armed with information on the range, mean, variability, and skewness of his/her variables, a researcher is in a better position to pick reasonable values and characterize them as above or below average, typical, or atypical.

A common pitfall occurs when examples intended as illustrations of typical values are, in fact, not typical. For example, the mean may not be appropriate for representing highly skewed distributions or other situations where few cases have the mean value. If the richest person in the nation happened to be one of ten people randomly chosen for an income survey, mean income for that sample would vastly exceed the population average, so the median or modal value would be a more representative choice. If half the respondents to a public opinion poll strongly agree with a proposed new law and the other half passionately oppose it, characterizing the ''average'' opinion as in the middle would be inappropriate. In such a case, a key point would be the polarized nature of the distribution. On the other hand, the mean would be the appropriate illustration of a typical value in which the observed values took on a normal distribution. Graphs such as simple histograms or Tukey's box-and-whisker plots can be particularly helpful ways to visualize distributions and reveal the presence of unreasonable values or outliers (David C. Hoaglin, Frederick Mosteller, and John W. Tukey 2000; Edward R. Tufte 2001).

To illustrate the substantive importance of a variable, a comparison example must be plausible: the differences between groups or changes across time must be feasible economically, behaviorally, politically, or in whatever arena the topic fits. For example, if voters are unlikely to approve more than a 0.7 percent increase in local property taxes, projecting the effects of a 1.0 percent increase will overestimate potential revenue.

If one is discussing the highest and lowest observed values, it is essential to explain that those values represent upper and lower bounds of a distribution and then include one or more smaller contrasts to illustrate more realistic changes. For instance, in 2007, wages in the US varied from minimum wage ($5.15 an hour) to hundreds of dollars per hour charged by elite attorneys and consultants. Estimating the expected reduction in the poverty rate resulting from a $1 per hour or $2 per hour increase in the minimum wage would be a more reasonable contrast than examining the change associated with the entire observed range of hourly wages. Finally, the application of example values that fall outside the data range also requires careful and transparent choices. This issue is probably most familiar for projecting future values based on historical patterns but also applies to regressions based on a limited age or income range to predict outcomes for other ages or incomes. In such cases, a description of the underlying assumptions and data ought to accompany the calculations. For example, a description of the effects of number of children on women's earnings from the Taiwan sample should clearly state that projections beyond six children would be out of range for those data, based on the distribution shown in Table 1b.

These points about choosing examples constitute an area that is ripe for abuse: advocates can artificially inflate apparent benefits or understate liabilities by using unrealistically large or small differences in their examples, as in the property tax illustration given above. An investigation into the real-world context of the research question and the distributions of one's variables will help identify pertinent, credible numeric contrasts for independent variables.

Matching examples to context

A critical facet of a numeric example or comparison is that it be relevant, meaning that the comparison should match its substantive context and likely application. Before coding variables or selecting numeric values to contrast, a researcher ought to identify conventional standards, cutoffs, or comparison values used in the field. This practice helps avoid model specifications that do not correspond well with associated policy or practice criteria for the topic under study. For instance, evaluations of children's nutritional status often use measures of the number of standard deviations above or below the mean for a standard reference population

(Robert J. Kuczmarski, Cynthia L. Ogden, Laurence M. Grummer-Strawn, Katherine M. Flegal, Shumei S. Guo, Rong Wei, Zuguo Mei, Lester R. Curtin, Alex F. Roche, and Clifford L. Johnson 2000), so using those measures and the same reference population facilitates a comparison of one's findings with those of other studies. As a second example, many countries use their national poverty lines to determine eligibility for social safety net programs. For instance, eligibility for the US's Medicaid and Food Stamps programs is based on multiples of the US's Federal Poverty Level (FPL), such as <133 percent of FPL, 134–185 percent of FPL, and so forth.[4] Using this kind of classification will yield results that can be translated more directly into policy or program recommendations than using purely empirical groupings, such as quartiles or standard deviations of the income distribution.

Avoiding decimal system biases

In a decimal (base ten) oriented society, people tend to think in increments of one or multiples of ten, yet there may be a more relevant or interesting contrast. Before using a 1-, 10-, or 100-unit difference, evaluate whether that difference suits the research question, taking into account theory, previous literature on the subject, the data, the scientific conversation, and common usage. Frequently, the ''choice'' of comparison unit is made by the statistical program used to do the analysis because the default increments are often one- or ten-unit contrasts. Depending on the research question or data, other contrasts may be of greater interest. For example, showing how much more food a Dutch family could buy with one euro (€) more in income per week would be a trivial result given today's weekly income (at the median of approximately €560) and current food prices in the Netherlands. A difference of €30 or €40 would be more informative. Finding out how much a proposed change in social benefits or the minimum wage would add to weekly income and then examining its effects on food purchases would be an even better strategy. However, as always, context matters: in a study of the Netherlands in the early twentieth century or some less-developed countries today, a one-unit contrast in weekly income would be well suited because it has meaningful implications for purchasing power.

The regression coefficient on a continuous independent variable reflects the effect of a one-unit increase in that variable on the dependent variable. For some variables, a one-unit increase in the independent variable is unrealistically large. For example, in the case of the proportion of workers in an occupation who are women (''PWOW'' in the following description of calculations), a one-unit increase constitutes the entire range, which by definition can be no lower than 0.0 and no higher than 1.0. (Recall that the units for *proportions* and *percentages* differ by a factor of 100, so authors

should take care to label and interpret the coefficients on such variables using the correct units).[5] A more reasonable increase in the PWOW would be on the order of 0.22 units (one standard deviation; Table 1b). One can then calculate the earnings penalty from working in an occupation that has a plausibly higher value for the PWOW by multiplying the value of 0.22 by the coefficient on this variable from Model I, Table 2. In this case, an increase of 0.22 in the PWOW is associated with roughly a NT$170 decrease in monthly earnings ($-770.7 * 0.22 = -170$).

Even if a one-unit increase is relevant, other contrasts may be better suited to the research question. For example, in the analysis of female earnings, a five-year increase in work experience might be of greater interest to demonstrate the implications of longer labor-force attachment. In Table 2, the positive coefficient on experience and the negative coefficient on experience-squared means that the marginal effect of experience on earnings decreases as we move farther out the experience scale. For example, going from zero to five years' experience is associated with more than five times the increase associated with moving from twenty to twenty-five years' experience (NT$1,882 versus NT$370).[6] This result implies a steeper experience-earnings profile in the early years that then levels off with additional experience.

Some analyses such as life-table calculations use ten-unit contrasts as the default – a poorly suited choice for research questions that require a more refined contrast. For instance, infant mortality declines precipitously in the hours, days, and weeks after birth. Ten-day age intervals are too wide to capture mortality variation in the first few weeks of life and too narrow in the months thereafter. For that topic, more appropriate groupings are the first day of life, the rest of the first week (six days), the remainder of the first month (twenty-one days), and the rest of the first year (337 days). Although these ranges are of unequal width, they satisfy both empirical and theoretical criteria for choosing suitable increments: First, mortality is relatively constant within each interval, satisfying a key empirical criterion whereby the value of the dependent variable is fairly homogeneous within the specified ranges of the independent variable. Second, those age ranges also correspond to theory about the causes of infant mortality at different ages (Ruth R. Puffer and Carlos V. Serrano 1973; T. J. Mathews, Marian F. McDorman, and Fay Menacker 2002).

While the increments may suggest themselves for some policy-oriented research questions – because, for example, social security legislation has already set some focal points such as eligibility based on multiples of a nation's poverty level – the increments may be less transparent for other types of questions, particularly in descriptive studies in a new research area. In such cases, exploring different empirical and theoretical criteria can help researchers arrive at and explain suitable contrasts for their research question.

PRESENTING STATISTICAL SIGNIFICANCE

An important part of balancing the presentation of substantive and statistical significance involves taking advantage of the complementary strengths of tables, charts, and prose for presenting empirical results. In this section, we describe the uses of different vehicles for conveying numeric information and offer some suggestions for enhancing clarity. Tables are the best way to report precise numeric values such as coefficients, standard errors, and model goodness of fit because they can be used to organize many detailed numeric values. Charts are an excellent vehicle for portraying the shape and size of relationships among variables, such as the net effects of interactions or polynomial relationships. Prose is the preferred tool for asking and answering the substantive questions underlying the statistical analysis.

Contents and format of statistical tables

A complete table of multivariate model results contains coefficients and standard errors for each variable in the model, as well as a quick means of communicating statistical significance such as *p*-values or symbols to denote conventional levels of statistical significance. Standard errors are considered essential because they provide an estimate of the extent of variation or uncertainty around the point estimate, allowing readers to exercise caution in interpreting a coefficient with a large estimated standard error. They also allow readers to calculate test statistics and confidence intervals, examine one-tailed and two-tailed tests, and test the statistical significance of differences between coefficients from different models (David Freedman, Robert Pisani, and Roger Purves 1998).

Although they should not be used as a substitute for standard errors, *p*-values can be included in the tables along with standard errors, as a way of rapidly assessing which coefficients are statistically significant, and if not, how closely they approach the standard cutoff of $p < 0.05$. A *p*-value answers the question: ''Is this variable statistically significantly associated with the outcome?'' without requiring much work on the reader's part because the test statistic has already been compared against the critical value. Alternatives for quickly communicating statistical significance include symbols such as asterisks or daggers or formatting such as boldface or color for $p < 0.05$ and $p < 0.01$. Symbols or formatting are particularly useful for tables presenting the results of several different models, and on slide presentations, because they avert the need for a separate column or row of detailed numbers.

In order for tables and charts to be truly useful for presenting numeric results (whether univariate, bivariate, or multivariate), there are several things authors can do to make them more accessible to readers. It is very

important that tables and charts be labeled clearly so readers can understand the information without reference to the text. Using the title, row labels, column headings, axis labels, legends, and notes, readers should be able to discern the purpose of the table or chart; the context of the data (who, when, and where); coding or units of measurement for every variable in the table; the type of statistics or statistical model; data sources; the definitions of pertinent terms, symbols, and abbreviations; and for multivariate models of a categorical dependent variable, the identity of the category or categories being modeled.

For continuous variables, tables should include units of measurement, the level of aggregation, and the system of measurement for every variable in the table. This seemingly lengthy list of items can usually be expressed in a few words, such as ''monthly earnings (NT$),'' or ''distance (kilometers).'' When possible, the units for the table should be generalized rather than repeated for each row and column. If the same units apply to most numbers in the table, they should be specified in the title (as in ''percent'' in Table 1a). If the units differ across rows or columns of a table, it is important that they be labeled in the pertinent row or column. A table of descriptive statistics for a study of access to micro-loans in India might include the value of the loan (in 100s of rupees), the term of the loan (in years), the distance to the lending institution (in kilometers), and the processing time (in weeks). Labeling the units of measurement is critical even if the concepts seem self-evident: without labels, readers might erroneously presume that earnings were measured annually or weekly rather than monthly or that the value of a loan was reported in dollars instead of rupees.[7] Closely related, upward or downward movements in exchange rates are especially prone to misinterpretation if the exchange rates are not clearly labeled (as, for example, €/US$ or US$/€). Finally, if the scale of a variable is changed (for example, by taking logarithms or dividing by 100), authors should take care that the label reflects that scale so that effect sizes can be interpreted correctly.

Abbreviations or acronyms should be kept to a minimum in headings, in favor of using short, meaningful phrases. Such labels in the text or tables are more appropriate than the eight-character variable names from statistical output, since there is no need for readers to learn the acronyms from the researcher's database. It is helpful for the methods section to define the concepts measured by each variable, so the brief labels become familiar.

Tables presenting multivariate models should report the units or coding of the dependent variable in the title. For example, a model of earnings should indicate whether they are monthly or weekly and measured in NT$ or yen. If different models within one table are specified with different transformations of a dependent variable, it is important that the units for

each model be labeled in the column headings, as in Table 2. For models of dichotomous or multichotomous dependent variables such as employment status, the category being modeled should be specified: being employed or being unemployed, for example. Both the included and omitted (reference) categories of each multichotomous nominal or ordinal independent variable should be identified, so readers can correctly interpret the effects. Each dummy variable should be named after the category it embodies rather than the general concept measured by the variable: ''Married'' not ''Marital status,'' for example. As obvious as this may seem, it is probably the single most common basic error in tables of multivariate model results.

It is essential that each column of effect estimates is labeled in order to convey whether it contains standardized or unstandardized coefficients, log-odds or odds ratios, probit coefficients or marginal probabilities, and so forth. Likewise, columns containing standard errors or other inferential statistical information should identify the type of statistic (χ^2, *F*-statistic, *t*-statistic, or *p*-value). Notes to the table should define symbols or formatting used to denote levels of statistical significance, such as $p < 0.05$ or $p < 0.01$. Tables 2, 4, and 5 provide examples of how to format complete, self-contained tables to present results of OLS, logit, and probit regressions, respectively. See Miller (2005) or Adelheid A. M. Nicol and Penny M. Pexman (1999) for more details on how to format univariate, bivariate, three-way, and multivariate tables.

As authors format their tables, they should take care not to overwhelm readers with an excessive number of digits or decimal places. In most cases, two decimal places are adequate for reporting estimated coefficients, but this could be increased to four decimal places if needed to display at least two ''significant digits.''[8] To avoid coefficients with many digits or a large number of leading zeroes, authors might consider changing the scale of the variables. For example, in the specification shown in Table 2, experience-squared is divided by 100 in order to keep the scale of its coefficient in line with other coefficients in the model. Odds ratios often need only two to three significant digits (for example, 1.68 or 0.22) to convey the size of the difference across groups. An exception is continuous independent variables, for which the odds ratio measures the effect of a one-unit increase in that variable. Again, consider changing scale or classifying values into groups to avert the need to report three or more decimal places. See Miller (2005: Chapter 4) for more on these topics.

For inferential statistical test information, there are additional considerations. The number of digits and decimal places for standard errors should be consistent with the number of digits and decimal places reported for the associated coefficients. Test statistics (for example, χ^2, *F*-statistic, or *t*-statistic) require two decimal places to compare them against critical values; *p*-values also conventionally include two decimal places.

Presenting statistical significance in the text

Authors commonly use tables to report statistical test results for all variables but then limit their text descriptions to only those results that are statistically significant, in most cases using $p < 0.05$ as the criterion. Emphasizing statistically significant findings in the text is especially useful when investigating several different independent variables, such as how education, work experience, tenure on the job, and demographic characteristics affect women's earnings. If only some traits associated with earnings are statistically significant, one will often emphasize those traits rather than giving equal prominence to all factors.

But highlighting a finding that is *not* statistically significant can be valuable if the result defies theory or prior studies or if a coefficient loses statistical significance with the introduction of possible mediating or confounding factors. For example, Charles Ballard and Marianne Johnson (2005) find that gender is among the main determinants of performance in an introductory microeconomics course at a particular US university, with female students answering on average 1.79 percent fewer exam questions correctly than their male counterparts ($p < 0.01$). However, the score disadvantage (coefficient) for female students drops to almost zero (-0.02) and is no longer statistically significant (s.e. $= 0.79$) once students' expectations about their ability to succeed and their secondary-school experience with economics are added to the model. The results indicate that the association between gender and performance in economics operates primarily through differences in expectations and prior experience.

To maintain a clear, linear story line, it is best to avoid writing about calculations needed for tests of differences between coefficients from the same model. Instead, theory can be used to identify in advance which tests are of interest for the research question, and then the researcher can compare coefficients behind the scenes, either by writing the statistical program to conduct the pertinent comparison of coefficients or by calculating the standard error of the difference using computerized output on the variances and co-variances of the estimated coefficients and then conducting the formal comparison (Freedman, Pisani, and Purves 1998). The results of those calculations should then be reported in the text. For example, it might be important to know whether the earnings returns to middle school and high school education differ from each other in Taiwan. In this case, for the test $\beta_{\text{middle school}} = \beta_{\text{high school}}$, the author should write ''predicted earnings for women with a middle school education were NT$2,934 lower than those for women with a high school education ($p < 0.01$).'' This kind of textual assessment allows the reader to see the results without having to wade through complicated explanations of the calculations.

WRITING ABOUT SUBSTANTIVE SIGNIFICANCE

Substantive significance in the results section

To communicate substantive significance of regression results, there are several key principles to keep in mind. First, authors should report and interpret the effect sizes for each independent variable of interest, showing what they mean in the context of the research question and data. For example, they should discuss whether the coefficients are in the expected direction, large or small, and above or below some pertinent threshold. Second, the direction (sign) and magnitude of effects should be conveyed, mentioning units of measurement for all variables involved. Third, it is important that the prominence of the research question be maintained by referring to the specific concepts involved rather than making generic references to ''the dependent variable'' or ''the coefficient.''

Another key point is that because of the potential confusion about the meaning of ''significant,'' it is preferable to restrict its use to the statistical sense when describing inferential statistical results. Many other adjectives such as ''considerable,'' ''appreciable,'' or even ''big'' can fill in ably to describe large effect sizes. Alternatively, if an author wishes to use the term ''significant'' when describing empirical results, it is helpful to accompany it with a modifier, such as ''*economically* significant'' or ''*statistically* significant,'' that clarifies which meaning is intended in that context.

Finally, in sections reporting multivariate model results, authors should use a phrase such as ''controlling for other variables in the model,'' ''*ceteris paribus*,'' or ''holding all else constant'' for the first coefficient they interpret, and then avoid repeating it when interpreting other coefficients from that model. A subheading such as ''*Multivariate results*'' can also be used to differentiate a section of multivariate results from presentation of bivariate or three-way findings, averting the need to state that information for every coefficient.

Table 3 presents illustrative sentences for common types of regression specifications, including those with standardized coefficients, logged and un-logged independent variables, and logged and un-logged dependent variables. Example sentences in Table 3 convey findings from Tables 2 (OLS), 4 (logit), and 5 (probit) regression models. For technical information about such models, see Allison (1999) or Damodar N. Gujarati (2002) for OLS models, and Powers and Xie (2000) or David Garson (2007) for logit and probit models.

To help authors structure sentences that are technically accurate, the ''units'' column of Table 3 includes information about whether each type of coefficient measures absolute or relative change, as well as the pertinent units involved for the independent and dependent variables. See Miller

Table 3 Example sentences for conveying substantive significance of coefficients from different types of regression models

Type of model	*Specification of variable*[a]	*Units*	*Example sentence*
Unstandardized coefficients	Lin-lin: $Y = \beta_0 + \beta_1 X_1$	Absolute change in the original units of the dependent variable for a one-unit (absolute) increase in the independent variable.[b]	*"Women living in urban areas earn on average roughly NT$ 1,010 more than those in rural areas* ($p < 0.01$; *Model I, Table 2)."*
	Lin-log: $Y = \beta_0 + \beta_1 \ln X_1$	Absolute change in the original units of the dependent variable for a 1 percent (relative) increase in the independent variable.	*"A 1 percent increase in monthly hours worked is associated with a NT$ 67 increase in monthly earnings* ($p < 0.01$; *Model I, Table 2)."*[c]
	Log-lin: $\ln Y = \beta_0 + \beta_1 X_1$	Percentage change in the dependent variable for a one-unit (absolute) increase in the independent variable.	*"For each additional child a woman has, her monthly earnings are reduced by 3.6 percent* ($p < 0.01$; *Model II, Table 2)."*[d]
	Log-log: $\ln Y = \beta_0 + \beta_1 \ln X_1$	Percentage change in the dependent variable for a 1 percent (relative) increase in the independent variable. This is the *elasticity*.	*"A 1 percent increase in monthly hours worked is associated with a 0.6 percent increase in monthly earnings (Model II, Table 2)."*
Standardized coefficients	Lin-lin: $Y = \beta_0 + \beta_1 X_1$	Change (in number of standard-deviation units) of the dependent variable for a one standard-deviation increase in the independent variable.	*"A one-standard-deviation increase in proportion female in the occupation is associated with a decrease of 0.020 standard deviations; or 2.0 percent of a standard deviation in monthly earnings* ($p < 0.05$; *Model I, Table 2)."*

(*continued*)

Table 3 (*Continued*)

Type of model	*Measure of effect size*	*Units*	*Example sentence*
Logit	Log-odds (β)	Multiples of log-odds that the dependent variable = 1 for a one unit increase in the independent variable.[b,e]	*Not recommended for conveying substantive significance.*
	Odds ratio[f]	Multiples of odds that the dependent variable = 1 for a one unit increase in the independent variable.[b,e]	*"Urban residents had 1.5 times the odds of employment as their rural counterparts ($p < 0.01$; Table 4)."*
	Percentage difference	Percentage difference in the odds that the dependent variable = 1 for a one unit increase in the independent variable.[b,e]	*"Urban women had 50 percent higher odds of employment than their rural counterparts."*
Probit	Probit coefficient	Change in the cumulative normal probability of the dependent variable, expressed in Z-scores (multiples of the standard deviation) for a one-unit increase in the independent variable.[b,e]	*Not recommended for conveying substantive significance.*
	Marginal probability[g]	Change in the probability that the dependent variable = 1 for a one-unit increase in the independent variable.[b,e]	*"The probability of employment was 0.029 lower for married than unmarried women ($p < 0.01$; Table 5), when all variables are set at their means."*

Notes:
[a]Terminology refers to the specification of the dependent and independent variables. "Lin" refers to an untransformed continuous variable in its original units. "Log" refers to a logarithmic transformation of a continuous variable, as in $\ln(Y)$ for the dependent variable, $\ln(X_i)$ for an independent variable. Hence a "lin-log" specification refers to one in which the dependent variable is untransformed and the independent variable is logged, as in the corresponding equation.
[b]For categorical independent variables, the comparison is between the group specified and the reference (omitted) category of the independent variable.
[c]For lin-log models, β_1 divided by 100 gives the absolute change in the dependent variable for a 1 percent increase in X_1 (Gujarati 2002).
[d]For log-lin models, $100 \times (e^{\beta} - 1)$ gives the percentage change in the dependent variable for a one-unit absolute increase in X_1 (Allison 1999).
[e]Logit and probit models can also include independent variables in logarithmic form. In such cases, the measure of effect size would be interpreted for a 1 percent increase in the independent variable.
[f]The odds ratio is calculated $OR = e^{\beta} = e^{\text{log-odds}}$.
[g]The marginal probability is calculated as $\phi(\beta' x)\beta$ where $\phi(t)$ is the standard normal density. See William H. Greene (2002) for more discussion.

(2005: Chapters 8, 9) for additional discussion of how to calculate and write about these and other types of quantitative comparisons.

Ordinary least squares regression

Ordinary least squares regression is used for continuous dependent variables, such as income in NT$. The following example applies the principles listed above to illustrate how to write sentences about coefficients from the OLS model shown in Table 2.

Poor: ''Residence is correlated with monthly earnings ($p < 0.01$).''
Comment: *This sentence names the concepts involved and conveys statistical significance but not the direction or magnitude of the association.*

Poor (version #2): ''The β for 'urban' is 1,008.4 with a standard error of 148.2 (Table 2).''
Comment: *This version simply reports the same information as the associated table and does not help readers interpret the meaning of the coefficient. It also leaves it to readers to perform the statistical test.*

Poor (version #3): ''Women who live in urban areas earn significantly more than those in rural areas.''
Comment: *In this version, it isn't clear whether "significant" is intended in the statistical sense or is meant to describe a large difference.*

Slightly better but still inadequate: ''Women who live in urban areas earn more than those in rural areas ($p < 0.01$).''
Comment: *This version conveys concepts (dependent and independent variables, including the reference category), direction, and statistical significance, but not the magnitude of the association.*

Best: ''Women living in urban areas earn on average NT$1,008 more per month than those in rural areas ($p < 0.01$).''
Comment: *Concepts, units, direction, magnitude, and statistical significance are all reported in one short, straightforward sentence.*

Standardized regression coefficients adjust for the fact that some variables have a much larger standard deviation than others; hence a one-unit absolute increase means different things for different independent variables. For example, a one-unit increase in the proportion of workers in an occupation who are women is much larger relative to its overall range and scale than a one-unit increase in the number of children under age 15. Standardized coefficients are measured in multiples of standard deviations, providing a consistent metric in which to compare coefficients on different variables and allowing assessment of the relative sizes of the associations of each independent variable with the dependent

variable (Sam Kash Kachigan 1991). They typically are not used for dummy variables, for which a one-standard-deviation increase lacks an intuitive interpretation (John Fox 1997). The standardized coefficient on the proportion of workers in an occupation who are women (PWOW) in Table 2 helps illustrate the point that in some cases, statistically significant effects are not large enough to be of much substantive interest. An increase of one full standard deviation in the PWOW is associated with only 2 percent of a standard-deviation decrease in monthly earnings ($p < 0.05$). Such a small change relative to the observed variation in the dependent variable suggests that changes in the proportion of workers in an occupation who are women are of very little economic importance in determining monthly earnings.

Table 3 includes example sentences for specifications involving one or more logged dependent and independent variables. Economists frequently transform their variables by taking logarithms for one of the following reasons: to correct for skewness in the distribution of a variable, reflect an underlying theoretical relationship that is linear in logs but non-linear in levels of the variables, capture the linear trend over time for a variable that grows at a constant rate, reduce the influence of outliers, or report coefficients that can be interpreted as percentages or elasticities. In the case of labor economics, for example, human capital earnings equations are typically specified with ln(earnings) as the dependent variable in order to correct for skewness in the earnings distribution. This specification also allows the interpretation of estimated coefficients as the percentage change in earnings associated with one-unit increases in independent variables, such as hours worked or years of work experience.

Multivariate models for categorical variables include logit and probit models. In the next two sections, we demonstrate how to write clear sentences to convey substantive significance of logit and probit coefficients.

Logit regression

The estimated coefficient (β_i) from a logistic regression is the change in the natural logarithm of the odds ratio (ln relative odds) of the outcome associated with a one-unit increase in the independent variable (X_i). Hence, logit coefficients communicate direction of association – in this case, which group has higher ($\beta_{\text{urban}} > 0$) or lower ($\beta_{\text{married}} < 0$) chances of being employed. But the logit coefficients capture the size of the association only relative to one another: although the researcher can assess which factors have larger or smaller effects on the dependent variable, the size is not interpretable in an intuitively meaningful way. As a consequence, the effect estimates from a logistic regression are conventionally expressed in terms of odds ratios for each independent variable, which are easily interpretable in multiples or percentage change in the odds of the outcome. Table 4 presents results from a logistic model of women's employment.

Table 4 Logistic regression results for models of employment, women aged 15–65 in Taiwan, 1992

	Log. (relative odds)		*Odds ratio*	
	Coeff. (β)	*Std. error of β*	*OR [exp(β)]*	*Std. error for OR*
Intercept	3.234**	0.118	NA	
Productivity characteristics				
Highest education level attended (Primary school or less)				
Middle school	-0.128	0.079	0.88	0.069
High school	0.078	0.104	1.08	0.112
Vocational school	0.018	0.089	1.02	0.091
Junior college	0.507**	0.112	1.66**	0.187
College or higher	0.392**	0.140	1.48**	0.207
Potential years post high-school experience				
Experience	-0.040**	0.007	0.96**	0.007
Experience^2/100	-0.050**	0.012	0.95**	0.012
Personal characteristics				
Live in urban area	0.387**	0.050	1.47**	0.074
Married	-0.185**	0.071	0.83**	0.059
Number of children <15 years	-0.125**	0.027	0.88**	0.024
School is a major activity	-7.739**	0.205	0.0004**	0.0001
Housework is a major activity	-5.115**	0.083	0.006**	0.0005
Business owner in the household	-2.520**	0.055	0.08**	0.004
Number of cases		24,293		
Likelihood ratio chi-square (df)		18,822.48 (13)**		
Pseudo R^2		0.61		

Notes: The data are from Taiwan's 1992 Manpower Utilization Survey, and we restrict the sample to all civilian women of working age. Some students and homemakers are participating in the labor force, so variables for school or housework status are appropriate identifying variables. Reference category for education in parenthesis.
*$p < 0.05$; **$p < 0.01$.
Source: DGBAS 1992.

To convey the direction and magnitude of associations from logit models in prose, authors should name the concepts under study rather than solely referring to odds ratios.

Poor: ''The odds ratio is 1.47 (s.e. = 0.074).''

Comment: *This version makes no reference to either the independent or dependent variables, so the results are completely divorced from the research question.*

Better but still inadequate: ''The odds ratio of employment is 1.47 ($p < 0.01$)''

Comment: *This version refers to the outcome under study, but does not specify which groups are being compared.*

Best: ''Urban residents had 1.5 times the odds of employment as their rural counterparts ($p < 0.01$).''

Comment: *This version clearly states the outcome (employment), contrast (urban compared to rural), direction, magnitude, and statistical significance of the association. The odds ratio is rounded to one decimal place to avoid cluttering the text with extra digits that do little to enhance an understanding of the size of the relationship.*

A drawback of odds ratios, however, is that as the odds ratio or the prevalence of the outcome among the reference group increase, odds ratios are an increasingly poor approximation of the corresponding relative risk (Jun Zhang and Kai F. Yu 1998; Miller 2005). The corresponding relative risk (RR) can be calculated from an odds ratio (OR) and the prevalence of the outcome in the reference group (p_r).

$$\mathrm{RR} = \mathrm{OR}/[(1 - p_r) + (\mathrm{OR} \times p_r)]$$

For instance, if the probability of employment among rural residents is 0.27 and the estimated odds ratio of employment for urban versus rural residents is 1.47, the corresponding relative risk is 1.30. In other words, the odds ratio overstates the true relative risk of employment for urban compared to rural residents in Taiwan by about 13 percent.

Probit regression

Probit models usually yield similar conclusions to those from logistic regression in terms of statistical significance, sign, and relative magnitude of effects, with probit coefficients generally in the neighborhood of 1.8 times the corresponding logit coefficient (Garson 2007). But probit coefficients are more difficult to interpret than odds ratios calculated from logit coefficients: a probit coefficient estimates the difference a one-unit increase in the independent variable will have on the cumulative normal probability of the dependent variable, expressed in Z-scores (multiples of the standard deviation). These are hardly user-friendly measures, particularly for applied audiences.

A more easily comprehensible measure of effect size can be calculated from probit coefficients by computing marginal probabilities, a step that can be performed by most statistical packages.[9] The marginal probability measures the effect of a one-unit increase in the independent variable on the probability that the dependent variable (for example, employment) equals one. Unlike OLS coefficients, marginal probabilities from probit models depend on the value of the independent variables X_i. They are typically calculated with the X_i set at their sample means, although calculating marginal probabilities for other values of selected independent

variables can be instructive (Powers and Xie 2000; Garson 2007). For example, Carole A. Green (2005) examines labor-force-participation decisions of men and women ages 65 and over in the US. She estimates the marginal probabilities of labor-force participation for hypothetical whites, blacks, and Hispanics with average characteristics for those in the sample under a variety of scenarios.

When reporting the results of a probit regression, it is helpful to state the assumptions about the values of other variables that were used when calculating marginal probabilities up front, so that this information does not need to be repeated with the interpretation of each marginal probability. If marginal probabilities were calculated for different scenarios (for example, several combinations of values of other variables), it is useful to explicitly convey how the assumptions change across scenarios being compared.

The following examples are drawn from Table 5, which presents regression results from a probit model of women's employment using the Taiwan data.

Poor: ''The marginal probability for 'married' was -0.029 ($p < 0.01$; Table 5).''

Comment: *This sentence adds little to the information already reported in the accompanying table. It also fails to mention the dependent variable.*

Better but still inadequate: ''The probability of employment for women differs by marital status, with a 0.029 point difference ($p < 0.01$; Table 5).''

Comment: *This sentence mentions the dependent variable and size of the difference, but does not specify which marital status group is more likely to be employed.*

Best: ''The probability of employment was 0.029 points lower for married than unmarried women ($p < 0.01$), when all variables are set at their means.''

Comment: *This version refers to the specific variables and categories involved and conveys the size, direction, and statistical significance of the association.*

Presenting statistical significance information to match the measure of effect size

An important aside: standard errors differ for standardized versus unstandardized OLS coefficients, for log-odds versus odds ratios from logit models, and for coefficients versus marginal probabilities from probit models. These differences occur because the metrics in which the respective variants are measured differ. Compare the standard errors of

Table 5 Probit regression results for models of employment, women aged 15–65 in Taiwan, 1992

	Probit coeff. (β)		*Marginal probability*	
	Coeff. (β)	*Std. error of β*	*Marg. prob.*	*Std. error of marg. prob.*
Intercept	1.694**	0.061	NA	
Productivity characteristics				
Highest education level attended (Primary school or less)				
Middle school	-0.078	0.043	-0.018	0.010
High school	0.027	0.056	0.006	0.014
Vocational school	0.006	0.047	0.001	0.011
Junior college	0.278**	0.059	0.074**	0.017
College or higher	0.212**	0.073	0.055**	0.021
Potential years post high-school experience				
Experience	-0.017**	0.004	-0.007**	0.001
Experience^2/100	-0.028**	0.007	-0.006**	0.002
Personal characteristics				
Live in urban area	0.211**	0.027	0.051**	0.007
Married	-0.120**	0.038	-0.029**	0.009
Number of children <15 years	-0.054**	0.015	-0.013**	0.004
School is a major activity	-3.960**	0.080	-0.313**	0.006
Housework is a major activity	-2.662**	0.037	-0.508**	0.006
Business owner in the household	-1.301**	0.028	-0.287**	0.007
Number of cases	24,293			
Likelihood ratio chi-square (df)	18,458.60 (13)**			
Pseudo R^2	0.60			

Notes: The data are from Taiwan's 1992 Manpower Utilization Survey, and we restrict the sample to all civilian women of working age. Some students and homemakers are participating in the labor force, so variables for school or housework status are appropriate identifying variables. Reference category in parenthesis.
*$p < 0.05$; **$p < 0.01$.
Source: DGBAS 1992.

the βs (logit coefficients) with the standard errors of the odds ratios, exp(β), in Table 4. For instance, the standard error of the β for ''business owner in the household'' is 0.055, whereas the standard error associated with the odds ratio for the same variable is 0.004. Reporting the standard error of the β for business owner with the associated odds ratio (OR = 0.08) would lead readers to falsely conclude that the OR was not significantly different from 1.0, in other words, that chances of being employed do not differ according to whether there is a business owner in the household. Likewise, the standard errors for the marginal probabilities presented in Table 5 differ from the standard errors for the corresponding probit coefficients. The key point here is that the *z*-statistics for the two variants of the effects' estimates of each variable are identical, leading to the same

conclusion about statistical significance whether one looks at standardized or unstandardized OLS coefficients, odds ratios or log-odds (for logit models), or coefficients or marginal probabilities (for probits).

As a consequence of these differences in metrics between coefficients and their transformed versions, the units of the statistical significance information (such as standard errors) should match the units of the corresponding effect estimates when presenting odds ratios or marginal probabilities as measures of effect size. Most statistical packages can generate the standard errors for odds ratios or marginal probabilities.

Using charts to present regression results

Charts are an effective way of portraying the direction and magnitude of associations through the slopes and shapes of curves or relative heights of bars. For example, Figure 1 shows the net effect of an interaction between gender and marital status from an OLS regression of monthly earnings using the full Taiwan sample of male and female employees. This type of chart averts the need for readers to mentally add together several coefficients from a table of regression results in order to calculate the overall pattern among the four gender/marital status combinations.[10] See Tufte (2001) or Miller (2005) for additional guidelines on charts to present regression results.

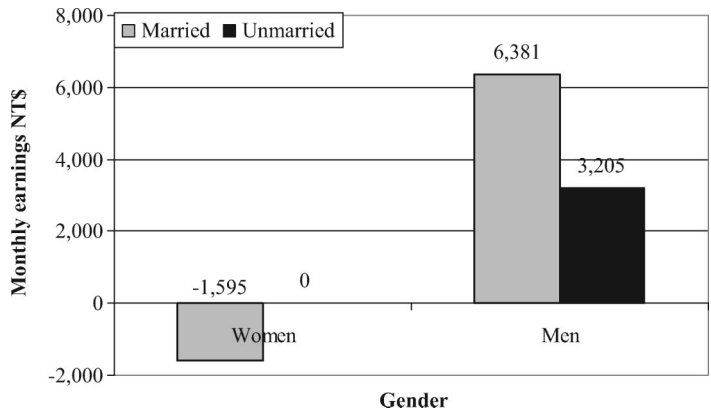

Figure 1 Predicted difference in monthly earnings (NT$) by gender and marital status, Taiwan, 1992

Notes: Based on models estimated from data from Taiwan's 1992 Manpower Utilization Survey, restricted to all civilian men and women of working age who are non-farm, paid employees. Values shown are compared to unmarried females. Model also controls for work experience, tenure, monthly hours, educational attainment, residence, and occupation characteristics.

Source: DGBAS 1992.

To explain the pattern clearly, the figure should then be accompanied by a description of the pattern:

Poor: ''The main effect of 'man' was 3,205 and the main effect of 'married' was −1,595, while the interaction term 'man and married' was 4,771 (all $p < 0.05$).''

Comment: *This description fails to mention dependent variable (earnings) and also does not explain that the main effect and interaction terms must be considered together in order to calculate the net effect of the interaction.*

Poor (version #2): ''Gender and marital status interacted in their effects on earnings.''

Comment: *This version conveys that there is an interaction and names the pertinent independent and dependent variables but does not explain the direction or size of the relationship.*

Better but still inadequate: ''The net effect of being a married man on earnings was NT$6,381 compared to unmarried women ($p < 0.01$; Figure 1).''

Comment: *This version reports the result of the calculation involving the two main effects terms and the interaction term that pertain to married men and correctly specifies the reference category (unmarried women). But, it does not interpret the meaning of that calculation or compare the other gender and marital status combinations.*

Best: ''As shown in Figure 1, men earn more than women regardless of marital status. But, the effect of marriage on earnings works in opposite directions for men than for women: although marriage confers a substantial earnings advantage for men (NT$3,176 extra per month for married compared to unmarried men), it is associated with a sizeable deficit for women (NT$1,595 less per month for married compared to unmarried women).

Comment: *This version captures the overall shape of the earnings pattern among the four gender/marital status categories, including direction, size, and units.*

Substantive significance in the discussion

In the concluding section of the paper, authors should explicitly return to the discussion of substantive significance, situating findings in their real-world context by returning to the original, broader research question. A restatement of economic importance in the discussion also helps ensure that readers who skim only the abstract and conclusions do not overlook this point. To help readers gain a sense of the ''importance'' of a causal factor, authors should use one of the following approaches to quantifying

the economic significance of proposed policies, programs, or other interventions.

One approach is to compare the sizes of the effects of the likely changes (policy or behavioral) associated with one or two important independent variables. For example, in an analysis of housing affordability, contrast effects of a 0.5 percent reduction in property taxes against a subsidy of 2 percent in mortgage interest rates. Closely related, one could compute the effect of a 0.5 percent reduction in property taxes and then compute how large a mortgage subsidy must be to produce the same effect, thus generating a set of precise equivalents that could then be evaluated for plausibility. Another approach is to provide evidence of whether the dependent variable is predicted to move across some economically important threshold in response to a likely change in an independent variable. For example, one could report how a given increase in average wages translates into movement above the poverty threshold to evaluate whether the effort and expense needed to induce that change is worthwhile. A third possibility is to report results of a cost-effectiveness analysis (Marthe R. Gold, Joanna E. Siegel, Louise B. Russell, and Milton C. Weinstein 1996). For instance, one could compare the cost-effectiveness of higher teacher salaries versus more teaching materials in improving education outcomes. Fourth, one could report and evaluate both the prevalence and consequences of a problem. For example, the expected value of insuring against a low-probability but severe disaster such as a 100-year flood might be lower than that of insuring against a higher-probability, less serious level of flooding. The net, combined effect of prevalence of a risk factor and the associated relative risk of an outcome can be quantified using attributable risk (David E. Lilienfeld and Paul D. Stolley 1994). Finally, one could examine the contribution of a variable to explained variance using the model R^2 and an F-test for change in model fit.

To maintain appropriate attention on the main relationships in their analysis (and to avoid an overly long discussion section), authors should limit such discussions of economic significance to the one or two main independent variables of interest for their research question. For example:

Poor: ''The association between being a supervisor and women's earnings was not very substantively significant.''

Comment: *This summary relies on the generic phrasing "substantively significant" without tying it to the research question at hand. It also fails to mention the direction, size, or statistical significance of the association and does not assess whether the change is big enough to matter.*

Better but still inadequate: ''Women who attain positions as managers or supervisors enjoy a monthly earnings premium of about NT$4,300 ($p < 0.01$).''

Comment: *This version is written with content and style appropriate to the results section, but fails to provide useful material for the discussion by putting that finding in a broader economic or social context.*

Best: ''Although the difference in women's earnings associated with being a supervisor or manager is large and highly statistically significant, the change applies to a fairly small share of women. Managers earn roughly 25 percent (NT$4,300) more per month than non-managers. But, the position is open only to those with high education and substantial work experience who can get past the thick glass ceiling in Taiwan, and less than 6 percent of women held such positions in 1992.''

Comment: *This description puts the results back in the context of the original research question of whether working as a manager is a reasonable path for increasing women's earnings. Both substantive and statistical significance are explicitly incorporated, and the interpretation is fleshed out by considering the plausibility of the change in the independent variable (being a manager).*

In analyses intended to inform policy or other interventions, it is also imperative to accurately convey the evidence about causality between the pertinent independent variables and the dependent variable. Verbs such as ''affect'' or ''cause'' and nouns such as ''consequences'' or ''effects'' all imply causality, while ''correlated'' or ''associated'' do not. Similar considerations apply to statements of hypotheses, so these should be phrased to convey whether the relationship is believed to be causal or merely a correlation.

CONCLUSIONS

In this article, we have explained and illustrated principles to help improve the communication of empirical research findings beyond the narrow statistical meaning of ''significant'' to a broader definition of what is economically important, consistent with the high standards set by editorial policies at the journal *Feminist Economics*. A balanced presentation of both substantive and statistical significance is critical for policy-makers and others who are not formally trained in statistics yet often constitute an appreciable part of the intended audience for an economic analysis. Such readers can mistakenly believe that ''importance'' is based only on the size of an association – the bigger the difference across groups, the more important the association. The guidelines suggested here will help authors avoid this type of misunderstanding, showing them how to maintain a focus on the underlying economic issues that matter for scholarly and policy discourse while also reporting the necessary statistical information.

We have demonstrated concrete approaches for presenting and distinguishing between these different facets of ''importance'' when writing about several common types of multivariate regression results. To help researchers identify contrasts that fit the context of their particular research question, we have provided strategies for interpreting coefficients using both empirical information and theoretical underpinnings associated with that question. In addition, we have explained how to use tables, charts, and prose together to create a clear presentation of both substantive and statistical significance of regression results.

A challenge in formulating appropriate guidelines is arriving at options that suit the presentation of different types of economic analyses, ranging from descriptive studies to hypothesis testing to program evaluations and policy proposals. It is our aim that these guidelines help researchers convey quantitative results more clearly, set standards that facilitate accumulation of new knowledge, generate findings that are relevant for policy reform, and address feminist critiques to put more emphasis on the substantive issues behind statistical analyses.

Jane E. Miller, Institute for Health, Health Care Policy, and Aging Research, Rutgers University, 30 College Avenue, New Brunswick, NJ 08901, USA
e-mail: jmiller@ifh.rutgers.edu

Yana van der Meulen Rodgers, Women's and Gender Studies Department, Rutgers University, 162 Ryders Lane, New Brunswick, NJ 08901, USA
e-mail: yrodgers@rci.rutgers.edu

ACKNOWLEDGMENTS

This paper is adapted in part from material in *The Chicago Guide to Writing about Multivariate Analysis* (Miller 2005), Chicago: The University of Chicago Press. We thank Deborah Carr, Mónica Parle, Anne Piehl, William Rodgers, Patricia Roos, Louise Russell, Kristen Springer, Diana Strassmann, Joseph Zveglich, and several anonymous reviewers for their useful comments, and Mike Alvarez for his helpful research assistance.

NOTES

[1] This statement is based on a review of online submission guidelines in February 2007. The ranking is from the 2005 Journal Citation Reports of the Social Sciences Citation Index.

[2] Another journal (*Journal of Financial Economics*) requires that authors report sample size, sample study period, sub-sample definition, and dimensions of numbers. Two journals (*American Economic Review* and *Journal of Law and Economics*) have instructions for which symbols to use to represent statistical significance.

[3] Continuous variables are measured in units such as years (for example, age or date) or currency (for example, income or price), including those that assume only integer values as well as those with decimal values. For continuous variables, the concept of a ''one-unit increase'' (essential to the interpretation of regression coefficients and other types of mathematical computations) is consistent with how those variables are measured. Categorical variables come in two types: Ordinal (''ordered'') variables have categories that can be ranked according to the values of those categories, such as primary, secondary, and higher education. For ordinal variables, neither the width of intervals nor the numeric distance between them can be assumed to be constant, so evaluating a ''one-unit increase'' (or other computations) is not meaningful (Daniel F. Chambliss and Russell K. Schutt 2003). Nominal (''named'') variables such as gender, marital status, or urban/rural residence are classified into categories with no inherent order, so again, mathematical computations involving numeric values of their categories do not make sense.

[4] The US government has established a wide-reaching set of programs to target individuals living in low-income households, including Medicaid (which sends healthcare providers reimbursements for medical services provided to eligible individuals) and the Food Stamp Program (which provides eligible individuals with coupons and electronic cards that they can use to purchase food). The Federal Poverty Level in the US is calculated based on family size and age composition (US Census Bureau 2007).

[5] For example, the proportion 0.22 is equivalent to 22 percent, not 0.22 percent.

[6] In the model shown in Table 2, the relationship between earnings and years of work experience is specified as a quadratic function of experience, so both coefficients must be considered together to estimate the net effect of a given increase in work experience. Applying the coefficients on experience and (experience-squared divided by 100) to an increase from no work experience to five years experience, we obtain $414.1*5 + (-756.0)*(5^2/100) = 1{,}881.5$.

[7] International currency labels are particularly important given that multinational organizations, such as the World Bank and the OECD, often report their cross-country data series in dollars rather than local currencies.

[8] In the phrase ''significant digits,'' the term ''significant'' has a different meaning from either the statistical or substantive interpretations of ''significant'' discussed earlier in the text. Here, it refers to precision of measurement and how that affects the appropriate number of digits in measured values (raw data) and calculations (Ralph H. Logan 1995).

[9] The marginal probability is calculated as $\phi(\beta' x)\beta$ where $\phi(t)$ is the standard normal density.

[10] Estimated coefficients from the OLS model of monthly earnings (using the full Taiwan sample of male and female employees) were: $\beta_{\text{man}} = 3{,}205$; $\beta_{\text{married}} = -1{,}595$; interaction: $\beta_{\text{man and married}} = 4{,}771$. To calculate the net effect of the interaction between gender and marital status requires calculation of comparisons of each of three groups against the reference category: unmarried women. To compare unmarried men against unmarried women, only the main effect β_{male} pertains. To compare married women against unmarried women, only the main effect β_{married} pertains. However, for married men, the net effect involves both main effects and the interaction term $\beta_{\text{man}} + \beta_{\text{married}} + \beta_{\text{man and married}}$, or $3{,}205 + (-1{,}595) + 4{,}771 = 6{,}381$, or NT$ 6,381 more per month than unmarried women. The model also included all other variables shown in Table 2 except number of children, which was not asked of men; the full set of estimated coefficients is available upon request.

REFERENCES

Allison, Paul D. 1999. *Multiple Regression: A Primer.* Thousand Oaks: Sage Publications.

Ballard, Charles and Marianne Johnson. 2005. ''Gender, Expectations, and Grades in Introductory Microeconomics at a US University.'' *Feminist Economics* 11(1): 95–122.

Chambliss, Daniel F. and Russell K. Schutt. 2003. *Making Sense of the Social World: Methods of Investigation.* Thousand Oaks, CA: Sage Publications.

Directorate-General of Budget, Accounting, and Statistics, Executive Yuan (DGBAS). 1992. *Manpower Utilization Survey Database.* Taipei: Directorate-General of Budget, Accounting, and Statistics.

Folbre, Nancy. 1995. ''Holding Hands at Midnight: The Paradox of Caring Labor.'' *Feminist Economics* 1(1): 73–92.

Fox, John. 1997. *Applied Regression Analysis, Linear Models and Related Methods.* Thousand Oaks, CA: Sage Publications.

Freedman, David, Robert Pisani, and Roger Purves. 1998. *Statistics,* 3rd ed. New York: W. W. Norton.

Garson, David. 2007. ''Log-Linear, Logit, and Probit Models.'' http://www2.chass.ncsu.edu/garson/pa765/logit.htm (accessed February 2007).

Gold, Marthe R., Joanna E. Siegel, Louise B. Russell, and Milton C. Weinstein, eds. 1996. *Cost-Effectiveness in Health and Medicine.* New York: Oxford University Press.

Green, Carole A. 2005. ''Race, Ethnicity, and Social Security Retirement Age in the US.'' *Feminist Economics* 11(2): 117–43.

Greene, William H. 2002. *Econometric Analysis,* 5th ed. Upper Saddle River, NJ: Prentice Hall.

Gujarati, Damodar N. 2002. *Basic Econometrics,* 4th ed. New York: McGraw-Hill/Irwin.

Hoaglin, David C., Frederick Mosteller, and John W. Tukey. 2000. *Understanding Robust and Exploratory Data Analysis.* New York: John Wiley and Sons.

Hoover, Kevin and Mark Siegler. 2008. ''Sound and Fury: McCloskey and Significance Testing in Economics.'' *Journal of Economic Methodology* 15(1): 1–37.

Jacobsen, Joyce P. and Andrew E. Newman. 1997. ''What Data Do Economists Use? The Case of Labor Economics and Industrial Relations.'' *Feminist Economics* 3(2): 127–30.

Kachigan, Sam Kash. 1991. *Multivariate Statistical Analysis: A Conceptual Introduction,* 2nd ed. New York: Radius Press.

Kuczmarski, Robert J., Cynthia L. Ogden, Laurence M. Grummer-Strawn, Katherine M. Flegal, Shumei S. Guo, Rong Wei, Zuguo Mei, Lester R. Curtin, Alex F. Roche, and Clifford L. Johnson. 2000. ''CDC Growth Charts: United States.'' *Advance Data from Vital and Health Statistics,* p. 314. Hyattsville, MD: National Center for Health Statistics.

Lilienfeld, David E. and Paul D. Stolley. 1994. *Foundations of Epidemiology,* 3rd ed. New York: Oxford University Press.

Logan, Ralph H. 1995. ''Significant Digits.'' http://members.aol.com/profchm/sig_fig.html (accessed December 2007).

MacDonald, Martha. 1995. ''Feminist Economics: From Theory to Research.'' *Canadian Journal of Economics* 28(1): 159–76.

Mathews, T. J., Marian F. MacDorman, and Fay Menacker. 2002. ''Infant Mortality Statistics from the 1999 Period Linked Birth/Infant Death Data Set.'' *National Vital Statistics Reports* 50(4). Hyattsville, MD: National Center for Health Statistics.

McCloskey, Deirdre N. 1998. *The Rhetoric of Economics,* 2nd ed. Madison: University of Wisconsin Press.

McCloskey, Deirdre N. and Stephen T. Ziliak. 1996. ''The Standard Error of Regressions.'' *Journal of Economic Literature* 34(1): 97–114.

Miller, Jane E. 2005. *The Chicago Guide to Writing about Multivariate Analysis.* Chicago: University of Chicago Press.

Nelson, Julie A. 1995. ''Feminism and Economics.'' *Journal of Economic Perspectives* 9(2): 131–48.

Nicol, Adelheid A. M. and Penny M. Pexman. 1999. *Presenting Your Findings: A Practical Guide for Creating Tables*. Washington, DC: American Psychological Association.

Powers, Daniel and Yu Xie. 2000. *Statistical Methods for Categorical Data Analysis*. San Diego, CA: Academic Press.

Puffer, Ruth R. and Carlos V. Serrano. 1973. *Patterns of Mortality in Childhood*. Washington, DC: World Health Organization.

Strassmann, Diana and Livia Polanyi. 1995. ''The Economist as Storyteller: What the Texts Reveal,'' in Edith Kuiper and Jolande Sap with Susan Feiner, Notburga Ott, and Zafiris Tzannatos, eds. *Out of the Margin: Feminist Perspectives on Economics*, pp. 29–50. London and New York: Routledge.

Thompson, Bruce. 2004. ''The 'Significance' Crisis in Psychology and Education.'' *Journal of Socio-Economics* 33(5): 607–13.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information,* 2nd ed. Cheshire, CT: Graphics Press.

US Census Bureau. 2007. ''How the Census Bureau Measures Poverty.'' http://www.census.gov/hhes/www/poverty/povdef.html (accessed November 2007).

Zellner, Arnold. 2004. ''To Test or Not to Test and If So, How? Comments on 'Size Matters.''' *Journal of Socio-Economics* 33(5): 581–6.

Zhang, Jun and Kai F. Yu. 1998. ''What's the Relative Risk? A Method of Correcting the Odds Ratio in Cohort Studies of Common Outcomes.'' *Journal of the American Medical Association* 280(19): 1690–91.

Ziliak, Stephen T. and Dierdre N. McCloskey. 2004a. ''Size Matters: The Standard Error of Regressions in the American Economic Review.'' *Journal of Socio-Economics* 33(5): 527–46.

———. 2004b. ''Significance Redux.'' *Journal of Socio-Economics* 33(5): 665–75.

———. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.

Zveglich, Joseph E., Yana V. Rodgers, and William M. Rodgers. 1997. ''The Persistence of Gender Earnings Inequality in Taiwan, 1978–1992.'' *Industrial and Labor Relations Review* 50(4): 594–609.