Econometric Models and Samples: Equivalence

A random sample is a sample in which the events or observations are randomly drawn from the population, universe or sample space, as explained in Appendix A. The events in a randomly drawn sample may or may not appear in random order. In contrast, an unsorted sample is a sample in which the events appear in random order. A sorted sample is a sample in which the events appear in nonrandom order. A random variable represented by the probability distribution of a sequence of randomly generated numbers will have no statistical significance or explanatory power, as shown in Appendix B.

For example, if a randomly drawn sample of common stocks is sorted by market capitalization (from low to high, or from high to low) and then partitioned into portfolios (e.g., quintiles, deciles, or other fractiles), this data sorting *per se* does not introduce any new information into any other variable in the sample.

But when a group-based variable is constructed from a sorted and partitioned sample by any variable other than a randomly generated number assigned to each observation, the variable used for sorting and grouping the observations introduces new information into the newly constructed variable. For example, when a sample of common stocks is sorted and grouped into portfolios using market capitalization as the portfolio formation variable and these portfolios are used to create a portfolio-based variable, then market capitalization is introduced into the portfolio-based variable.

Testing and Estimating Econometric Models

One of the estimation and testing methods for econometric models is the Classical Linear Regression Model estimated by Ordinary Least Squares techniques pursuant to the least-squares principle and the Gauss-Markov Theorem, the workhorse of econometrics.

The statistical significance of each explanatory variable that is directly specified in a given econometric model equation is measured by the probability level or by the Student's *t*-statistic, with a critical value equal to conventional levels of probability. Typically, the critical value is probability = 5%, two-tail test. The statistical significance of the combination of explanatory variables that are directly specified in a given model equation is measured by the *F*-statistic in the *F* test.

The explanatory power of a univariate, monocausal, econometric model (one explanatory variable) is measured by the coefficient of determination, R^2, which ranges from a minimum of zero to a maximum of 100. For a multivariate econometric model (two or more explanatory variables), the explanatory power is measured by the adjusted coefficient of determination, to allow for interaction between the explanatory variables. In the case of a univariate econometric model, the coefficient of determination is equal to the square of the Pearson's product-moment correlation coefficient. The correlation coefficient ranges from a minimum of negative one for perfect inverse correlation to a maximum of positive one for perfect direct correlation, and zero is no correlation.

Empirical Benchmarks for Econometric Models

In the doctoral dissertation entitled *Capital Market Efficiency of Firms Financing Research and Development* by Robert D. Coleman, 1996, [Coleman (1996)] empirical lower and upper benchmarks are used for comparison with the explanatory variables specified in the econometric models. The lower benchmark variable, *LEXI*, is designed to approximate a randomly-generated number. *LEXI* is the lexical order of the company name of each common stock in the sample used for econometric model testing and estimation of parameters. The upper benchmark variable, *MKEQ*, is designed to

approximate a tautology. *MKEQ* is the market equity of each common stock in the sample, where market equity is equal to share price multiplied by the number of shares outstanding for each company.

LEXI was used instead of a randomly-generated number because *LEXI* would not approach zero explanatory power as closely as would a randomly-generated number. In addition, *LEXI* is more highly visible and memorable, thereby emphasizing that positive steps were taken in the research design to draw attention to this important characteristic of sample construction.

The theoretical lower and upper limits of the explanatory power of an econometric model are zero percent and 100%, as indicated by the coefficient of determination, R^2, for simple linear regression or the adjusted coefficient of determination or multiple coefficient of determination, adjusted R^2, for multiple linear regression. The range of the theoretical R^2 is zero percent to 100%, and the range of the empirical R^2 is a low positive percent greater than zero percent to a high positive percent less than 100%. The lower and upper benchmarks are provided by separate econometric models, in contrast to separate explanatory variables specified in the same econometric model with the non-benchmark explanatory variables.

The main appearances of the *LEXI* variable in the subject doctoral dissertation are summarized in Table 1. The market proxies for this study were provided by the CRSP database of monthly prices, capitalization changes, and dividends, as shown in Table 2:

As reported in Coleman (1996, Table 27, Part A and Part B, page 193), the Pearson's product-moment correlation coefficient between *BSPRD* and *LEXI* is equal to 0.06 or 6%.

BSPRD is the estimated CAPM market-beta factor based on the S&P 500 with reinvested dividends as the market proxy.

In a univariate econometric model to explain *BSPRD*, with *LEXI* specified as the sole explanatory variable, the coefficient of determination, R^2, is equal to (0.06 x 0.06) = 0.0036 or 0.36%. This R^2 is calculated by commercial statistical analysis software packages using proprietary computational algorithms. Any grouping of a randomly-drawn sample before model testing and estimation of parameters with related statistics may result in a positive value for the R^2 statistic. There is a logical reason for the positive explanatory power in a simple linear regression of *BSPRD* on *LEXI*, even if it is counter-intuitive. The non-zero explanatory power can be considered a statistical artifact.

Econometric Equivalence

Econometric equivalence here means that an explanatory variable in an econometric model can be tested, and the qualitative result of the statistical test will be the same (either accept or reject the null hypothesis, at the critical level of probability with either one-tail or two-tail test) regardless of whether the variable is directly specified in the model equation or indirectly included in a group-based variable that is specified in the model equation, as shown in Appendix C and Appendix D.

The formulas for the calculation of the parameters and statistics for an econometric model equation are presented in introductory econometrics textbooks. Simple univariate regression models ($Y = a + bX + e$, where Y is the dependent variable to be explained, "a" is the intercept, "b" is the regression coefficient, X is the explanatory variable, and e is the stochastic disturbance term or error term) can be estimated with small samples (e.g., five data points) using pencil, paper and no calculator. Multivariate regression models

can be estimated with large samples (thousands of data points) using computational algorithms and large-memory, high-speed, electronic computers. Competing proprietary algorithms achieve either greater computational speed with no loss in precision or greater precision with no loss in speed. Commercial professional statistical software packages have very high accuracy of calculations, and the degree of accuracy can be considered the same for most practical applications.

Table 1. LEXI Benchmark Explanatory Variable

Page	Section or Table	Item
98	Table 6. Measures and Formulas	Benchmarks
101	Table 8. Descriptive Statistics	BSPRD and LEXI
115	The Explanatory Benchmarks	Lower Benchmark: LEXI Upper Benchmark: MKEQ
123	Table 11. Glossary and Definitions: Basic Variables	Market Proxies: VWRXG, EWRXG, SPRXG, and SPRDG
193	Table 27. Correlation Analysis	LEXI
228	Table 37. Univariate CAPM	Benchmarks: Case 12: MKEQ Benchmarks: Case 13: LEXI
229	Table 38. Univariate CAPM	Market Proxies: VWRXG, EWRXG, SPRXG, and SPRDG
264	Table 52. Dynamic Return-Generative Process	Case 25: Model 1 Case 25: Model 2
265	Table 53. Dynamic Return-Generative Process	Case 25: Model 1 Case 25: Model 2
266	Table 53. Dynamic Return-Generative Process	Legend
267	Table 54. Dynamic Return-Generative Process	Case 25: Model 1 Case 25: Model 2
268	Table 54. Dynamic Return-Generative Process	Legend
269	Table 55. Dynamic Return-Generative Process	Case 25: Model 1 Case 25: Model 2
271	Table 56. Dynamic Return-Generative Process	Case 5: Model 1 Case 5: Model 2
Source: Coleman, Robert D., *Capital Market Efficiency of Firms Financing Research and Development*, May 1996. Ph.D. dissertation. Dallas, TX: The University of Texas at Dallas.		

Table 2. Stock Market Proxies

Proxy	Definition
VWRXG	Value-weighted NYSE/AMEX/NASDAQ full nominal total gross return without dividends reinvested
EWRXG	Equal- weighted NYSE/AMEX/NASDAQ full nominal total gross return without dividends reinvested
SPRXG	S&P 500 Index full nominal total return without reinvested dividends
SPRDG	S&P 500 Index full nominal total return with reinvested dividends
Source: Coleman, Robert D., *Capital Market Efficiency of Firms Financing Research and Development*, May 1996. Ph.D. dissertation. Dallas, TX: The University of Texas at Dallas, Table 11, page 123.	

APPENDIX A

GROUPED SAMPLES

Gujarati, Damodar N., *Basic Econometrics*, 2/e, 1988, New York: McGraw-Hill.

Appendix A. A Review of Some Statistical Concepts
A.2. Sample Space, Sample Points, and Events
Pages 624-625
The set of all possible outcomes of a random, or chance, experiment, is called the *population*, or *sample space*, and each member of this sample space is called a sample point. … An *event* is a subset of the sample space. … Events are said to be *mutually exclusive* if the occurrence of one event precludes the occurrence of another event. … Events are said to be (collectively) *exhaustive* if they exhaust all the possible outcomes of an experiment.

A.3. Probability and Random Variables
Pages 625-626
See the text.

Bailey, Kenneth D., *Methods of Social Research*, 1987, 3/e, New York: The Free Press.

Chapter 4. Measurement

Level of Measurement
Page 61
S. S. Stevens (1951) constructed a widely adopted classification of levels of measurement in which he speaks of nominal measurement, ordinal measurement, interval measurement, and ratio measurement.

Chapter 5. Survey Sampling
Probability Sampling
Page 87
Sampling methods can be classified into to those that yield *probability* samples and those that yield *nonprobability* samples. In the former type of sample the probability of selection of each respondent is known. In the latter type, the probability of selection is not known.

Random Sampling
Pages 87-88
Probably the best-known form of probability sample is the random sample. In a random sample each person in the universe has an equal person of being chosen for the sample, and every collection of persons, of the same size, has an equal probability of becoming the actual sample. This is true regardless of the similarities or differences among them, as long as they are members of the same universe.

All that is required to conduct a random sample, after an adequate sampling frame is constructed, is to select persons without showing bias for any personal characteristics. Notice that the adequacy of the random sample depends on the adequacy of the sampling frame.

Another factor is that sampling for surveys is usually sampling without replacement. … Sampling without replacement is called *simple random sampling*.
Simple random sampling is usually considered adequate if the chances of selection are equal at any given stage in the sampling process.

The usual procedure in random sampling is to assign a number to each person or sampling unit in the sampling frame, so that one cannot be biased by labels, names, or other identifying criteria.

Random sampling has the advantage of canceling out biases and providing a statistical means for estimating sampling errors.

Chapter 16. Data Reduction, Analysis, Interpretation, and Application
Table Presentations
Page 371
Statistical analysis is generally presented either in equation form or in a table or graph of some sort.

Univariate Presentation
Page 371-372
In a descriptive study, especially an exploratory one, the researcher may be more concerned with describing the extent of occurrence of a phenomenon than with studying its correlates. In such a case a univariate presentation is in order. … One useful and easy presentation is the *range* of scores, which is defined as the highest score minus the lowest score.

In addition to the range, the researcher can present averages or measures of central tendency such as the mean, median, and mode. … In addition to the mean it is helpful to compute a measure of dispersion such as the variance.

Other succinct measures that can be given without presenting all scores are the frequency distribution and grouped data. The frequency distribution is a listing of the frequency with which each score occurs. .. For an interval variable with many possible scores, such as income, even presentation of a frequency distribution may not be feasible. In this case the researcher may wish to group the data into categories and present the frequency of scores within each category. Such a grouped frequency distribution is obviously a compromise. It provides frequencies of each group of scores from low to high, but provides no information on ranges or variations in scores within each group. … One has to compromise by providing few enough groups so that the data is manageable without making each group too broad.

Hypothesis Testing
Page 381
Statistics that are used to infer the truth or falsity of a hypothesis are called inferential statistics, in contrast to descriptive statistics, which do not seek to make an inference but merely provide a description of the sample data.

The general inference to be tested is that some phenomenon that is true for a sample is also true for the population from which the sample was drawn.

Another distinction often made is between *parametric* and *nonparametric* statistics. Nonparametric statistics are those used when the variables being analyzed are either nominal or ordinal, and interval measurement may not be assumed. Thus nonparametric statistics are also called *order* statistics. The name "nonparametric" stems from the fact that these statistics are not based on assumptions about the parameters of the distribution (the normal or bell-shaped distribution is not assumed, for example). However, this does not mean that no assumptions are necessary for using nonparametric statistics. … Parametric statistics are used when interval measurement can be assumed.

Blalock, Hubert M., Jr., *Social Statistics*, revised second edition, 1979, New York: McGraw-Hill.

Table on inside of front cover (with one data cell populated):

Measurement level of first variable	Single variable procedures	Two-Variable (bivariate) procedures			
		Measurement level of second variable			
		Dichotomy	Nominal (*c* categories)	Ordinal	Interval and ratio
Dichotomy					
Nominal (*r* categories)					
Ordinal					
Interval and ratio					Correlation and regression Chaps. 17, 18

Chapter 4. Interval Scales: Frequency Distributions and Graphic Presentation
Page 41
In the following two chapters we shall be concerned with methods of summarizing data in a more compact manner so that they may be described by several numbers representing measures of typicality and degree of homogeneity.

4.1. Frequency Distributions: Grouping the Data
Page 41
If interval-scale data are to be summarized in a similar manner, however, an initial decision must be made as to the categories that will be used. Since the data will ordinarily

be distributed in a continuous fashion, with few or no large gaps between adjacent scores, the classification scheme may be somewhat arbitrary. If will be necessary to decide how many categories to use and where to establish the cutting points. Unfortunately, there are no simple rules for accomplishing this since the decision will depend on the purposes served by the classification.

Chapter 9. Probability
9.1. A Priori Probabilities
Page 116
Let us call any outcome or set of outcomes of an experiment an *event*, with the set of all possible outcomes under the null hypothesis being referred to as the *sample space*. An event can be simple (nondecomposable) or compound (a combination of simple events). … It is conventional to use the term *success* whenever the event under consideration occurs and *failure* when it does not occur.[2] ([2] This technical use of the terms *success* and *failure* need not conform to general usage.)

9.5. Independence and Random Sampling
Pages 139-140
All the statistical tests to be discussed in this text make use of the assumption that there is independence between events and that therefore conditional probabilities do not have to be used when multiplying probabilities. In other words, it is assumed that there is independence of selection within a sample—the choice of one individual having no bearing on the choice of another individual to be included in the sample. There are many instances in which this important assumption is likely to be violated, however. One should therefore develop the habit of always asking himself whether or not the independence assumption is actually justified in any given problem.

Statisticians often obtain what is called a *random sample* (or *simple random sample*) in order to meet the required assumption of independence as well as to give every individual in the population an equal chance of appearing in the sample. … A random sample has the property *not only of giving each individual an equal chance of being selected but also of giving each combination of individuals an equal chance of selection*.

Strictly speaking, since we practically always sample without replacement, the assumption of independence is not quite met.

Although the problems introduced by failure to replace are not serious ones, the failure to give every *combination* of individuals an equal chance of appearing in the sample may result in a serious violation of the independence assumption.

APPENDIX B

Source: Coleman, Robert D., 2005, "Circling the Square". In: "Asset Pricing Circularity", research paper, page 10.

Table. Estimated OLS Regression of Logically Circular Equations for Unit Rectangles. Explained Variable (DV), Explanatory Variables (IV), estimated coefficients (B), Student's *t*, and R-square coefficient of multiple determination. Sample Size = 64.

DV	IV	B	t	IV	B	t	R-SQ
P	L	2.00	8				50%
P	L	2.00	2e15	W	2.00	2e15	100%
A	L	5.50	7				46%
A	L	5.50	19	W	5.50	19	92%
C	L	0.13	7				45%
C	L	0.13	16	W	0.13	16	90%
C	A	0.02	44				97%
C	P	0.06	23				90%
C	P	0.002	0.46	A	0.02	12	97%
C	1/P	-19.24	11				67%
C	1/P	-2.29	3	A	0.02	26	97%
C	O	0.08	1.66				4%
C	1/O	-0.30	1.55				4%
C/O*	1/O	1.16	10				58%
C	S	-0.26	5				27%
C	S	-0.25	5	O	0.06	1.39	29%
C/O*	S/O	-0.23	5	1/O	1.62	15	68%
C	D	0.17	15				78%
C	D	0.17	14	O	0.01	0.51	78%
C/O*	D/O	0.17	14	1/O	-0.14	1.46	90%
C	R	0.46	56				98%
C	R	0.46	55	O	-0.0001	0.11	98%
C/O*	R/O	0.46	55	1/O	-0.12	5	99%
C	X	-0.004	1.85				5%
C/X*	1/X	0.00	0.00				1%

* Intercept constrained to equal zero.
Note: t (60 df, 2-tailed test): 2.00 = 5% probability and 2.66 = 1% probability.

LEGEND

A = area = L*W
C = compactness = A/P
D = diagonal = sqrt(L**2+W**2)
L = length (2, …, 9)
O = oddness (1, 2, 3 or 4)
P = perimeter = (L+W)*2
R = radius of circle = sqrt(A/π)
S = sides ratio = long side/short side
W = width (2, …, 9)
X = random integer (1, …, 100)

APPENDIX C

STATISTICAL REGRESSION PROCEDURES

Source: SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2, Cary, NC: SAS Institute, Inc.

Chapter 36. The REG Procedure (pages 1351-1456)
Page 1352:
ABSTRACT
The REG procedure fits linear regression models by least-squares. Subsets of independent variables that "best" predict the dependent or response variable can be determined by various model-selection methods.

INTRODUCTION
PROC REG is one of the many regression procedures in the SAS System. REG is a general-purpose procedure for regression, while other SAS regression procedures have more specialized applications. … SAS/ETS procedures are specialized for applications in time-series or simultaneous systems. These other SAS/STAT and SAS/ETS regression procedures are summarized in Chapter 1, "Introduction to Regression Procedures," which also contains an overview of regression techniques and defines many of the statistics computed by REG and other regression procedures.

Page 1353:
PROC REG performs the following regression techniques with flexibility:

- handles multiple MODEL statements
- provides nine model-selection methods
- allows interactive changes both in the model and the data used to fit the model
- allows linear inequality restrictions on parameters
- tests linear hypotheses and multivariate hypotheses
- generates scatter plots of data and various statistics
- "paints" or highlights scatter plots
- produces partial regression leverage plots
- computes collinearity diagnostics
- prints predicted values, residuals, studentized residuals, confidence limits, and influence statistics and can output these items to a SAS data set
- can use correlations or crossproducts for input
- write the crossproducts matrix to an output SAS data set
- performs weighted least-squares regression

Nine model-selection methods are available in PROC REG. The simplest method is also the default, where REG fits the complete model you specify.

Page 1354:
Least-Squares Estimation

REG uses the principle of least squares to produce estimates that are the best linear unbiased estimates (BLUE) under classical statistical assumptions (Gauss 1809; Markov 1900).

Page 1357:
Although there are numerous statements and options available in REG, many analyses use only a few of them.

PROC REG <options>;	required statement
<label>: **MODEL** dependents=<regressors> < / options>;	required statement for model fitting; can be used interactively
BY variables; **FREQ** variables; **ID** variable; **VAR** variables; **WEIGHT** variable;	[each of these five statements, when used,] must appear before the first RUN statement
RUN;	the RUN statement

In the above [selective] list, angle brackets denote optional specifications, and vertical bars denote a choice of one of the specifications separated by the vertical bars. In all cases, *label* is optional.

The PROC REG statement is required. To fit a model to the data, the MODEL statement is required. The BY, FREQ, ID, VAR, and WEIGHT statements are optionally specified once for the entire PROC REG step and must appear before the first RUN statement.

Page 1358:
The statements used with the REG procedure in addition to the PROC REG statement are the following [selective] (in alphabetic order):

BY specifies variables to define subgroups for the analysis.

MODEL specifies the dependent and independent variables in the regression model, requests a model selection method, prints predicted values, and provides details on the estimates (according to which options are selected).

Pages 1360-1361:
BY Statement

BY variables

A BY statement can be used with PROC REG to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Use the SORT procedure with a similar BY statement to sort the data.
- Use the BY statement options NOTSORTED or DESCENDING in the BY statement for the REG procedure.
- Use the DATASETS procedure (in base SAS software) to create an index on the BY variables.

When a BY statement is used with PROC REG, interactive processing is not possible; that is, once the first RUN statement is encountered, processing proceeds for each BY group in the data set, and no further statements are accepted by the procedure. A BY statement that appears after the first RUN statement is ignored.

COMMENTS

In private correspondence with the author, a member of the SAS technical service staff added the following clarification.

Explanatory variables in your model cannot go on the BY statement in any SAS procedure. The sole purpose of the BY statement is to provide you with the separate estimation of the model for each level of the BY group. For example, if you were to run the following:

```
PROC REG;
    MODEL   gnp = year   manufact   service   pop;
    BY   country;
RUN;
```

then you would get a separate model for each country.

That is not the same as running the following:

```
PROC REG;
    MODEL   gnp = country   year   manufact   service   pop;
RUN;
```

which would fit one model to the entire dataset.

When you insert the BY statement in PROC REG, that changes how the standard errors will be calculated for the terms in the MODEL statement, and that could change the statistical conclusion drawn. Instead of calculating the standard errors from all the data in the dataset, the BY statement forces the calculation of standard errors within each BY-group separately. That difference in the calculation of standard errors could (and should) force different conclusions, if the data warrant it.

In the example, assume the dataset includes:

360 observations for the US,
240 observations for Canada,
120 observations for Mexico, and
720 observations for all countries in total.

In SAS, when a BY statement is inserted between a PROC REG statement and the first subsequent RUN statement, and the BY grouping-variable is "country", then the equation in the MODEL statement will be fitted for each value of the "country" variable in the dataset, taking each country-value in turn, but using only the subset of the 720-observation dataset that applies to each country-value in order to make the calculations for the associated "country".

With the BY statement inserted, therefore, the SAS calculations of the Standard Error (SE) and calculations of the test statistic (e.g., Student's *t*-statistic) will use the value of "n" for number of independent observations as follows:

country-value = US n = 360.
country-value = Canada n = 240.
country-value = Mexico n = 120.

In contrast, with no BY statement inserted, the SAS calculations of SE and Student's *t*-statistic will use the value of "n" for number of independent observations as follows:

country-values = US, Canada, Mexico n = 720.

Why does this make any difference? Generally, a higher value of "n" results in a lower value of SE, and SE appears in the denominator of the Student's *t*-statistic. In addition, "n" appears in the numerator of the Student's *t*-statistic. Therefore, a higher value of "n" increases the Student's *t*-statistic in two ways: by decreasing the denominator, and by increasing the numerator.

In effect, the BY statement in SAS is a productivity technique and not a statistical technique. The alternative to using the BY statement is to partition the sample into sub-samples. In the above example, the 720-observation dataset could be divided into three subsets (360, 240, and 120 observations), one subset for each country. Then the regression could be run three separate times, one time on each of the three country datasets, but this is less efficient than running the regression only once.

APPENDIX D

Source: Coleman, Robert D., 2006, "Single-Equation Simultaneity Paradox", research paper, pages 23-26 and 49.

REGRESSION EXAMPLE: EQUIVALENCE

An explanatory variable can be introduced into an econometric regression model in at least four ways. We are concerned with group-based variables. To simplify our discussion, we use only two explanatory variables. We use *DIV* as the group-formation variable because it has the widest range of the three explanatory variables. We sort the sample in ascending order of *DIV* with the smallest observation ranked number one and then divide the sample into three *DIV* fractile groups that are as closely equal in size as possible without being equal: Low (n = 1 to 39), Middle (n = 40 to 79), and High (n = 80 to 121). Then we run five regression models.

The first regression model is run overall with group interactions:

$$PDV_i = a_i + b_{1i}(RIV_i) + b_{2i}(GROUP_i) + b_{3i}(RG_i) + u_i,\ i = 1, \ldots, 121 \qquad (1)$$

where *i* indexes individual observations, *GROUP* = 1, 2 or 3, formed on *DIV*, and *RG* = (*RIV*)(*GROUP*).

The second regression model is run overall with dummy variables:

$$PDV_i = a_i + b_{1i}(RIV_i) + b_{2i}(DUMG2_i) + b_{3i}(DUMG3_i) + b_{4i}(RD2_i) + b_{5i}(RD3_i) + u_i,\ i = 1, \ldots, 121 \qquad (2)$$

where *RD2* = (*RIV*)(*DUMG2*), *RD3* = (*RIV*)(*DUMG3*), *DUMG2* [1 = Yes or 0 = No] is the dummy variable for *GROUP* = 2, and *DUMG3* [1 = Yes or 0 = No] is the dummy variable for *GROUP* = 3. The *DUMG1*, *DUMG2* and *DUMG3* dummy variables

represent Low, Middle and High *DIV*, and *DUMG1* is omitted to avoid the dummy-variable trap.

The third regression model is run by group formed on *DIV*:

$PDV_{ig} = a_g + b_g(RIV_{ig}) + u_{ig},\ i = 1, \ldots, 121;\ g = 1, 2, 3$	(3)

where *g* indexes *GROUP* formed on *DIV*, resulting in three separate regressions, one on each *DIV* group.

The fourth regression model directly and explicitly specifies *DIV* as a separate explanatory variable:

$PDV_i = a_i + b_{1i}(RIV_i) + b_{2i}(DIV_i) + u_i,\ i = 1, \ldots, 121$	(4)

The fifth regression model is the same as the third regression model except that is it has no grouping of observations:

$PDV_i = a + b(RIV_i) + u_i,\ i = 1, \ldots, 121$	(5)

The regression models in Eq. (1), Eq. (2), Eq. (4) and Eq (5) contain the same information, and they will return the same results for the same sample. The regression model in Eq. (3) contains additional information about the number of groups, which is not meaningful information but rather is somewhat arbitrarily chosen. More importantly, Eq. (3) specifies groups or sub-samples, and this reduction in sample size alone will change the calculation of standard errors and the test statistics. Properly interpreted, the full-sample regressions [**Supplement: Tables A4-A8**] return equivalent statistical tests and *R*-square's, as summarized in **Table C1** and **Table C2**. The base case for comparison is the regression equation that specifies only *RIV* in Eq. (5).

As it turns out, Eq. (3) results in the same statistical conclusions as do the other model equations. It may appear superficially that Models 3A, 3B and 3C do not include *DIV*

because it is not specified as an explanatory variable, neither directly by itself nor indirectly by entailment in another variable. Nevertheless, the influence of *DIV* is transmitted indirectly through its use in grouping the observations to form sub-samples.

It also may appear that Models 1 and 2 do not include *DIV* because it is not explicitly specified in the model. But here again, close scrutiny of the specified explanatory variables will reveal its presence. When an interval-scale variable is used to form groups, it is reduced to a category-scale variable, resulting in a reduction of the effective sample size and thus less efficient estimation.

A group-based factor can introduce an explanatory variable as shown in the estimated models below. Bold-face emphasis indicates a parameter that is statistically significant at the 5% level of probability with non-directional (two-tail) test.

$PDV = -1.59 + \mathbf{6.80}RIV + \mathbf{4.88}GROUP - \mathbf{0.66}RG$ where *GROUP* is formed on *DIV*; *RG* = (*RIV*)(*GROUP*); N = 121.	(6)
$PDV = 1.24 + \mathbf{7.01}RIV + \mathbf{11.40}DUMG2 + \mathbf{10.60}DUMG3 - \mathbf{2.78}RD2 - \mathbf{1.95}RD3$ where *DUMG2* is Group 2; *DUMG3* is Group 3; *RD2* = (*RIV*)(*DUMG2*); *RD3* = (*RIV*)(*DUMG3*); N = 121.	(7)
$PDV = 1.24 + \mathbf{7.01}RIV$ for Group 1: Low *DIV*; N=39 [$i = 1, \ldots, 39$].	(8)
$PDV = \mathbf{12.64} + \mathbf{4.23}RIV$ for Group 2: Middle *DIV*; N=40 [$i = 40, \ldots, 79$].	(9)
$PDV = \mathbf{11.84} + \mathbf{5.05}RIV$ for Group 3: High *DIV*; N=42 [$i = 80, \ldots, 121$].	(10)
$PDV = \mathbf{-0.05} + \mathbf{3.16}RIV + \mathbf{1.57}DIV$	(11)

for all groups: Low, Middle and High *DIV*; N = 121.	
$PDV = \mathbf{3.41} + \mathbf{6.61}RIV$ for all Groups: Low, Middle and High *DIV*; N = 121.	(12)

Properly interpreted, Eq. (6), Eq. (7) and Eq. (11) are equivalent because they all contain the same information about *PDV*, *RIV* and *DIV*, even though *DIV* is not directly specified in any of the models except Eq. (11). Eq. (12) with no *DIV* influence is the base case for comparison. In addition, Eq. (6), Eq. (7) and Eq. (11) are equivalent to Eq. (13).

$PDV_g = \mathrm{a} + \mathrm{b}(RIV_g) + e_g$ where g indexes groups formed on *DIV*, $g = 1, \ldots, G$; N = 121.	(13)

The number of groups is chosen in advance to partition a given sample into useful case-specific sub-sample sizes.

TABLE D1. MODEL AND SAMPLE: TESTS

Eq. No(s).	Linear Regression Model and Sample Description	Explanatory Variables		Adjusted R-squared
		Significant*	Total	
1, 6	*DIV* Group with Interaction Term	3	3	0.98
2, 7	*DIV* Group with Dummy Variables	5	5	0.99
3, 8	No Group; Low *DIV* Sub-Sample	1	1	0.93
4, 9	No Group; Middle *DIV* Sub-Sample	1	1	0.86
5, 10	No Group; High *DIV* Sub-Sample	1	1	0.96
11	No Group; *DIV* Specified	2	2	1.00
12, 13	No Group; *DIV* Not Specified	1	1	0.96
* Student's *t*-statistic at 5% level of probability with non-directional (two-tail) test.				

Note: The model and full sample combinations have the same qualitative result of either reject or fail to reject the null hypothesis that there is no relation between the each explanatory variable and the dependent variable. In addition, these model and full sample combinations have nearly the same explanatory power, ranging from 96% to 100%, except for the middle-fractile group at 86%. The model and sub-sample combinations effectively fit a separate model to each sub-sample.

TABLE D2. MODEL AND SAMPLE: ESTIMATES

Eq. No.	***PDV* =**					
	Intercept	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**
6	–1.59	+**6.80** *RIV*	+**4.88** *GROUP*	**–0.66** *RG*		
7	+1.24	+**7.01** *RIV*	+**11.40** *DUMG2*	+**10.60** *DUMG3*	**–2.78** *RD2*	**–1.95** *RD3*
8	+1.24	+**7.01** *RIV*				
9	+**12.64**	+**4.23** *RIV*				
10	+**11.84**	+**5.05** *RIV*				
11	**–0.05**	+**3.16** *RIV*	+**1.57** *DIV*			
12	+**3.41**	+**6.61** *RIV*				
Bold parameter values are significant at 5%, two-tail test.						