MODULE ONE, PART ONE: DATA-GENERATING PROCESSES

William E. Becker
Professor of Economics, Indiana University, Bloomington, Indiana, USA
Adjunct Professor of Commerce, University of South Australia, Adelaide, Australia
Research Fellow, Institute for the Study of Labor (IZA), Bonn, Germany
Editor, *Journal of Economic Education*
Editor, *Social Scinece Reseach Network: Economic Research Network Educator*

This is Part One of Module One. It highlights the nature of data and the data-generating process, which is one of the key ideas of modern day econometrics. The difference between cross-section and time-series data is presented and followed by a discussion of continuous and discrete dependent variable data-generating processes. Least-squares and maximum-likelihood estimation is introduced along with analysis of variance testing. This module assumes that the user has some familiarity with estimation and testing previous statistics and introductory econometrics courses. Its purpose is to bring that knowledge up-to-date. These contemporary estimation and testing procedures are demonstrated in Parts Two, Three and Four, where data are respectively entered into LIMDEP, STATA and SAS for estimation of continuous and discrete dependent variable models.

CROSS-SECTION AND TIME-SERIES DATA

In the natural sciences, researchers speak of collecting data but within the social sciences it is advantageous to think of the manner in which data are generated either across individuals or over time. Typically, economic education studies have employed cross-section data. The term cross-section data refer to statistics for each in a broad set of entities in a given time period, for example 100 Test of Economic Literacy (TEL) test scores matched to time usage for final semester 12^{th} graders in a given year. Time-series data, in contrast, are values for a given category in a series of sequential time periods, i.e., the total number of U.S. students who completed a unit in high school economics in each year from 1980 through 2008. Cross-section data sets typically consist of observations of different individuals all collected at a point in time. Time-series data sets have been primarily restricted to institutional data collected over particular intervals of time.

More recently empirical work within education has emphasized panel data, which are a combination of cross-section and time-series data. In panel analysis, the same group of individuals (a cohort) is followed over time. In a cross-section analysis, things that vary among individuals, such as sex, race and ability, must either be averaged out by randomization or taken into account via controls. But sex, race, ability and other personal attributes tend to be constant from one time period to another and thus do not distort a panel study even though the assignment of individuals among treatment/control groups is not random. Only one of these four modules will be explicitly devoted to panel data.

CONTINUOUS DEPENDENT (TEST SCORE) VARIABLES

Test scores, such as those obtained from the TEL or Test of Understanding of College Economics (TUCE), are typically assumed to be the outcome of a continuous variable Y that may be generated by a process involving a deterministic component (e.g., the mean of Y, μ_y, which might itself be a function of some explanatory variables $X_1, X_2 \ldots X_k$) and the purely random perturbation or error term components v and ε:

$$Y_{it} = \mu_y + v_{it} \quad \text{or} \quad Y_{it} = \beta_1 + \beta_2 X_{it2} + \beta_3 X_{it3} + \beta_4 X_{it4} + \varepsilon_{it},$$

where Y_{it} is the test score of the i^{th} person at time t and the it subscripts similarly indicate observations for the i^{th} person on the X explanatory variables at time t. Additionally, normality of the continuous dependent variable is ensured by assuming the error term components are normally distributed with means of zero and constant variances: $v_{it} \sim N(0, \sigma_v^2)$ and $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$.

As a continuous random variable, which gets its normal distribution from epsilon, at least theoretically any value is possible. But as a test score, Y is only supported for values greater than zero and less than the maximum test score, which for the TUCE is 30. In addition, multiple-choice test scores like the TUCE can only assume whole number values between 0 and 30, which poses problems that are addressed in these four modules.

The change score model (also known as the value-added model, gain score model or achievement model) is just a variation on the above basic model:

$$Y_{it} - Y_{it-1} = \lambda_1 + \lambda_2 X_{it2} + \lambda_3 X_{it3} + \lambda_4 X_{it4} + u_{it},$$

where $Y_{it\text{-}1}$ is the test score of the i^{th} person at time $t-1$. If one of the X variables is a bivariate dummy variable included to capture the effect of a treatment over a control, then this model is called a difference in difference model:

[(mean treatment effect at time t) – (mean control effect at time t)] –
[(mean treatment effect at time t−1) − (mean control effect at time t−1)]
$= [E(Y_{it} | treatment = 1) - E(Y_{it} | treatment = 0)] - [E(Y_{it\text{-}1} | treatment = 1) - E(Y_{it\text{-}1} | treatment = 0)]$
$= [E(Y_{it} | treatment = 1) - E(Y_{it\text{-}1} | treatment = 1)] - [E(Y_{it} | treatment = 0) - E(Y_{it\text{-}1} | treatment = 0)]$
= the lambda on the bivariate treatment variable.

Y_{it} is now referred to as the post-treatment score or posttest and $Y_{it\text{-}1}$ is the pre-treatment score or pretest. Again, the dependent variable $Y_{it} - Y_{it\text{-}1}$ can be viewed as a continuous random variable, but for multiple-choice tests, this difference is restricted to whole number values and is bounded by the absolute value of the test score's minimum and maximum.

This difference in difference model is often used with cross-section data that ignores time-series implications associated with the dependent variable (and thus the error term) involving two periods. For such models, ordinary least-squares estimation as performed in

EXCEL and all other computer programs is sufficient. However, time sequencing of testing can cause problems. For example, as will be demonstrated in Module Three on sample selection, it is not a trivial problem to work with observations for which there is a pretest (given at the start of the term) but no posttest scores because the students dropped out of the class before the final exam was given. Single equation least-squares estimators will be biased and inconsistent if the explanatory variables and error term are related because of time-series problems.

Following the lead of Hanushek (1986, 1156-57), the change-score model has been thought of as a special case of an allegedly superior regression involving a lagged dependent variable, where the coefficient of adjustment (λ_0^*) is set equal to one for the change-score model:

$$Y_{it} = \lambda_0^* Y_{it-1} + \lambda_1^* + \lambda_2^* X_{it2} + \lambda_3^* X_{it3} + \lambda_4^* X_{it4} + \omega_{it} .$$

Allison (1990) rightfully called this interpretation into question, arguing that these are two separate models (change score approach and regressor variable approach) involving different assumptions about the data generating process. If it is believed that there is a direct causal relationship $Y_{it-1} \Rightarrow Y_{it}$ or if the other explanatory X variables are related to the $Y_{it\text{-}1}$ to Y_{it} transition, then the regressor variable approach is justified. But, as demonstrated to economic educators as far back as Becker (1983), the regressor variable model has a built-in bias associated with the regression to the mean phenomenon. Allison concluded, "The important point is that there should be no automatic preference for either model and that the only proper basis for a choice is a careful consideration of each empirical application . . . . In ambiguous cases, there may be no recourse but to do the analysis both ways and to trust only those conclusions that are consistent across methods." (p. 110)

As pointed out by Allison (1990) and Becker, Greene and Rosen (1990), at roughly the same time, and earlier by Becker and Salemi (1977) and later by Becker (2004), models to avoid are those that place a change score on the left-hand side and a pretest on the right. Yet, educational researchers continue to employ this inherently faulty design. For example, Hake (1998) constructed a "gap closing variable (g)" as the dependent variable and regressed it on the pretest:

$$g = gap\ closing = \frac{posttest\ score - pretest\ score}{maximum\ \ score - pretest\ score} = f(pretest\ score ...)$$

where the pretest and posttest scores where classroom averages on a standardized physics test, and maximum score was the highest score possible. Apparently, Hake was unaware of the literature on the gap-closing model. The outcome measure g is algebraically related to the starting position of the student as reflected in the pretest: g falls as the *pretest score* rises, for $maximum\ score \geq posttest\ score \geq pretest\ score$.[i] Any attempt to regress a posttest-minus-pretest change score, or its standardized gap-closing measure g on a pretest score yields a biased estimate of the pretest effect.[ii]

As an alternative to the change-score models [of the type posttest − pretest= f(treatement, . . .) or posttest = f(pretest, treatment, …)], labor economics have turned to a

difference-in-difference model employing a panel data specification to assess treatment effects. But not all of these are consistent with the change score models discussed here. For example, Bandiera, Larcinese and Rasul (2010) wanted to assess the effect in the second period of providing students with information on grades in the first period. In the first period, numerical grade scores were assigned to each student for course work, but only those in the treatment were told their scores, and in the second period numerical grade score were given on essays. That is, the treatment dummy variable reflected whether or not the student obtained grade information (feedback) on at least 75 percent of his or her course work in the first period, and zero if not. This treatment dummy then entered in the second period as an explanatory variable for the essay grade.

More specifically, Bandiera, Larcinese and Rasul estimated the following panel data model for the i^{th} student, enrolled on a degree program offered by department d, in time period t,

$$g_{idct} = \alpha_i + \beta [F_c \times T_t] + \gamma T_t + \delta X_c + \sum_{d'} \mu_{d'} TD_{id'} + \varepsilon_{idct}$$

where g_{idct} is the i^{th} student's grade in department d for course (or essay) c at time t and α_i is a fixed effect that captures time-invariant characteristics of the student that affect his or her grade across time periods, such as his or her underlying motivation, ability, and labor market options upon graduation. Because each student can only be enrolled in one department or degree program, α_i also captures all department and program characteristics that affect grades in both periods, such as the quality of teaching and the grading standards. F_c is a equal to one if the student obtains feedback on his or her grade on course c and T_t identifies the first or second time period, X_c includes a series of course characteristics that are relevant for both examined courses and essays, and all other controls are as previously defined. $TD_{id'}$ is equal to one if student i took any examined courses offered by department d' and is zero otherwise; it accounts for differences in grades due to students taking courses in departments other than their own department d. Finally, ε_{idct} is a disturbance term.

As specified, this model does not control for past grades (or expected grades), which is the essence of a change-score model. It should have been specified as either

$$g_{idct} = \alpha_i + \omega g_{idct-1} + \beta [F_c \times T_t] + \gamma T_t + \delta X_c + \sum_{d'} \mu_{d'} TD_{id'} + \varepsilon_{idct}$$

or

$$g_{idct} - g_{idct-1} = \alpha_i + \omega g_{idct-1} + \beta [F_c \times T_t] + \gamma T_t + \delta X_c + \sum_{d'} \mu_{d'} TD_{id'} + \varepsilon_{idct}$$

Obviously, there is no past grade for the first period and that is in part why a panel data set up has historically not been used when only "pre" and "post" measures of performance are available. Notice that the treatment dummy variable coefficient β is inconsistently estimated with bias if the relevant past course grades in the second period essay-grade equation are omitted. As discuss in Module Three on panel data studies, bringing in a lagged dependent variable into panel data analysis poses more estimation problems. The thing emphasized here is that a change-score model must be employed in assessing a treatment effect. In Module Four,

propensity score matching models are introduced for a means of doing this as an alternative to the least squares method employed in this module.

DISCRETE DEPENDENT VARIABLES

In many problems, the dependent variable cannot be treated as continuous. For example, whether one takes another economics course is a bivariate variable that can be represented by $Y =$ 1, if yes or 0, if not, which is a discrete choice involving one of two options. As another example, consider count data of the type generated by the question how many more courses in economics will a student take? 0, 1, 2 … where increasing positive values are increasingly unlikely. Grades provide another example of a discrete dependent variable where order matters but there are no unique number line values that can be assigned. The grade of A is better than B but not necessarily by the same magnitude that B is better than C. Typically A is assigned a 4, B a 3 and C a 2 but these are totally arbitrary and do not reflect true number line values. The dependent variable might also have no apparent order, as the choice of a class to take in a semester – for example, in the decision to enroll in economics 101, sociology 101, psychology 101 or whatever, one course of study cannot be given a number greater or less than another with the magnitude having meaning on a number line.

In this module we will address the simplest of the discrete dependent variable models; namely, those involving the bivariate dependent variable in the linear probability, probit and logit models.

Linear Probability Model

Consider the binary choice model where $Y_i = 1$, with probability P_i, or $Y_i = 0$, with probability $(1–P_i.)$. In the linear probability regression model $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $E(\varepsilon_i) = 0$ implies $E(Y_i \mid x_i) = \beta_1 + \beta_2 x_i$, where also $E(Y_i \mid x_i) = (0)[1-(P_i \mid x_i)] + (1)(P_i \mid x_i) = P_i \mid x_i$. Thus, $E(Y_i \mid x_i) = \beta_1 + \beta_2 x_i = P_i \mid x_i$, which we will write simply as P_i. That is, the expected value of the 0 or 1 bivariate dependent variable, conditional on the explanatory variable(s), is the probability of a success ($Y = 1$). We can interpret a computer-generated, least-squares prediction of $E(Y|x)$ as the probability that $Y = 1$ at that x value.

In addition, the mean of the population error in the linear probability model is zero:

$$\begin{aligned} E(\varepsilon) &= (1-\beta_1-\beta_2 x)P + (0-\beta_1-\beta_2 x)(1-P) \\ &= P-\beta_1-\beta_2 x = P - E(Y \mid x) = 0 \text{ for } P = E(Y \mid x) \end{aligned}$$

However, the least squares $\hat{Y}$ can be negative or greater than one, which makes it a peculiar predictor of probability. Furthermore, the variance of epsilon is

$$\text{var}(\varepsilon) = P_i[1-(\beta_1+\beta_2 x_i)]^2 + (1–P_i)(\beta_1+\beta_2 x_i)^2 = P_i(1-P_i)^2 + (1–P_i)P_i^2 = P_i(1–P_i),$$

which (because P_i depends on x_i) means that the linear probability model has a problem of heteroscedasticity.

An adjustment for heteroscedasticity in the linear probability model can be made via a generalized least-squares procedure but the problem of constraining $\beta_1 + \beta_2 x_i$ to the zero – one interval cannot be easily overcome. Furthermore, although predictions are continuous, epsilon cannot be assumed to be normally distributed as long as the dependent variable is bivariate, which makes suspect the use of the computer-generated *t* statistic. It is for these reasons that linear probability models are no longer widely used in educational research.

Probit Model

Ideally, the estimates of the probability of success ($Y = 1$) will be consistent with probability theory with values in the 0 to 1 interval. One way to do this is to specify a probit model, which is then estimated by computer programs such as LIMDEP, SAS and STATA that use maximum likelihood routines. Unlike least squares, which selects the sample regression coefficient to minimize the squared residuals, maximum likelihood selects the coefficients in the assumed data-generating model to maximize the probability of getting the observed sample data.

The probit model starts by building a bridge or mapping between the 0s and 1s to be observed for the bivariate dependent variable and an unobservable or hidden (latent) variable that is assumed to be the driving force for the 0s and 1s:

$$I_i^* = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i = X_i \beta \text{ , where } \varepsilon_{it} \sim N(0,1) .$$

and $I^* > 0$ implies $Y = 1$ and $I^* \leq 0$ implies $Y = 0$ and
$P_i = P(Y = 1 \mid X_i) = G(I_i^* > 0) = G(Z_i \leq X_i \beta)$.

$G(\ )$ and $g(\ )$ are the standard normal distribution and density functions, and
$P(Y = 1) = \int_{-\infty}^{X\beta} g(t)dt$.

Within economics the latent variable I^* is interpreted as net utility or propensity to take action. For instance, I^* might be interpreted as the net utility of taking another economics course. If the net utility of taking another economics course is positive, then I^* is positive, implying another course is taken and $Y = 1$. If the net utility of taking another economics course is negative, then the other course is not taken, I^* is negative and $Y = 0$.

The idea behind maximum likelihood estimation of a probit model is to maximize the density L with respect to β and σ where the likelihood function is

$$L = f(\varepsilon) = (2\pi\sigma^2)^{-n/2} \exp(-\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} / 2\sigma^2)$$
$$= (2\pi\sigma^2)^{-n/2} \exp[-(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) / 2\sigma^2]$$

The calculation of $\partial L / \partial \beta$ is not convenient but the logarithm (ln) of the likelihood function is easily differentiated

$$\partial \ln L / \partial \beta = L^{-1} \partial L / \partial \beta .$$

Intuitively, the strategy of maximum likelihood (ML) estimation is to maximize (the log of) this joint density for the observed data with respect to the unknown parameters in the beta vector, where σ is set equal to one. The probit maximum likelihood computation is a little more difficult than for the standard classical regression model because it is necessary to compute the integrals of the standard normal distribution. But computer programs can do the ML routines with ease in most cases if the sample sizes are sufficiently large. See William Greene, *Econometric Analysis* (5th Edition, 2003, pp. 670-671) for joint density and likelihood function that leads to the likelihood equations for $\partial \ln L / \partial \beta$.

The unit of measurement and thus the magnitude of the probit coefficients are set by the assumption that the variance of the error term ε is unity. That is, the estimated probit coefficients along a number line have no meaning. If the explanatory variables are continuous, however, the probit coefficients can be employed to calculate a marginal probability of success at specific values of the explanatory variables:

$\partial p(x) / \partial x = g(X\beta)\beta_x$, where $g(\ )$ is density $g(z) = \partial G(z) / \partial z$.

Interpreting coefficients for discrete explanatory variables is more cumbersome as demonstrated graphically in Becker and Waldman (1989) and Becker and Kennedy (1992).

Logit Model

An alternative to the probit model is the logit model, which has nearly identical properties to the probit, but has a different interpretation of the latent variable *I**. To see this, again let

$$P_i = E(Y = 1 \mid X_i) .$$

The logit model is then obtained as an exponential function

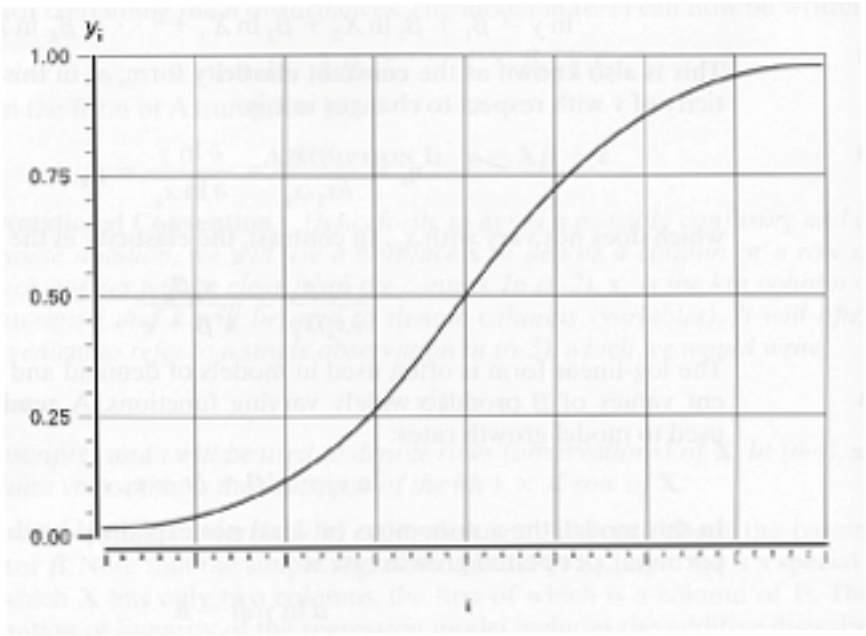
$$P_i = 1/(1 + e^{-X_i\beta}) = 1/(1 + e^{-z_i}) = e^{z_i} /(1 + e^{z_i}) \text{ ; thus,}$$
$$1 - P_i = 1 - e^{z_i} /(1 + e^{z_i}) = 1/(1 + e^{z_i}) \text{, and}$$
$$P_i /(1 - P_i) = e^{z_i} \text{, which is the odd ratio for success } (Y = 1)$$

The log odds ratio is the latent variable logit equation

$$I_i^* = \ln\left(\frac{P_i}{1-P_i}\right) = z_i = X_i\beta\,.$$

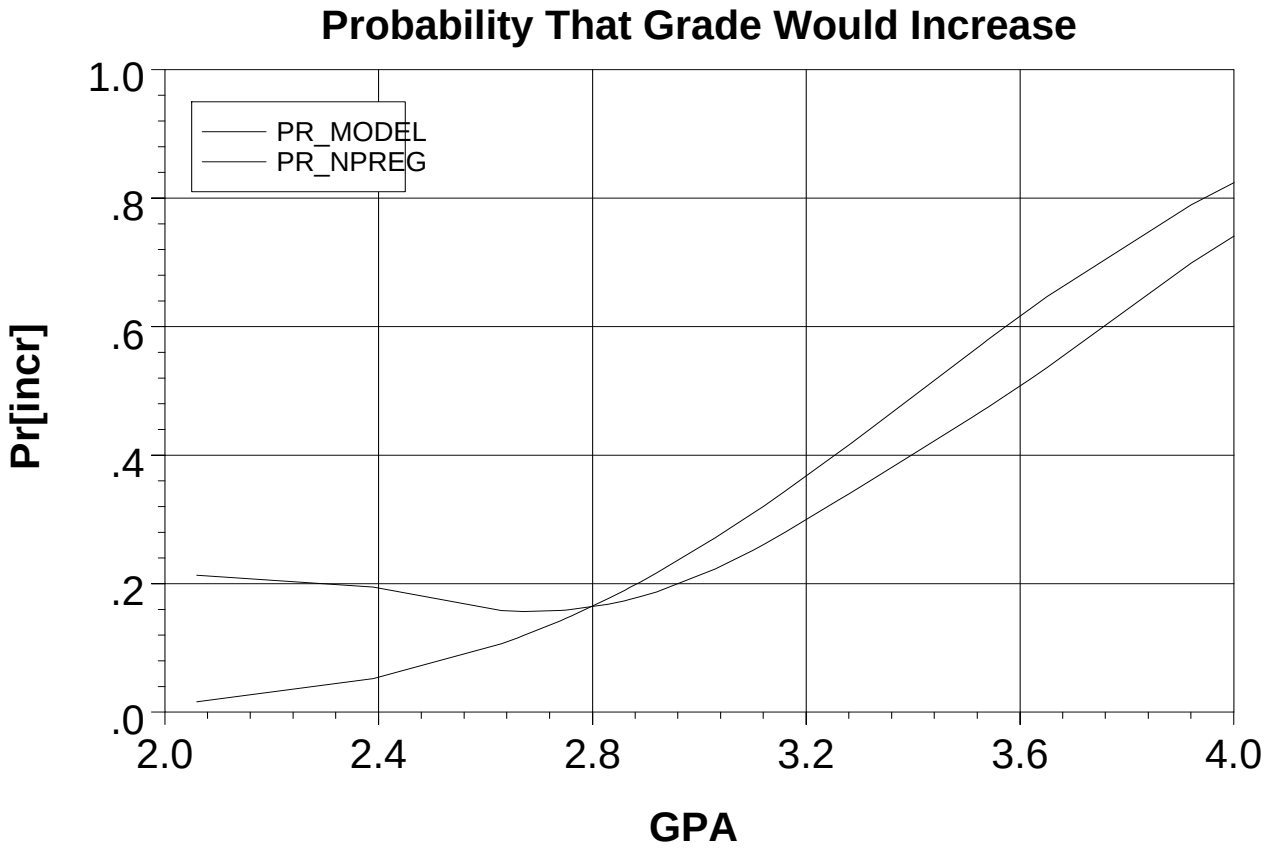
A graph of the logistic function *G(z) = exp(z)/*[1+*exp(z)*] looks like the standard normal, as seen in the following figure, but does not rise or fall to 1.00 and 0.00 as fast:

Graph of Logistic Function



Nonparametrics

As outlined in Becker and Greene (2001), recent developments in theory and computational procedures enable researchers to work with nonlinear modeling of all sorts as well as nonparametric regression techniques. As an example of what can be done consider the widely cited economic education application in Spector and Mazzeo (1980). They estimated a probit model to shed light on how a student's performance in a principles of macroeconomics class relates to his/her grade in an intermediate macroeconomics class, after controlling for such things as grade point average (GPA) going into the class. The effect of GPA on future performance is less obvious than it might appear at first. Certainly it is possible that students with the highest GPA would get the most from the second course. On the other hand, perhaps the best students were already well equipped, and if the second course catered to the mediocre (who had more to gain and more room to improve) then a negative relationship between GPA and increase in grades (GRADE) might arise. A negative relationship might also arise if artificially high grades were given in the first course. The below figure provides an analysis similar to that done by Spector and Mazzeo (using a subset of their data).



In this figure, the horizontal axis shows the initial grade point average of students in the study. The vertical axis shows the relative frequency of the incremental grades that increase from the first to the second course. The solid curve shows the estimated relative frequency of grades that improve in the second course using a probit model (the one used by the authors). These estimates suggest a positive relationship between GPA and the probability of grade improvement in the second macroeconomics throughout the GPA range. The dashed curve in the figure provides the results using a much less-structured nonparametric regression model.[iii] The conclusion reached with this technique is qualitatively similar to that obtained with the probit model for GPAs above 2.6, where the positive relationship between GPA and the probability of grade improvement can be seen, but it is materially different for those with GPAs lower than 2.6, where a negative relationship between GPA and the probability of grade improvement is found. Possibly these poorer students received gift grades in the introductory macroeconomics course.

There are other alternatives to least squares that economic education researchers can employ in programs such as LIMDEP, STATA and SAS. For example, the least-absolute-deviations approach is a useful device for assessing the sensitivity of estimates to outliers. It is likely that examples can be found to show that even if least-squares estimation of the conditional mean is a better estimator in large samples, least-absolute-deviations estimation of the conditional median performs better in small samples. The critical point is that economic education researchers must recognize that there are and will be new alternatives to modeling and

estimation routines as currently found in *Journal of Economic Education* articles and articles in the other journals that publish this work, as listed in Lo, Wong and Mixon (2008). In this module and in the remaining three, only passing mention will be given to these emerging methods of analysis. The emphasis will be on least-squares and maximum-likelihood estimations of continuous and discrete data-generating processes that can be represented parametrically.

INDIVIDUAL OBSERVATIONS OR GROUP AVERAGES: WHAT IS THE UNIT OF ANALYSIS?

In Becker (2004), I called attention to the implications of working with observations on individuals versus working with averages of individuals in different groupings. For example, what is the appropriate unit of measurement for assessing the validity of student evaluations of teaching (as reflected, for example, in the relationship between student evaluations of teaching and student outcomes)? In the case of end-of-term student evaluations of instructors, an administrator's interest may not be how students as individuals rate the instructor but how the class as a whole rates the instructor. Thus, the unit of measure is an aggregate for the class. There is no unique aggregate, although the class mean or median response is typically used.[iv] For the assessment of instructional methods, however, the unit of measurement may arguably be the individual student in a class and not the class as a unit. Is the question: how is the i^{th} student's learning affected by being in a classroom where one versus another teaching method is employed? Or is the question: how is the class's learning affected by one method versus another? The answers to these questions have implications for the statistics employed and interpretation of the results obtained.[v]

Hake (1998) reported that he has test scores for 6,542 individual students in 62 introductory physics courses. He works only with mean scores for the classes; thus, his effective sample size is 62, and not 6,542. The 6,542 students are not irrelevant, but they enter in a way that I did not find mentioned by Hake. The amount of variability around a mean test score for a class of 20 students versus a mean for 200 students cannot be expected to be the same. Estimation of a standard error for a sample of 62, where each of the 62 means receives an equal weight, ignores this heterogeneity.[vi] Francisco, Trautman, and Nicoll (1998) recognized that the number of subjects in each group implies heterogeneity in their analysis of average gain scores in an introductory chemistry course. Similarly, Kennedy and Siegfried (1997) made an adjustment for heterogeneity in their study of class size on student learning in economics.

Fleisher, Hashimoto, and Weinberg (2002) considered the effectiveness (in terms of student course grades and persistences) of 47 foreign graduate student instructors versus 21 native English speaking graduate student instructors in an environment in which English is the language of the majority of their undergraduate students. Fleisher, Hashimoto, and Weinberg recognized the loss of information in using the 92 mean class grades for these 68 graduate student instructors, although they did report aggregate mean class grade effects with the corrected heterogeneity adjustment for standard errors based on class size. They preferred to look at 2,680 individual undergraduate results conditional on which one of the 68 graduate student instructors each of the undergraduates had in any one of 92 sections of the course. To

ensure that their standard errors did not overstate the precision of their estimates when using the individual student data, Fleisher, Hashimoto, and Weinberg explicitly adjusted their standard errors for the clustering of the individual student observations into classes using a procedure akin to that developed by Moulton (1986).[vii]

Whatever the unit of measure for the dependent variable (aggregate or individual) the important point here is recognition of the need for one of two adjustments that must be made to get the correct standard errors. If an aggregate unit is employed (e.g., class means) then an adjustment for the number of observations making up the aggregate is required. If individual observations share a common component (e.g., students grouped into classes) then the standard errors reflect this clustering. Computer programs such as LIMDEP (NLOGIT), SAS and STATA can automatically perform both of these adjustments.

ANALYSIS OF VARIANCE (ANOVA) AND HYPOTHESES TESTING

Student of statistics are familiar with the *F* statistic as computed and printed in most computer regression routines under a banner "Analysis of Variance" or just ANOVA. This *F* is often presented in introductory statistics textbooks as a test of the overall all fit or explanatory power of the regression. I have learned from years of teaching econometrics that it is better to think of this test as one of all population model slope coefficients are zero (the explanatory power is not sufficient to conclude that there is any relations between the *x*s and *y* in the population) versus the alternative that at least one slope coefficient is not zero (there is some explanatory power). Thinking of this *F* statistic as just a joint test of slope coefficients, makes it easier to recognize that an *F* statistics can be calculated for any subset of coefficients to test for joint significance within the subset. Here I present the theoretical underpinnings for extensions of the basic ANOVA to tests of subsets of coefficients. Parts two three and four provide the corresponding commands to do these tests in LIMDEP, STATA and SAS.

As a starting point to ANOVA consider the *F* statistics that is generated by most computer programs. This *F* calculation can be viewed as a decomposition or partitioning of the dependent variable into two components (intercept and slopes) and a residual:

$$\mathbf{y} = \mathbf{i}b_1 + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e}$$

where $\mathbf{i}$ is the column of 1's in the X matrix associated with the intercept b_1 and X_2 is the remaining *(k–1)* explanatory *x* variables associated with the *(k–1)* slope coefficients in the $\mathbf{b}_2$ vector. The total sum of squared deviations

$$\text{TotSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i)^2 - n\bar{y}^2 = (\mathbf{y}'\mathbf{y} - n\bar{y}^2)$$

measures the amount of variability in *y* around $\bar{y}$, which ignoring any effect of the *x*s (in essence the b_2 vector is assumed to be a vector of zeros). The residual sum of squares

$$\text{ResSS} = \sum_{i=1}^{n}(e_i)^2 = \mathbf{e}'\mathbf{e}$$

measures the amount of variability in y around $\hat{y}$, which lets b_1 and b_2 assume their least squares values.

Partitioning of y in this manner enables us to test the contributions of the xs to explaining variability in the dependent variable. That is,

$$H_0 : \beta_2 = \beta_3 = ...\beta_k = 0 \text{ versus } H_A : \text{at least one slope coefficient is not zero.}$$

For calculating the F statistic, computer programs use the equivalent of the following:

$$F = \frac{[(\mathbf{y'y} - n\bar{y}^2) - \mathbf{e'e}]/[(n-1)-(n-K)]}{\mathbf{e'e}/(n-K)} = \frac{[(\mathbf{y'y} - n\bar{y}^2) - \mathbf{e'e}]/(K-1)}{\mathbf{e'e}/(n-K)} = \frac{(\text{TotSS} - \text{ResSS})/(K-1)}{\text{ResSS}/(n-K)}$$

This F is the ratio of two independently distributed Chi-square random variables adjusted for their respective degrees of freedom. The relevant decision rule for rejecting the null hypothesis is that the probability of this calculated F value or something greater, with $K - 1$ and $n - K$ degrees of freedom, is less than the typical (0.10, 0.05 or 0.01) probabilities of a Type I error.

Calculation of the F statistic in this manner, however, is just a special case of running two regressions: a restricted and an unrestricted. One regression was computed with all the slope coefficients set equal (or restricted) to zero so Y is regressed only on the column of ones. This **restricted regression** is the same as using $\bar{Y}$ to predict Y regardless of the values of the xs. This **restricted residual sum of squares**, $\mathbf{e}_r'\mathbf{e}_r$, is what is usually called the **total sum of squares**, TotSS = $\mathbf{y'y} - n\bar{y}^2$. The unrestricted regression allows all of the slope coefficients to find their values to minimize the residual sum of squares, which is thus called the **unrestricted residual sum of squares,** $\mathbf{e}_u'\mathbf{e}_u$, and is usually just list in a computer printout as the residual sum of squares ResSS= $\mathbf{e'e}$.

The idea of a restricted and unrestricted regression can be extended to test any subset of coefficients. For example, say the full model for a posttest Y is

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i .$$

Let's say the claim is made that x_3 and x_4 do not affect Y. One way to interpret this is to specify that $\beta_3 = \beta_4 = 0$, but $\beta_2 \neq 0$. The dependent variable is again decomposed into two components but now x_1 is included with the intercept in the partitioning of the $\mathbf{X}$ matrix:

$$\mathbf{y} = \mathbf{X_1 b_1} + \mathbf{X_2 b_2} + \mathbf{e} .$$

where $\mathbf{X_1}$ is the $n \times 2$ matrix, with the first column containing ones and the second observations on x_1 ($\mathbf{b_1}$ contains the y intercept and x_1 slope coefficient) and $\mathbf{X_2}$ is the $n \times 2$ matrix, with two columns for x_3 and x_4 ($\mathbf{b_2}$ contains x_3 and x_4 slope coefficients). If the claim about x_3 and x_4 not

belonging in the explanation of Y is true, then the two slope coefficients in $\mathbf{b_2}$ should be set to zero because the true model is the restricted specification

$$Y_i = \beta_1 + \beta_2 x_{i2} + \varepsilon_i .$$

The null hypotheses is $H_0 : \beta_3 = \beta_4 = 0$; i.e., x_2 might affect Y but x_3 and x_4 do not affect Y.

The alternative hypothesis is $H_A : \beta_3 \neq 0$ or $\beta_4 \neq 0$; i.e.,. x_3 and x_4 both affect Y.

The F statistic to test the hypotheses is then

$$F = \frac{[\mathbf{e}_\mathbf{r}'\mathbf{e}_\mathbf{r} - \mathbf{e}_u'\mathbf{e}_u]/[(n-K_r)-(n-K_u)]}{\mathbf{e}_u'\mathbf{e}_u/(n-K_u)} ,$$

where the restricted residual sum of squares $\mathbf{e}_\mathbf{r}'\mathbf{e}_\mathbf{r}$ is obtained from a simple regression of Y on x_2, including a constant, and the unrestricted sum of squared residuals $\mathbf{e}_u'\mathbf{e}_u$ is obtained from a regression of Y on x_2, x_3 and x_4, including a constant.

In general, it is best to test the overall fit of the regression model before testing any subset or individual coefficients. The appropriate hypotheses and F statistic are

$$H_0 : \beta_2 = \beta_3 = ... = \beta_K = 0 \quad (\text{or } H_0 : R^2 = 0)$$
$$H_A : \text{at least one slope coefficient is not zero} \quad (\text{or } H_0 : R^2 \neq 0)$$

$$F = \frac{[(\mathbf{y}'\mathbf{y} - n\bar{y}^2) - \mathbf{e}'\mathbf{e}]/(K-1)}{\mathbf{e}'\mathbf{e}/(n-K)} .$$

If the calculated value of this F is significant, then subsets of the coefficients can be tested as

$$H_0 : \beta_s = \beta_t = ... = 0$$
$$H_A : \text{at least one of these slope coefficient is not zero}$$

$$F = \frac{[\mathbf{e}_\mathbf{r}'\mathbf{e}_\mathbf{r} - \mathbf{e}_u'\mathbf{e}_u]/[(K_u - q)]}{\mathbf{e}_u'\mathbf{e}_u/(n-K_u)} , \text{ for } q = k \text{ – number of restrictions.}$$

The restricted residual sum of squares $\mathbf{e}_\mathbf{r}'\mathbf{e}_\mathbf{r}$ is obtained by a regression on only the q xs that did not have their coefficients restricted to zero. Any number of subsets of coefficients can be tested in this framework of restricted and unrestricted regressions as summarized in the following table.

SUMMARY FOR ANOVA TESTING

PANEL A. TRADITIONAL ANOVA FOR TESTING
$R^2 = 0$ versus $R^2 \neq 0$

Sum of Squares	Source	Degrees of Freedom	Mean Square
Total (to be explained)	$\mathbf{y'y} - n\bar{y}^2$	$n - 1$	s_y^2
Residual or Error (unexplained)	$\mathbf{e'e}$	$n - k$	s_e^2
Regression or Model (explained)	$\mathbf{b'X'y} - n\bar{y}^2$	$k - 1$	

$$F = \frac{R^2/(K-1)}{(1-R^2)/(n-K)} = \frac{[1-(\text{ResSS/TotSS})]/(K-1)}{(\text{ResSS/TotSS})/(n-K)} = \frac{(\text{TotSS-ResSS})/(K-1)}{\text{ResSS}/(n-K)}$$

PANEL B. RESTRICTED REGRESSION FOR TESTING ALL THE SLOPES $\beta_2 = \beta_3 = ... = \beta_K = 0$

Sum of Squares	Source	Degrees of Freedom	Mean Square
Restricted (all slopes = 0)	$\mathbf{e}_\mathbf{r}'\mathbf{e}_\mathbf{r} = \mathbf{y'y} - n\bar{y}^2$	$n - 1$	s_y^2
Unrestricted	$\mathbf{e}_u'\mathbf{e}_u = \mathbf{e'e}$	$n - k$	s_e^2
Improvement	$\mathbf{b'X'y} - n\bar{y}^2$	$k - 1$	

$$F = \frac{[\text{Restricted ResSS}(slopes = 0) - \text{Unrestricted ResSS}](K-1)}{\text{Unrestricted ResSS}/(n-k)}$$

PANEL C. RESTRICTED REGRESSION FOR TESTING A SUBSET OF COEFFICIENTS $\beta_s = \beta_t = ... = 0$

Sum of Squares	Source	Degrees of Freedom
Restricted ($\beta_s = \beta_t = ... = 0$)	$\mathbf{e}_\mathbf{r}'\mathbf{e}_\mathbf{r}$	$n - q$, for $q =$ k – number of restrictions
Unrestricted	$\mathbf{e}_u'\mathbf{e}_u$	$n - k$
Improvement	$\mathbf{e}_\mathbf{r}'\mathbf{e}_\mathbf{r} - \mathbf{e}_u'\mathbf{e}_u$	$K - q$

$$F = \frac{[\text{Restricted ResSS}(subset = 0) - \text{Unrestricted ResSS}](K-q)}{\text{Unrestricted ResSS}/(n-k)}$$

The *F* test of subsets of coefficients is ideal for testing interactions. For instance, to test for the treatment effect in the following model both β_4 and β_5 must be jointly tested against zero:

$$ChangeScore = \beta_1 + \beta_2 female + \beta_3 femaletreatment + \beta_4 treatment + \beta_5 GPA + \varepsilon$$

$$H_o : \beta_4 = \beta_5 = 0 \qquad H_A : \beta_4 \text{ or } \beta_5 \neq 0$$

where "*ChangeScore*" is the difference between a student's test scores at the end and beginning of a course in economics, *female* = 1, if female and 0 if male, "*treatment*" = 1, if in the treatment group and 0 if not, and "*GPA*" is the student's grade point average before enrolling in the course.

The *F* test of subsets of coefficients is also ideal for testing for fixed effects as reflected in sets of dummy variables. For example, in Parts Two, Three and Four an *F* test is performed to check whether there is any fixed difference in test performance among four classes taking economics using the following assumed data generating process:

$$post = \beta_1 + \beta_2 pre + \beta_3 class1 + \beta_4 class2 + \beta_5 class3 + \varepsilon$$

$$H_o : \beta_3 = \beta_4 = \beta_5 = 0 \qquad H_A : \beta_3, \beta_4 \text{ or } \beta_5 \neq 0$$

where "post" is a student's post-course test score, "pre" is the student's pre-course test score, and "class" identifies to which one of the four classes the students was assigned, e.g., class3 = 1 if student was in the third class and class3 = 0 if not. The fixed effect for students in the fourth class (*class*1, *class*2 and *class*3 are zero)) is captured in the intercept β_1.

It is important to notices in this test of fixed class effects that the relationship between the *post* and *pre* test (as reflected in the slope coefficient β_2) is assumed to be the same regardless of the class to which the student was assigned. The next section described a test for any structural difference among the groups.

TESTING FOR A SPECIFICATION DIFFERENCE ACROSS GROUPS

Earlier in our discussion of the difference in difference or change score model, a 0-1 bivariate dummy variable was introduced to test for a difference in intercepts between a treatment and control group, which could be done with a single coefficient *t* test. However, the expected difference in the dependent variable for the two groups might not be constant. It might vary with the level of the independent variables. Indeed, the appropriate model might be completely different for the two groups. Or, it might be the same.

Allowing for any type of difference between the control and experimental variables implies that the null and alternative hypotheses are

H_0: $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \beta$
H_A: $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$,

where the $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are KX1 column vectors containing the K coefficients $\beta_1, \beta_2, \beta_3, \ldots \beta_K$ for the control $\boldsymbol{\beta}_1$ and the experimental $\boldsymbol{\beta}_2$ groups . Let $\mathbf{X}_1$ and $\mathbf{X}_2$ contain the observations on the explanatory variables corresponding to the $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, including the column of ones for the constant β_1. The unrestricted regression is captured by two separate regressions:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}.$$

That is, the unrestricted model is estimated by fitting the two regressions separately. The unrestricted residual sum of squares is obtained by adding the residuals from these two regressions. The unrestricted degrees of freedom are similarly obtained by adding the degrees of freedom of each regression.

The restricted regression is just a regression of y on the xs with no group distinction in beta coefficients:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} [\boldsymbol{\beta}] + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \end{bmatrix}.$$

That is , the restricted residual sum of squares is obtained from a regression in which the data from the two groups are pooled and a single set of coefficients is estimated for the pooled data set.

The appropriate F statistic is

$$F = \frac{[\text{Restricted ResSS}(\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2) - \text{Unrestricted ResSS}]/K}{\text{Unrestricted ResSS}/[n - 2K]},$$

where unrestricted ResSS = residuals sum of squares from a regression on only those in the control plus residuals from a regression on only those in the treatment groups.

Thus, to test for structure change over J regimes, run separate regressions on each and add up the residuals to obtain the unrestricted residual sum of squares, ResSS_u,with df = $n - JK$. The restricted residual sum of squares is ResSS_r, with df = $n - K$.

$$H_0 : \boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \ldots = \boldsymbol{\beta_J} \text{ and } H_a : \boldsymbol{\beta}'s \text{ are not equal}$$

$$F = \frac{(\text{ResSS}_\text{r} - \text{ResSS}_\text{u}) / K(J-1)}{\text{ResSS}_\text{u} / (n - JK)}$$

This form of testing for a difference among groups is known in economics as a Chow Test. As demonstrated in Part Two using LIMDEP and Parts Three and Four using STATA and SAS, any number of subgroups could be tested by adding up their individual residual sums of squares and degrees of freedom to form the unrestricted residual sums of squares and matching degrees of freedom.

OTHER TEST STATISTICS

Depending on the nature of the model being estimated and the estimation method, computer programs will produce alternatives to the *F* statistics for testing (linear and nonlinear) restrictions and structural changes. What follows is only an introduction to these statistics that should be sufficient to give meaning to the numbers produced based on our discussion of ANOVA above.

The **Wald (*W*) statistic** follows the Chi-squared distribution with *J* degrees of freedom, reflecting the number of restrictions imposed:

$$W = \frac{(\mathbf{e}_r'\mathbf{e}_r - \mathbf{e}_u'\mathbf{e}_u)}{\mathbf{e}_u'\mathbf{e}_u / n} \sim \chi^2(J) \ .$$

If the model and the restriction are linear, then

$$W = \frac{nJ}{n-k} F = \frac{J}{1-(k/n)} F,$$

which for large *n* yields the asymptotic results

$$W = JF \ .$$

The **likelihood ratio (LR) test** is formed by twice the difference between the log-likelihood function for an unrestricted regression (L_{ur}) and its value for the restricted regression (L_r).

$$LR = 2(L_{ur} - L_r) \geq 0 \ .$$

Under the null hypothesis that the J restrictions are true, LR is distributed Chi-square with *J* degrees of freedom.

The relationship between the likelihood ratio test and Wald test can be shown to be

$$LR = \frac{n(\mathbf{e}_r'\mathbf{e}_r - \mathbf{e}_u'\mathbf{e}_u)}{\mathbf{e}_u'\mathbf{e}_u} - \frac{n(\mathbf{e}_r'\mathbf{e}_r - \mathbf{e}_u'\mathbf{e}_u)^2}{2\mathbf{e}_u'\mathbf{e}_u} \leq W .$$

The **Lagrange multiplier test (LM)** is based on the gradient (or score) vector

$$\begin{bmatrix} \partial L/\partial \beta \\ \partial L/\partial \sigma^2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\boldsymbol{\varepsilon}/\sigma^2 \\ -(n/2\sigma^2) + (\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/2\sigma^4) \end{bmatrix}.$$

where, as before, to evaluate this score vector with the restrictions we replace $\mathbf{e} = \mathbf{y} - \mathbf{Xb}$ with $\mathbf{e_r} = \mathbf{y} - \mathbf{Xb_r}$. After sufficient algebra, the Lagrange statistic is defined by

$$LM = n\mathbf{e}_r'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_r/\mathbf{e}_r'\mathbf{e}_r = nR^2 \sim \chi^2(J) ,$$

where R^2 is the conventional coefficient of determination from a regression of $\mathbf{e}_r$ on $\mathbf{X}$, where e_r has a zero mean (i.e., only slopes are being tested). It can also be shown that

$$LM = \frac{nJ}{(n-k)[1 + JF/(n-k)]} F = \frac{W}{1 + (W/n)} .$$

Thus, $LM \leq LR \leq W$.

DATA ENTRY AND ESTIMATION

I like to say to students in my classes on econometrics that theory is easy, data are hard – hard to find and hard to get into a computer program for statistical analysis. In this first of four parts in Module One, I provided an introduction to the theoretical data generating processes associated with continuous versus discrete dependent variables. Parts Two, Three and Four concentrate on getting the data into one of three computer programs: LIMDEP (NLOGIT), STATA and SAS. Attention is also given to estimation and testing within regressions employing individual cross-sectional observations within these programs. Later modules will address complications introduced by panel data and sources of endogeneity.

REFERENCES

Allison, Paul D. (1990). "Change Scores as Dependent Variables in Regression Analysis," *Sociological Methodology*, Vol. 20: 93-114.

Bandiera, Oriana, Valentino Larcinese and Imron Rasul (2010). "Blissful Ignorance? Evidence from a Natural Experiment on the Effect of Individual Feedback on Performance , " IZA Seminar, Bonn Germany, December 5, 2009. January 2010 version downloadable at http://www.iza.org/index_html?lang=en&mainframe=http%3A//www.iza.org/en/webcontent/events/izaseminar_description_html%3Fsem_id%3D1703&topSelect=events&subSelect=seminar

Becker, William E. (2004). "Quantitative Research on Teaching Methods in Tertiary Education," in W. E. Becker and M. L. Andrews (eds), *The Scholarship of Teaching and Learning in Higher Education: Contributions of the Research Universities*, Indiana University Press: 265-309.

Becker, William E. (Summer 1983). "Economic Education Research: Partr III, Statistical Estimation Methods," *Journal of Economic Education*, Vol. 14 (Sumer): 4-15

Becker, William E. and William H. Greene (2001). "Teaching Statistics and Econometrics to Undergraduates," *Journal of Economic Perspectives*, Vol. 15 (Fall): 169-182.

Becker, William E., William Greene and Sherwin Rosen (1990). "Research on High School Economic Education," *American Economic Review*, Vol. 80, (May): 14-23, and an expanded version in *Journal of Economic Education*, Summer 1990: 231-253.

Becker, William E. and Peter Kennedy (1992). "A Graphical Exposition of the Ordered Probit," with P. Kennedy, *Econometric Theory*, Vol. 8: 127-131.

Becker, William E. and Michael Salemi (1977). "The Learning and Cost Effectiveness of AVT Supplemented Instruction: Specification of Learning Models," *Journal of Economic Education* Vol. 8 (Spring) : 77-92.

Becker, William E. and Donald Waldman (1989). "Graphical Interpretation of Probit Coefficients," *Journal of Economic Education*, Vol. 20 (Fall): 371-378.

Campbell, D., and D. Kenny (1999). *A Primer on Regression Artifacts*. New York: The Guilford Press.

Fleisher, B., M. Hashimoto, and B. Weinberg. 2002. "Foreign GTAs can be Effective Teachers of Economics." *Journal of Economic Education*, Vol. 33 (Fall): 299-326.

Francisco, J. S., M. Trautmann, and G. Nicoll. 1998. "Integrating a Study Skills Workshop and Pre-Examination to Improve Student's Chemistry Performance." *Journal of College Science Teaching*, Vol. 28 (February): 273-278.

Friedman, M. 1992. "Communication: Do Old Fallacies Ever Eie?" *Journal of Economic Literature,* Vol. 30 (December): 2129-2132.

Greene, William (2003). *Econometric Analysis*. 5th Edition, New Jersey: Prentice Hall.

Hake, R. R. (1998). "Interactive-Engagement versus Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses." *American Journal of Physics*, Vol. 66 (January): 64-74.

Hanushek, Eric A. (1986). "The Economics of Schooling: Production and Efficiency in Public Schools," *Journal of Economic Literature*, 24(September)): 1141-1177.

Lo, Melody, Sunny Wong and Franklin Mixon (2008). "Ranking Economics Journals Economics Departments, and Economists Using Teaching-Focused Research Productivity." *Southern Economics Journal* 2008, 74(January): 894-906.

Moulton, B. R. (1986). "Random Group Effects and the Precision of Regression Estimators." *Journal of Econometrics*, Vol. 32 (August): 385-97.

Kennedy, P., and J. Siegfried. (1997). "Class Size and Achievement in Introductory Economics: Evidence from the TUCE III Data." *Economics of Education Review*, Vol. 16 (August): 385-394.

Kvam, Paul. (2000). "The Effect of Active Learning Methods on Student Retention in Engineering Statistics." *American Statistician*, 54 (2): 136-40.

Ramsden, P. (1998). "Managing the Effective University." *Higher Education Research & Development*, 17 (3): 347-70.

Salemi, Michael and George Tauchen. 1987. "Simultaneous Nonlinear Learning Models." In W. E. Becker and W. Walstad, eds., *Econometric modeling in economic education research*, pp. 207-23. Boston: Kluwer-Nijhoff.

Spector, Lee C. and Michael Mazzeo (1980). "Probit Analysis and Economic Education" *Journal of Economic Education,* Vol. 11(Spring), 11(2): 37-44.

Wainer, H. 2000. "Kelley's Paradox." *Chance*, 13 (Winter): 47-48.

ENDNOTES

[i] Let the change or gain score be $\Delta y = [y_1 - y_0]$, which is the posttest score minus the pretest score, and let the maximum change score be $\Delta y_{\max} = [y_{\max} - y_0]$, then

$$\frac{\partial(\Delta y / \Delta y_{\max})}{\partial y_0} = \frac{-(y_{\max} - y_1)}{(y_{\max} - y_0)^2} \leq 0, \text{for } y_{\max} \geq y_1 \geq y_0$$

[ii] Let the posttest score (y_1) and pretest score (y_0) be defined on the same scale, then the model of the i^{th} student's pretest is

$$y_{0i} = \beta_0 (ability)_i + v_{0i},$$

where β_0 is the slope coefficient to be estimated, v_{0i} is the population error in predicting the i^{th} student's pretest score with ability, and all variables are measured as deviations from their means. The i^{th} student's posttest is similarly defined by

$$y_{1i} = \beta_1 (ability)_i + v_{1i}$$

The change or gain score model is then

$$y_{1i} - y_{0i} = (\beta_1 - \beta_0) ability + v_{1i} - v_{0i}$$

And after substituting the pretest for unobserved true ability we have

$$\Delta y_i = (\Delta\beta / \beta_0) y_{0i} + v_{1i} - v_{0i}[1 + (\Delta\beta / \beta_0)]$$

The least squares slope estimator $(\Delta b / b_0)$ has an expected value of

$$E(\Delta b / b_0) = E(\sum_i \Delta y_1 y_{0i} / \sum_i y_{0i}^2)$$
$$E(\Delta b / b_0) = (\Delta\beta / \beta_0) + E\{\sum_i [v_{1i} - v_{0i} - v_{01}(\Delta\beta / \beta_0)]_{1i} y_{0i} / \sum_i y_{0i}^2\}$$
$$E(\Delta b / b_0) \leq (\Delta\beta / \beta_0)$$

Although v_{1i} and y_{0i} are unrelated, $E(v_{1i}\ y_{0i}) = 0$, v_{0i} and y_{0i} are positively related, $E(v_{0i}\ y_{0i}) > 0$; thus, $E(\Delta b / b_0) \leq \Delta\beta / \beta_0$. Becker and Salemi (1977) suggested an instrumental variable technique to address this source of bias and Salemi and Tauchen (1987) suggested a modeling of the error term structure.

Hake (1998) makes no reference to this bias when he discusses his regressions and correlation of average normalized gain, average gain score and posttest score on the average pretest score. In

http://www.consecol.org/vol5/iss2/art28/, he continued to be unaware of, unable or unwilling to specify the mathematics of the population model from which student data are believed to be generated and the method of parameter estimation employed. As the algebra of this endnote suggests, if a negative relationship is expected between the gap closing measure

$$g = (posttest - pretest)/(maxscore - pretest)$$

and the pretest, but a least-squares estimator does not yield a significant negative relationship for sample data, then there is evidence that something is peculiar. It is the lack of independence between the pretest and the population error term (caused, for example, by measurement error in the pretest, simultaneity between *g* and the pretest, or possible missing but relevant variables) that is the problem. Hotelling received credit for recognizing this endogenous regressor problem (in the 1930s) and the resulting regression to the mean phenomenon. Milton Friedman received a Nobel prize in economics for coming up with an instrumental variable technique (for estimation of consumption functions in the 1950s) to remove the resulting bias inherent in least-squares estimators when measurement error in a regressor is suspected. Later Friedman (1992, p. 2131) concluded: "I suspect that the regression fallacy is the most common fallacy in the statistical analysis of economic data ..." Similarly, psychologists Campbell and Kenny (1999, p. xiii) stated: "Regression toward the mean is a artifact that as easily fools statistical experts as lay people." But unlike Friedman, Campbell and Kenny did not recognize the instrumental variable method for addressing the problem.

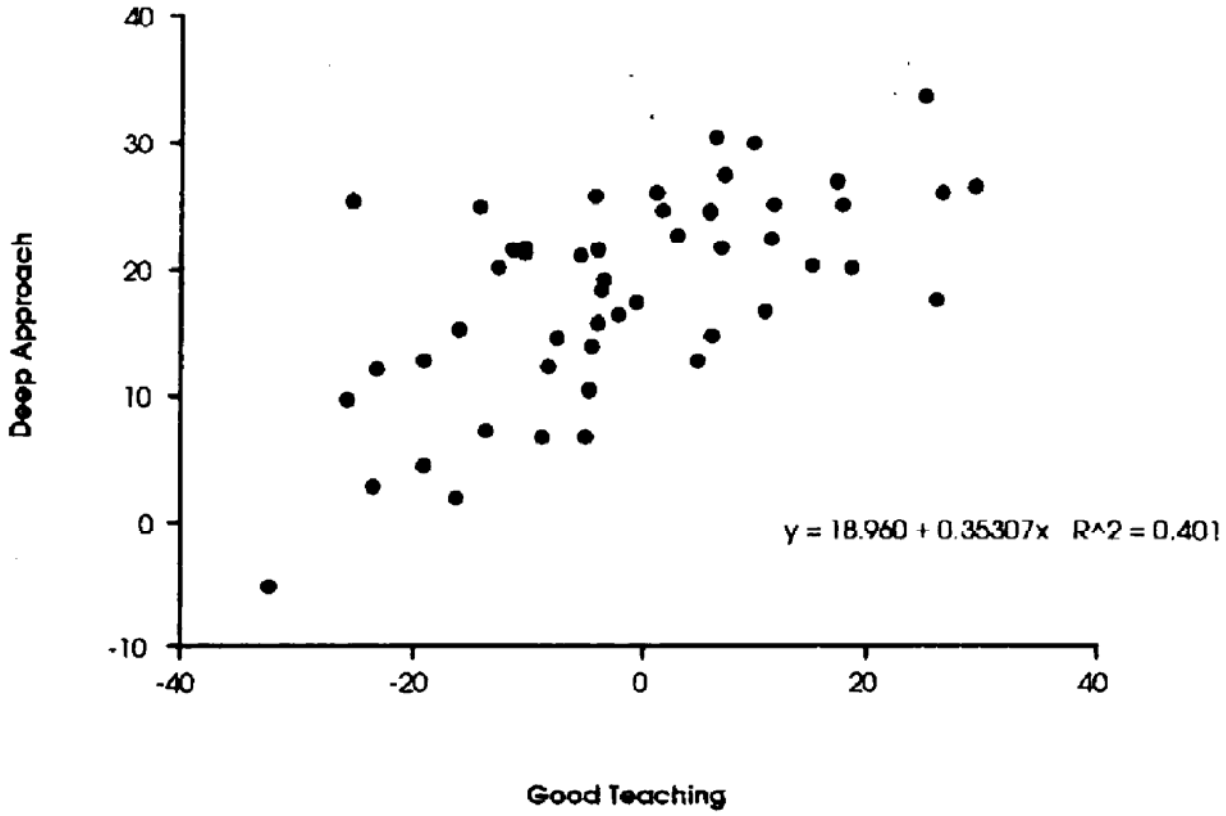
In an otherwise innovative study, Paul Kvam (2000) correctly concluded that there was insufficient statistical evidence to conclude that active-learning methods (primarily through integrating students' projects into lectures) resulted in better retention of quantitative skills than traditional methods, but then went out on a limb by concluding from a scatter plot of individual student pretest and posttest scores that students who fared worse on the first exam retain concepts better if they were taught using active-learning methods. Kvan never addressed the measurement error problem inherent in using the pretest as an explanatory variable. Wainer (2000) called attention to others who fail to take measurement error into account in labeling students as "strivers" because their observed test scores exceed values predicted by a regression equation.

[iii] The plot for the probability model was produced by first fitting a probit model of the binary variable GRADE, as a function of GPA. This produces a functional relationship of the form $\text{Prob}(\text{GRADE} = 1) = \Phi(\alpha + \beta\text{GRADE})$, where estimates of α and β are produced by maximum likelihood techniques. The graph is produced by plotting the standard normal distribution function, $\Phi(\alpha + \beta\text{GRADE})$ for the values of GRADE in the sample, which range between 2.0 and 4.0, then connecting the dots. The nonparametric regression, although intuitively appealing because it can be viewed as making use of weighted relative frequencies, is computationally more complicated. [Today the binomial probit model can be fitted with just about any statistical package but software for nonparametric estimation is less common. LIMDEP (NLOGIT) version 8.0 (Econometric Software, Inc., 2001) was used for both the probit and nonparametric estimations.] The nonparametric approach is based on the assumption that there is some as yet

unknown functional relationship between the Prob(GRADE = 1) and the independent variable, GPA, say Prob(Grade = 1 | GPA) = F(GPA). The probit model based on the normal distribution is one functional candidate, but the normality assumption is more specific than we need at this point. We proceed to use the data to find an approximation to this function. The form of the 'estimator' of this function is $F(\text{GPA}^*) = \Sigma_{\text{i = all observations}}\, w(\text{GPA}^* - \text{GPA}_\text{i})\text{GRADE}_\text{i}$. The weights, '$w$(.),' are positive weight functions that sum to 1.0, so for any specific value GPA*, the approximation is a weighted average of the values of GRADE. The weights in the function are based on the desired value of GPA, that is GPA*, as well as all the data. The nature of the computation is such that if there is a positive relationship between GPA and GRADE =1, then as GPA* gets larger, the larger weights in the average shown above will tend to be associated with the larger values of GRADE. (Because GRADE is zeros and ones, this means that for larger values of GPA*, the weights associated with the observations on GRADE that equal one will generally be larger than those associated with the zeros.) The specific form of these weights is as follows: $w(\text{GPA}^* - \text{GPA}_\text{i}) = (1/A)\times(1/h)K[(\text{GPA}^* - \text{GPA}_\text{i})/h]$. The '$h$' is called the smoothing parameter, or bandwidth, K[.] is the 'kernel density function' and A is the sum of the functions, ensuring that the entire expression sums to one. Discussion of nonparametric regression using a kernel density estimator is given in Greene (2003, pp. 706-708). The nonparametric regression of GRADE on GPA plotted in the figure was produced using a logistic distribution as the kernel function and the following computation of the bandwidth: let r equal one third of the sample range of GPA and let s equal the sample standard deviation of GPA. The bandwidth is then $h = .9\times\text{Min}(r,s)/n^{1/5}$. (In spite of their apparent technical cache, bandwidths are found largely by experimentation. There is no general rule that dictates what one should use in a particular case, which is unfortunate because the shapes of kernel density plots are heavily dependent upon them.)

[iv] Unlike the mean, the median reflects relative but not absolute magnitude; thus, the median may be a poor measure of change. For example, the series 1, 2, 3 and the series 1, 2, 300 have the same median (2) but different means (2 versus 101).

[v] To appreciate the importance of the unit of analysis, consider a study done by Ramsden (1998, pp. 352-354) in which he provided a scatter plot showing a positive relationship between a y-axis index for his "deep approach" (aimed at student understanding versus "surface learning") and an x-axis index of "good teaching" (including feedback of assessed work, clear goals, etc.):



Ramsden's regression ($y = 18.960 + 0.35307x$) seems to imply that a decrease (increase) in the good teaching index by one unit leads to a 0.35307 decrease (increase) in the predicted deep approach index; that is, good teaching positively affects deep learning. But does it?

Ramsden (1998) ignored the fact that each of his 50 data points represent a type of institutional average that is based on multiple inputs; thus, questions of heteroscedasticity and the calculation of appropriate standard errors for testing statistical inference are relevant. In addition, because Ramsden reports working only with the aggregate data from each university, it is possible that within each university the relationship between good teaching (x) and the deep approach (y) could be negative but yet appear positive in the aggregate.

When I contacted Ramsden to get a copy of his data and his coauthored "Paper presented at the Annual Conference of the Australian Association for Research in Education, Brisbane (December 1997)," which was listed as the source for his regression of the deep approach index on the good teaching index in his 1998 published article, he confessed that this conference paper never got written and that he no longer had ready access to the data (email correspondence August 22, 2000).

Aside from the murky issue of Ramsden citing his 1997 paper, which he subsequently admitted does not exist, and his not providing the data on which the published 1998 paper is allegedly based, a potential problem of working with data aggregated at the university level can be seen

with three hypothetical data sets. The three regressions for each of the following hypothetical universities show a negative relationship for *y* (deep approach) and *x* (good teaching), with slope coefficients of –0.4516, –0.0297, and –0.4664, but a regression on the university means shows a positive relationship, with slope coefficient of +0.1848. This is a demonstration of "Simpson's paradox," where aggregate results are different from dissaggregated results.

University One

$\hat{y}(1) = 21.3881 - 0.4516x(1)$ Std. Error = 2.8622 $R^2 = 0.81$ n = 4
y(1): 21.8 15.86 26.25 14.72
x(1): -4.11 6.82 -5.12 17.74

University Two

$\hat{y}(2) = 17.4847 - 0.0297x(2)$ Std. Error = 2.8341 $R^2 = 0.01$ n = 8
y(2): 12.60 17.90 19.00 16.45 21.96 17.1 18.61 17.85
x(2): -10.54 -10.53 -5.57 -11.54 -15.96 -2.1 -9.64 12.25

University Three

$\hat{y}(3) = 17.1663 - 0.4664x(3)$ Std. Error = 2.4286 $R^2 = 0.91$ n = 12
y(3): 27.10 2.02 16.81 15.42 8.84 22.90 12.77 17.52 23.20 22.60 25.90
x(3): -23.16 26.63 5.86 9.75 11.19 –14.29 11.51 –0.63 –19.21 –4.89 –16.16

University Means

$\hat{y}(means) = 18.6105 + 0.1848x(means)$ Std. Error = 0.7973 $R^2 = 0.75$ n = 3
y(means): 19.658 17.684 17.735
x(means): 3.833 -6.704 -1.218

[vi] Let y_{it} be the observed test score index of the i^{th} student in the t^{th} class, who has an expected test score index value of μ_{it}. That is, $y_{it} = \mu_{it} + \varepsilon_{it}$, where ε_{it} is the random error in testing such that its expected value is zero, $E(\varepsilon_{it}) = 0$, and variance is σ^2, $E(\varepsilon_{it}^2) = \sigma^2$, for all i and t. Let $\bar{y}_t$ be the sample mean of a test score index for the t^{th} class of n_t students. That is, $\bar{y}_t = \bar{\mu}_t + \bar{\varepsilon}_t$ and $E(\bar{\varepsilon}_t^2) = \sigma^2 / n_t$. Thus, the variance of the class mean test score index is inversely related to class size.

[vii] As in Fleisher, Hashimoto, and Weinberg (2002), let y_{gi} be the performance measure of the i^{th} student in a class taught by instructor g, let F_g be a dummy variable reflecting a characteristics of the instructor (e.g., nonnative English speaker), let x_{gi} be a $(1 \times n)$ vector of the student's

observable attributes, and let the random error associated with the i^{th} student taught by the g^{th} instructor be ε_{gi} . The performance of the i^{th} student is then generated by

$$y_{gi} = F_g\gamma + x_{gi}\beta + \varepsilon_{gi}$$

where γ and β are parameters to be estimated. The error term, however, has two components: one unique to the i^{th} student in the g^{th} instructor's class (u_{gi}) and one that is shared by all students in this class (ξ_g): $\varepsilon_{gi} = \xi_g + u_{gi}$. It is the presence of the shared error ξ_g for which an adjustment in standard errors is required. The ordinary least squares routines employed by the standard computer programs are based on a model in which the variance-covariance matrix of error terms is diagonal, with element σ_u^2 . The presence of the ξ_g terms makes this matrix block diagonal, where each student in the g^{th} instructor's class has an off-diagonal element σ_ξ^2 .

In (May 11, 2008) email correspondence, Bill Greene called my attention to the fact that Moulton (1986) gave a specific functional form for the shared error term component computation. Fleisher, Hashimoto, and Weinberg actually used an approximation that is aligned with the White estimator (as presented in Parts Two, Three and Four of this module), which is the "CLUSTER" estimator in STATA. In LIMDEP (NLOGIT), Moulton's shared error term adjustment is done by first arranging the data as in a panel with the groups contained in contiguous blocks of observations. Then, the command is "REGRESS ; ... ; CLUSTER = spec. $" where "spec" is either a fixed number of observations in a group, or the name of an identification variable that contains a class number. The important point is to recognize that heterogeneity could be the result of each group having its own variance and each individual within a group having its own variance. As discussed in detail in Parts Two, Three and Four, heteroscedasticity in general is handled in STATA with the "ROBUST" command and in LIMDEP with the "HETRO" command.