postpositivist social research methodology, their value has been challenged by others.

—Martyn Hammersley

REFERENCES

Alvesson, M., & Sköldberg, K. (2000). *Reflexive methodology.* London: Sage.

Becker, H. S. (1967). Whose side are we on? *Social Problems, 14,* 239–247.

Bourdieu, P., & Waquant, L. J. D. (1992). *An invitation to reflexive sociology.* Chicago: University of Chicago Press.

Czyzewski, M. (1994). Reflexivity of actors versus reflexivity of accounts. *Theory, Culture and Society, 11,* 161–168.

Gouldner, A. V. (1973). *For sociology.* Harmondsworth, UK: Penguin.

Hammersley, M. (1999). Sociology, what's it for? A critique of the grand conception. *Sociological Research Online, 4*(3). Retrieved from http://www.socresonline.org.uk/socresonline/4/3/hammersley.html

Hammersley, M., & Atkinson, P. (1995). *Ethnography: Principles in practice.* London: Routledge.

Lynch, M. (2000). Against reflexivity as an academic virtue and source of privileged knowledge. *Theory, Culture and Society, 17*(3), 26–54.

Paechter, C. (1996). Power, knowledge and the confessional in qualitative research. *Discourse: Studies in the Politics of Education, 17*(1), 75–84.

Troyna, B. (1994). Reforms, research and being reflexive about being reflective. In D. Halpin & B. Troyna (Eds.), *Researching education policy.* London: Falmer.

Woolgar, S. (Ed.). (1988). *Knowledge and reflexivity.* London: Sage.

REGRESSION

Regression is the most widely used data analysis technique in the nonexperimental social sciences. In a regression model, a DEPENDENT VARIABLE, commonly labeled Y, is a function of one or more INDEPENDENT VARIABLES, commonly labeled X_1, X_2, and so on. The independent variables are assumed to explain, or at least predict, the phenomenon measured by the dependent variable. The simplest of regression models posits only one independent variable: $Y = a + b_1X_1 + e$. This bivariate regression says variable Y is a linear additive function of variable X_1, plus a constant (the fixed value represented by a) plus error (represented by e). Of principal interest is the influence of X_1 on Y, the regression coefficient represented by b_1. To obtain numerical estimates for the values a and b_1, a straight line is fitted to the observations on X_1 and Y by the method of ORDINARY LEAST SQUARES (OLS). A more complicated regression model posits two independent variables: $Y = a + b_1X_1 + b_2X_2 + e$. This multiple regression model says Y is a linear function of X_1 and X_2 (plus a constant term and an error term). A MULTIPLE REGRESSION ANALYSIS equation always has two or more independent variables. Below, the mechanics of bivariate and multivariate regression are reviewed in a DATA example. Then, assumptions, history, and recent developments are considered.
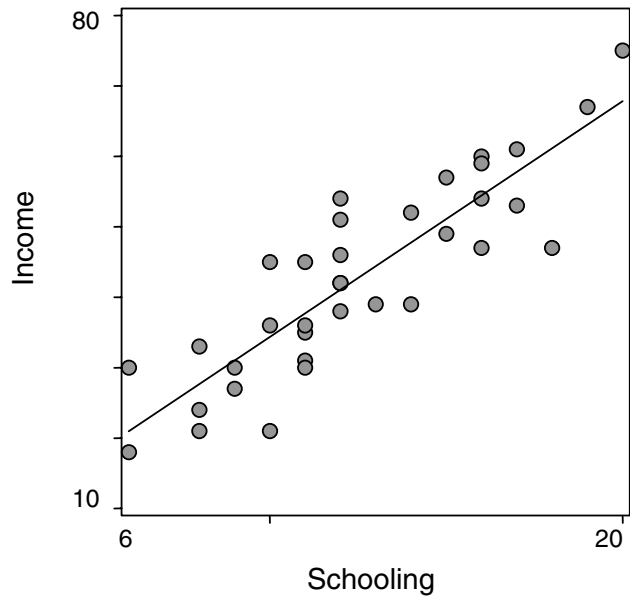
BIVARIATE REGRESSION

Social scientists often want to examine the relationship between two variables. To illustrate, an empirical example is unfolded. Imagine that educational sociologists in a small midwestern college seek information on the annual earnings of their students well after graduation. They want to know many things, including the influence of basic demographic forces, such as the socioeconomic background of the students' families. In particular, they seek to establish the link, if any, between the educational attainment of parents and the later job earnings of the child. To gather the necessary data on these and other research questions, they draw a RANDOM SAMPLE from the alumni list of 350 students who graduated 15 years ago. Because of cost considerations, they sample 1 out of 10, yielding a sample size of 35. Table 1 contains some of the data from the survey administered to them. In column 1 is the CODE number given the student. In column 2 is the parent education variable, measured as the number of years of formal schooling completed by the most educated parent (mother or father). In column 3 is the student income variable, measured as the reported gross annual income (in thousands of dollars) the student earned 15 years after graduation. In column 4 is the gender of the student, scored 1 = male or 0 = female. These data are listed as they would appear when entered into a typical computer STATISTICAL PACKAGE.

The research question at hand is whether parent education (variable X_1) helps account for later student income (variable Y). The dominant HYPOTHESIS is that they are positively related. The more educated the parental environment, the more child opportunities and incentives for learning and advancement and, ultimately, higher income on the job. Does an analysis of the data support the hypothesis? In bivariate regression, the first step is inspection of a SCATTERPLOT, as in Figure 1.

Table 1   A Data Set of Three Variables

Case	Variable 1 Schooling	Variable 2 Income	Variable 3 Gender
1	6	18	0
2	6	30	0
3	8	21	0
4	8	24	0
5	8	33	1
6	9	27	0
7	9	30	0
8	10	36	1
9	10	45	1
10	10	21	0
11	11	35	0
12	11	45	1
13	11	31	0
14	11	30	1
15	11	36	0
16	12	42	0
17	12	51	1
18	12	54	1
19	12	38	1
20	12	46	0
21	12	42	1
22	13	39	0
23	14	52	1
24	14	39	0
25	15	57	1
26	15	49	1
27	16	60	1
28	16	47	0
29	16	59	1
30	16	54	0
31	17	61	1
32	17	53	1
33	18	47	0
34	19	67	1
35	20	75	0

Figure 1   Scatterplot With Regression Line

Each dot, or point, in the plot represents the scores of a particular student on variables X_1 and Y. For example, student number 35 has a parent education score of 20 years and an income of $75,000. She is represented by the dot in the upper right-hand corner. Indeed, the point for each student locates itself at the intersection of an imaginary perpendicular line from his or her Y value (on the vertical axis) and of an imaginary perpendicular line from his or her X value (on the horizontal axis). Visual inspection of the scatter of points suggests that X_1 is positively related to Y; that is, lower values of X_1 tend to appear in the lower left-hand corner with lower values of Y, higher values of X_1 tend to appear in the upper right-hand corner with higher values

of Y. Overall, the constellation of dots, to the extent it suggests any geometric form, suggests a linear one. Put another way, a straight line would seem able to touch more of the points than any plausible curve. The formula for a straight line is $Y = a + bX$. The letter a is the CONSTANT value, where the line intercepts the y-axis, and the letter b denotes the SLOPE of the line. In Figure 1, a straight line is actually fitted to this scatterplot and expressed in the regression equation, predicted $Y = 0.88 + 3.35X_1$.

The regression line in Figure 1 is not arbitrarily drawn. Rather, it is the line of best fit, calculated from the LEAST SQUARES principle. Observe that some points fall on the line, and others do not. One sees positive errors (above the line) and negative errors (below the line). If all these errors (measured as the vertical distance of each point to the line) were simply added up, they would sum to zero, with the signs ($+$ and $-$) canceling out. However, when these errors are squared, there is no canceling out, and a measure of total error is produced, the sum of the squares of these errors (SSE). The sum of these squared errors is "least," the lowest possible value, hence the phrase "least squares." From calculus, the unique values for the intercept (a) and slope (b) that minimize the SSE are derived, and they constitute the ordinary least squares (OLS) solution. In our example, no other combination of INTERCEPT and slope values would produce a line that fit the data so well as the combination (0.88, 3.35) does.

The regression line serves as a PREDICTION EQUATION. Say the symbol for a predicted Y value is $\hat{Y}$, and $X = 12$. Then,

$$\hat{Y} = 0.88 + 3.35X$$
$$= 0.88 + 3.35(12)$$
$$= 0.88 + 40.2$$
$$\hat{Y} = 41.08.$$

The model predicts the student should earn about $41,000 annually, if the highest educational attainment of the parent was 12 years of schooling. Suppose now that $X = 13$.

$$\hat{Y} = 0.88 + 3.35(13)$$
$$= 0.88 + 43.55$$
$$\hat{Y} = 44.43.$$

One observes that when the educational attainment variable went up one year, the predicted value of income went up by 3.35, the slope value. This illustrates the general interpretation of the slope: b indicates the expected change in Y for a unit change in X_1. The intercept also has a general interpretation: a indicates the EXPECTED VALUE of Y when $X_1 = 0$. When there are no X_1 values at 0, as is the case with these data, the intercept has a mathematical role (it is a constant that must be added to complete a prediction) but no substantive meaning.

The regression line can be used for predictions, but they are not perfect, as the points distant from the line demonstrate. How well the linear model actually fits the data is measured with the R-SQUARED (R^2), which assesses the variation in the dependent variable accounted for by the independent variable. It is a summary statistic ranging from 1.0 to .00. In the example, $R^2 = 0.76$, suggesting parent educational background accounts for 76% of the variation in student income.

MULTIPLE REGRESSION

Analysis with multiple regression extends the foregoing, by allowing for multiple independent variables. This is an important extension, for two reasons. First, a more complete explanation of a phenomenon becomes possible, because multiple causes, predictors, or determinants can be entertained simultaneously. Second, the specific effect of a particular X on Y can be better understood, because the influence of other, perhaps CONFOUNDING, variables can be removed through statistical controlling. Consider the example. Other factors besides parent education undoubtedly help shape student income. A multiple regression model is in order, to take these other variables into account. There is a measure on at least one of these other variables—gender. Therefore, the revised theory is that student income (Y) is influenced by parent education (X_1) and gender (X_2). Here are the OLS estimates for this multiple regression model:

$$\hat{Y} = 0.44 + 3.12X_1 + 6.9X_2.$$

On the basis of these results, the interpretation is somewhat altered. The impact of an additional year of parent education appears slightly diminished, compared with the earlier bivariate result ($3.12 < 3.35$). This reduction in the slope coefficient can be attributed to statistically "holding constant" X_2. Furthermore, gender itself seems to have an independent effect on income. Specifically, males can expect to earn about $7,000 more than females, once parent education differences are controlled for. Overall, the $R^2 = 0.82$, indicating that these two independent variables together account for 82% of the variation in income. This is a boost in the R^2 of six percentage points, moving from the bivariate to the multivariate model (a gain that drops a bit with the ADJUSTED R-SQUARED).

ASSUMPTIONS

Regression analysis of nonexperimental social science data aims to reveal structural links—causal, explanatory, predictive—between variables in the real world. More formally, the goal is to estimate POPULATION parameters, the fixed intercept and slope values that reliably join the variables under study. These estimates are accurate to the extent that key assumptions are met. If the assumptions are ignored, regression results may have no reality beyond the numbers typed on a page. Granting the Table 1 data, do the regression estimates presented reveal the actual connection between student income, parent education, and gender? Yes, to the extent that they are supported by the assumptions presented below.

Most regression analysis is carried out on a SAMPLE, rather than on the target population as a whole. With estimation of the sample equation, say, $Y = a + b_1X_1 + b_2X_2 + e$, inferences are made to the

parameters in the population equation, say, $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. The sample must be a scientific PROBABILITY SAMPLING of sufficient size. Furthermore, because the samples are only samples, SIGNIFICANCE TESTING must be applied, to rule out chance results. For example, with the sample of 35 students out of the population of 350, we observe a slope coefficient of 3.12 in the multiple regression equation. A significance test (.05, two-tailed) applied to that regression coefficient allows us to reject the NULL HYPOTHESIS that the link between parent education and income is zero in the population, that is, $\beta_1 = 0$.

Besides assumptions of proper sampling and significance testing, there are the classical linear multiple regression assumptions, which can be variously stated. Broadly, these include no SPECIFICATION error, no measurement error, no perfect COLLINEARITY, and no error term problems. With respect to specification, the essential ideas are that the model has the right independent variables, and their relationship to the dependent variable is linear. With respect to measurement, the variables should be quantitative and accurately scored. With respect to collinearity, no independent variable is allowed to be a perfect linear function of all the others. With respect to the error term, it should not be related to the independent variables either as variance (to preserve HOMOSKEDASTICITY) or as a variable (because the data are nonexperimental), and it should not be related to itself (to avoid AUTOCORRELATION). When these assumptions are met, the OLS estimator is BLUE, or BEST LINEAR UNBIASED ESTIMATOR. Certain analysts prefer to think of the classical linear regression assumptions as lying along a continuum, say, from 1 to 10, where 1 means *not met at all* and 10 means *perfectly met*. In that view, the closer the results to 10, the closer the inferences are to reality.

OLD DIRECTIONS AND NEW

Sir Francis Galton (1822–1911) can perhaps be considered the founder of regression analysis, because he was the first to perceive that a straight line could make sense of the whirl of points in a scatterplot. He and his disciple Karl Pearson (1857–1936) made extensive investigations of scatterplots relating the characteristics of fathers and sons. For example, Pearson examined the heights of 1,078 father-and-son combinations and found that very tall fathers tended to have shorter sons. This was "to regress toward the mean," or what Galton called "regression to mediocrity." The straight line that depicted this regression effect was named the "regression line." Modern regression and CORRELATION methods were launched by the papers published on genetics in the journal *Biometrika*, particularly the work of Pearson coming out in 1903.

OLS has been the central method of regression estimation. The least squares method was discovered independently by French mathematician Adrien Marie Legendre (1752–1833) and German mathematician Carl Friedrich Gauss (1777–1855). When the classical multiple regression assumptions are met, OLS estimators cannot be improved upon. A great deal of work, especially since the 1960s, has gone into REGRESSION DIAGNOSTICS, which attempt to test assumptions and provide corrections. However, certain models are intrinsically nonlinear, and their parameters cannot be efficiently estimated with OLS. A case in point is when the dependent variable is dichotomous, say, a "yes" versus "no" vote choice. Here, a MAXIMUM LIKELIHOOD ESTIMATION technique, such as LOGISTIC REGRESSION, is generally preferred. Logistic regression approaches, including POLYTOMOUS ones, have recently been a vigorous area of research. Although these approaches are, in some ways, very far from OLS, the logic of the modeling, and many of the problems of inference due to violation of assumptions, is the same.

—Michael S. Lewis-Beck

REFERENCES

Kmenta, J. (1997). *Elements of econometrics* (2nd ed.). Ann Arbor: University of Michigan Press.

Lewis-Beck, M. S. (1980). *Applied regression: An introduction.* Beverly Hills, CA: Sage.

Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. W. (1996). *Applied linear regression models.* New York: Irwin.

REGRESSION COEFFICIENT

The term *REGRESSION* refers to a group of popular methodologies that aims to explain variation in a DEPENDENT VARIABLE (Y) using a set of k INDEPENDENT VARIABLES (X_k). In the context of ORDINARY LEAST SQUARES (OLS), a regression coefficient ($\hat{\beta}_k$) is associated with each independent variable ($X_1, X_2, \ldots, X_k$) and gives the expected direct effect of a one-unit change in that independent variable on the dependent variable, holding all other independent variables constant. The regression coefficient is