in-variable models, and item-response theory illustrate models that incorporate latent variables.

The basic statistical concept of latent variables analysis is simple. These variables refer to an abstract level of analysis that cannot be directly observed and measured. In order to estimate the numerical values of the parameters from empirical data, we must use observable indicators to link the unobservable conceptual variables. An example of a formative model, the measurement model for socioeconomic status (SES), may make clearer this distinction between conceptually abstract and observable levels of analysis. A researcher may observe the variables income, educational level, and neighborhood as indicators (manifest variables) of SES (latent variable). Latent variable models provide a means to parse out measurement errors by combining across observed variables (using correlations among variables), and they allow for the estimation of complex causal models. Those measurement errors may include faulty respondent memory or systematic errors made in the survey process.

Latent variable analysis is parallel to factor analysis. In modern test-theory models, the relation between the latent variable and the observed score (item responses) is mathematically explicit. The form for the relation is a generalized regression function of the observed scores on the latent variable. This regression function may differ in form—a linear pattern for the factor models and a logistic one for the probabilistic models (Mellenbergh 1994). Researchers should decide whether to treat the underlying latent variable(s) as continuous or discrete. Further discussion can be found in Tom Heinen's demonstrations (1996).

In psychological studies, researchers may adopt a reflective model rather than a formative model because it is the standard conceptualization of measurement in psychology. This model specifies a pattern of covariation between the indicators, which can be fully explained by a regression on the latent variable. That is, the indicators are independent after conditioning on the latent variable (this is the assumption of local independence). An example of a reflective model in the latent variable of depression may use item responses on items like, "I am sad all the time," "I often feel helpless," and "I often feel my life is empty." In the reflective model of depression, it implies that a depressed person will be more inclined to answer the question affirmatively than a mentally healthy person. In ordinary language interpretation, depression comes first and "leads to" the item responses. In the mathematical term, it implies a regression of the indicators on the latent variable, while in the SES model (a formative model), the relationship between indicators and the latent variable is reversed. In other words, variation in the SES indicators now precedes variation in the latent variable; SES changes as a result of an increase in income and/or education and not the other way around.

In sum, latent variable theory signifies both realism and constructivism. Latent variables of the formative model are more a summary of the observed variables, while a reflective model implies entity realism about the latent variable. A causal implication between observable indicators and the latent variable thus is not a strong assumption. It is suggested that researchers be cautious when interpreting the relation in empirical studies (Borsboom et al. 2003).

SEE ALSO *Depression, Psychological; Factor Analysis; Realism; Regression Analysis; Social Science; Sociology; Structural Equation Models; Variables, Random*

BIBLIOGRAPHY

Borsboom, Denny, Gideon J. Mellenbergh, and Jaap van Heerden. 2003. The Theoretical Status of Latent Variables. *Psychological Review* 110 (2): 203–219.

Heinen, Tom. 1996. *Latent Class and Discrete Latent Trait Models: Similarities and Differences.* Thousand Oaks, CA: Sage.

Mellenbergh, Gideon J. 1994. Generalized Linear Item Response Theory. *Psychological Bulletin* 115: 300–307.

Cheng-Hsien Lin

VARIABLES, PREDETERMINED

This entry explains when a variable is a predetermined variable and how identification and inference require a variable to be predetermined. In social science, researchers often try to explain a phenomenon or event using one or more explanatory variables. For example, how much an individual earns can be explained (to some degree) by his or her education level, and how much an individual consumes can be explained by his or her income and wealth. In many cases, a social scientist will formulate a model in which one variable is a function of another variable. For example, the following is a model that relates consumption to income and wealth:

$$\text{Consumption} = c_0 + c_1 \cdot \text{income} + c_2 \cdot \text{wealth}$$

where c_0, c_1, and c_2 are numbers. For example, $c_0 = \$10,000$, $c_1 = 0.7$, and $c_2 = 0.05$. This model implies that a one-dollar increase of income causes consumption to increase by $\$c_1$ (that is, consumption increases by seventy cents if income increases by one dollar). In order to estimate this model, we need to extend the model with an error term. This error term captures variables other than income or wealth. Let

$$\text{Consumption} = c_0 + c_1 \cdot \text{income} + c_2 \cdot \text{wealth} + \varepsilon$$

where the error term ε is assumed to be uncorrelated with income and wealth. If ε is assumed to be uncorrelated with income and wealth, then income and wealth are exogenous variables. No correlation means that we cannot use the regressors to predict the error term, that is, $E(\varepsilon|\text{income, wealth}) = 0$. If all the explanatory variables are exogenous variables, then the coefficients can be given a causal interpretation. Suppose that a social science researcher does not have access to data on wealth and, therefore, estimates the model

$$\text{Consumption} = d_0 + d_1 \cdot \text{income} + u.$$

Note that the new error term u consists of the old error term ε plus $c_2 \cdot$ wealth. Wealth and income are correlated so that income is correlated with u. Therefore, we cannot give a causal interpretation to d_1. In particular, an estimate of d_1 is likely to overstate the effect of income on consumption. Suppose that we have data on consumption and income for N individuals, {Consumption$_i$, income$_i$} where $i = 1, \ldots, N$. Consider the least squares estimator for d_1. This estimator minimizes $\Sigma_i(\text{Consumption}_i - d_0 - d_1 \cdot \text{income}_i)^2$ with respect to d_0 and d_1. The least squares estimator for d_1 has the following form,

$$\begin{aligned}\hat{d}_1 &= \frac{\Sigma_i\,(\text{income}_i - \overline{\text{income}}) \cdot \text{Consumption}_i}{\Sigma_i\,(\text{income}_i - \overline{\text{income}})^2}\\ &= d_1 + \frac{\Sigma_i\,(\text{income}_i - \overline{\text{income}}) \cdot \mu_i}{\Sigma_i\,(\text{income}_i - \overline{\text{income}})^2}\end{aligned}$$

where $\overline{\text{income}}$ denotes the mean of income, $\overline{\text{income}} = \frac{1}{N}\sum_i$. If income and the error term u are uncorrelated, then (1) the expectation of the last term, $\frac{\sum_i(\text{income}_i - \overline{\text{income}})\cdot u_i}{\sum_i(\text{income}_i - \overline{\text{income}})^2}$, is zero so that $E(\hat{d}_1) = d_1$, and (2) this last term is very small for large N (the technical term is that $\frac{\sum_i(\text{income}_i - \overline{\text{income}})\cdot u_i}{\sum_i(\text{income}_i - \overline{\text{income}})^2}$ converges in probability to zero so that $\hat{d}_1$ is a consistent estimator for d_1). However, in this example, the error term u_i depends on wealth. Wealth and income are correlated so that the assumption exogeneity (i.e., that all regressors are uncorrelated with the error term) is violated. As a result, the estimate for d_1 cannot be given a causal interpretation. In particular, the expectation of the estimator, $E\hat{d}_1$, will be larger than 0.7 because of the positive correlation between income and wealth.

The exogeneity assumption is very strong and can be relaxed somewhat. Consider the following model that describes the squared daily return of a stockmarket (e.g., the daily return of the Standard & Poor's 500 index),

$$\text{Squared Return}_t = f_0 + f_1 \cdot \text{Squared Return}_{t-1}.$$

As before, this model can be extended to include an error term,

$$\text{Squared Return}_t = f_0 + f_1 \cdot \text{Squared Return}_{t-1} + v_t.$$

Suppose there are T data points so that $t = 1, 2, \ldots, T$. Rather than assuming that the correlation between the squared return and the error term is zero, that is, that $E(v_t|\text{Squared Return}_1, \ldots, \text{Squared Return}_T) = 0$ for all t, we now make the weaker assumption that, given the past values of the squared return, the expectation of the error term is zero, that is, $E(v_t|\text{Squared Return}_1, \ldots, \text{Squared Return}_{t-1}) = 0$ for all t. Note that the past values of the squared return for error term v_t consist of the squared return of the first period, Squared Return$_1$, through period t – 1, Squared Return$_{t-1}$. Regressors that have the property that the error term has zero expectation given past values of the regressor are called *predetermined regressors* or *predetermined variables*. Consider the least squares regressor again to see how predeterminedness helps the estimator,

$$\begin{aligned}\hat{f}_1 &= \frac{\Sigma_t\,(\text{Squared Return}_{t-1} - \overline{\text{Squared Return}}) \cdot \text{Consumption}_t}{\Sigma_t\,(\text{Squared Return}_{t-1} - \overline{\text{Squared Return}})^2}\\ &= f_1 + \frac{\Sigma_t\,\text{Squared Return}_{t-1} \cdot \nu_t - \overline{\text{Squared Return}}\,\Sigma_t\,\nu_t}{\Sigma_t\,(\text{Squared Return}_{t-1} - \overline{\text{Squared Return}})^2}\\ &= f_1 + \frac{T}{\Sigma_t\,(\text{Squared Return}_{t-1} - \overline{\text{Squared Return}})^2}\ \frac{\Sigma_t\,\text{Squared Return}_{t-1} \cdot \nu_t}{T}\\ &- \frac{T}{\Sigma_t\,(\text{Squared Return}_{t-1} - \overline{\text{Squared Return}})^2}\ \frac{\overline{\text{Squared Return}}\,\Sigma_t\,\nu_t}{T}.\end{aligned}$$

The term $\frac{\sum_t \text{Squared Return}_{t-1}\cdot v_t}{T}$ is zero in expectation since $E(v_t|\text{Squared Return}_1, \ldots, \text{Squared Return}_{t-1}) = 0$. Moreover, for large T, this term, as well as $\frac{\overline{\text{Squared Return}}\sum_t v^t}{T}$, will be small so that the estimate $\hat{f}_1$ is close to the true value f_1. This model of squared returns is an ARCH (auto regressive conditional heteroscedasticity) model and can be used to study volatility. In particular, a large decline of the stockmarket in period $t - 1$ means that the stockmarket will be more volatile in period t. Tim Bollerslev, Robert Engle, and Daniel Nelson (1994) discuss other ARCH models.

An endogenous regressor has the property that $E(v_t|\text{Squared Return}_1, \ldots, \text{Squared Return}_{t-1}) \neq 0$. Thus, an endogenous regressor cannot be a predetermined regressor. Endogeneity (i.e., having an endogenous regressor) occurs if there is a third unobserved variable that affects both the regressor and the error term. For example, how much an individual earns can be partly explained by his or her education. Data on earnings and education levels are not hard to collect, but reliable data on intelligence are difficult to obtain. For this reason, earnings are usually regressed on the education so that intelligence is part of the error term. However, intelligence will also affect education levels so that the regressor education and the error term are correlated. In other words, there is an unobserved variable that affects both the regressor and the error term so that $E(v_t|\text{Squared Return}_1, \ldots, \text{Squared Return}_{t-1}) \neq 0$. Therefore, least squares cannot be used to estimate the effect of education on income. Econometricians have developed another technique, namely, two-stage least squares.

In nonlinear models, a slightly different definition of exogeneity and predeterminedness is sometimes used. In particular, the regressors are exogenous if the regressors and the error term are statistically independently distributed. That is, if the density of the error term conditional on the regressors, $p(\text{error term}|\text{regressors})$ is the same as the unconditional density of the error term, $p(\text{error term}|\text{regressors})$. Similarly, the regressors are predetermined if the density of the error term of period t conditional on the past regressors, $p(\text{error term}_t|\text{regressors}_1, \ldots, \text{regressor}_{t-1})$, is the same as the unconditional density of the error term, $p(\text{error term}_t)$ for all t. Robert De Jong and Tiemen Woutersen (2006) use these definitions when they estimate a model to predict monetary policy.

SEE ALSO *Autoregressive Models; Causality; Econometric Decomposition; Identification Problem; Probability; Regression; Regression Analysis; Statistics*

BIBLIOGRAPHY

Bollerslev, Tim, Robert F. Engle, and Daniel B. Nelson. 1994. ARCH Models. In *Handbook of Econometrics*, Vol. 4, ed. Robert. F. Engle and Daniel McFadden, 2961–3031. Amsterdam: North Holland.

De Jong, Robert, and Tiemen Woutersen. 2006. Dynamic Time Series Binary Choice. Working Paper. Baltimore: MD: Johns Hopkins University.

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

Stock, James H., and Mark W. Watson. 2006. *Introduction to Econometrics*. 2nd ed. Boston: Addison-Wesley.

Tiemen Woutersen

VEBLEN, THORSTEIN
1857–1929

Thorstein Bunde Veblen, an economist and sociologist (social critic and social and cultural theorist), was born to a Norwegian immigrant couple and grew up in rural Minnesota. He attended Yale University for graduate work in philosophy, where he met the sociologist William Graham Sumner (1840–1910). Upon graduation Veblen was not able to find academic employment. He eventually went to Cornell University to study economics, then taught at the University of Chicago (1891), later moving to Stanford University (1906), the University of Missouri (1911), and the New School for Social Research (1919). Veblen's troubles with university administrations stemmed from his disregard for the norms of dress for "proper professors," his uncommon living conditions (he lived in a shack of his own construction at one point), his classroom presentations (he often spoke softly in monotone or displayed unorthodox behavior), and his unconcealed extramarital affairs. At one point in his career he taught a class entirely in the Icelandic language to make the point that modern education was useless. Veblen saw himself as outside both Norwegian and American cultures and specifically asked those who knew him not to write his biography after his death.

Veblen posited certain human instinctual drives (mediated by cultural norms) that allow for technological and social advance, social organization, and social evolution: the instinct of workmanship, which is the most productive instinct for well-being, being an underlying creative impulse to manipulate the world with productive labor; the instinct of parenting, which leads to a concern for the well-being of others and an identification with community; and the instinct of idle curiosity, which leads to the development of knowledge. His use of the word *instinct* does not correspond to standard understandings from biology. Rather, he used *instincts* as socially refracted modifications of desire (Veblen 1914).

SOCIETY

Influenced by Darwin's theory of evolution, Veblen was interested in the historical and evolutionary development of society and argued that humans interpret the world using categories based in biographic and historically shaped "habits of the mind," which in turn are the basis for cultural norms passed on through socialization. Activities formed around these norms Veblen called "institutions," with changes in productive activity leading to changes in society (Veblen 1914).

Veblen formulated a scale of three evolutionary stages of society based on changes in material forms of production: "savagery" (a peaceable, isolated, and stable society); "barbarianism" (a warlike and conquest-oriented society, hierarchical and dominated by religion, with distinct predatory and industrious classes and a surplus of wealth); and "civilization" (a modern, economically developed society that is rational and instrumental, with machine technology, mass production, and a high division of labor). For Veblen, the business class in modern society is "predatory" in that its livelihood is based on the acquisition of personal wealth and competitive capitalist profit