Glossary for econometrics and epidemiology

F Imlach Gunasekara, K Carter and T Blakely

J Epidemiol Community Health 2008;62;858-861
doi:10.1136/jech.2008.077461

Glossary for econometrics and epidemiology

F Imlach Gunasekara, K Carter, T Blakely

Department of Public Health, Wellington School of Medicine and Health Sciences, University of Otago, Wellington, New Zealand

Correspondence to: Dr F Imlach Gunasekara, Department of Public Health, Wellington School of Medicine and Health Sciences, University of Otago, PO Box 7343, Wellington South 6242, New Zealand; Fiona.gunasekara@otago.ac.nz

Accepted 16 June 2008

ABSTRACT

Epidemiologists and econometricians are often interested in similar topics—socioeconomic position and health outcomes—but the different languages that epidemiologists and economists use to interpret and discuss their results can create a barrier to mutual communication. This glossary defines key terms used in econometrics and epidemiology to assist in bridging this gap.

Econometrics is the application of statistical methods to observed data for the purposes of empirical research, forecasting or testing of economic theory.[1] The disciplines of epidemiology and economics share much common ground. Both are interested in the impacts of socioeconomic position (such as income and education) and fiscal policy on people's lives. Epidemiology, as the study of disease patterns in human populations, will often seek socioeconomic causes for disease, which may be remedied and assist in disease prevention or mitigation. Economists frequently include health indicators in their analyses,[2–4] recognising the importance of health for an individual's welfare and for economic development.

As epidemiologists move more towards causal modelling, the econometrics literature is a source of knowledge because econometricians have been doing this for years, are aware of potential bias from unmeasured confounders and have advanced methods for tackling endogeneity problems, especially with longitudinal data.

To date there has been relatively little interdisciplinary dialogue between econometricians and epidemiologists, resulting in the emergence of methodological and linguistic rifts that prove difficult to cross for those unfamiliar with the other discipline. Both camps are concerned with similar concepts and associations but describe them with different terminology. Unless we attempt to bridge the gap between epidemiology and economics, we miss out on a rich resource of information that is within our reach. We also risk duplicating research that has already been done or is being done in parallel, which is wasteful or even unethical.

Some of the distance between the disciplines might be because epidemiologists often conceive of theories and relationships visually, using diagrams and tools such as causal diagrams,[5 6] whereas economists rely more heavily on written equations and statistical language as entry points (eg defining types of bias in terms of correlations with error terms). Consider the outcome variable, y, sometimes called the "left-hand side variable" by economists and statisticians because written regression equations are the normal way of conceptualising relationships between variables. In contrast, epidemiologists draw causal diagrams with the y variable on the right-hand side because it is practice for causal connections to be drawn from left to right. Also the language of econometrics can be novel to an epidemiologist—for example, what exactly is the definition of endogeneity in its various contexts and what is the difference between omitted variable bias, unmeasured confounding and unobserved heterogeneity? Not an issue exclusive to just epidemiology, biostatistics and econometrics, multiple different terms are often used for a single concept. For example, other terms used for exposure variable include independent variable, treatment, indicator variable, right-hand variable, predictor or predictive variable, control variable, determinant, explanatory variable, covariate and regressor. Similarly, synonyms for outcome variable include dependent variable, response variable, left-hand variable, responding variable, explained variable, endpoint and regressand.

This glossary introduces and defines key terms used within econometrics, and terms from both econometrics and epidemiology that have common ground. When the glossary term is a primarily an econometric term, we first give a definition from commonly used econometrics textbooks by Wooldridge.[1 7] Where multiple synonyms for the same concept exist (in either discipline), these have been noted in the glossary, with the reader referred to the definition given under one terminology. Most examples are discussed in terms of simple ordinary least squares (OLS) models. Our motivation for writing this glossary began with the development of a longitudinal study (or panel study in econometrics), with a particular interest in the dynamic association of income and health over time.

COLLINEARITY

See *multicollinearity*.

CONFOUNDING

A confounder, a term commonly used in epidemiology, is associated with an exposure or risk factor for the outcome and with the outcome independent of the exposure, but is not on the causal pathway between the exposure and the outcome.[8]

Confounding occurs when the relationship between an exposure variable and the outcome variable is contaminated, so that the measure of association between these two variables is actually also capturing the effect of a third variable, the confounding variable. For example, many factors are associated with income (exposure) and health (outcome), such as education, employment or

wealth—it may be that these factors explain part of the observed association.

Confounding is not a term often used in econometric language (exceptions can occur in a health context, eg Zimmerman and Katon[9] refer to "some third variable(s)—possibly unobserved" as a possible cause of the association between low income and depression). However, when economists refer to '*omitted variables*', this is largely (but not exclusively; see later entry) referring to a subset of confounders, those problematic unmeasured confounders that are not included in the analysis. Confounders are also sometimes referred to generically in an econometric model as control variables.

ENDOGENEITY

Endogeneity arises when an explanatory variable is correlated with the *error term*. According to Wooldridge,[7] this may occur because of *omitted variable* bias (or unmeasured *confounding*), *simultaneity* (or *reverse causality*) or *measurement error*; however, endogeneity is sometimes used to refer only to *simultaneity*[10] and *measurement error* is usually treated separately, especially if it is assumed that this error is random (eg that people are as likely to underrate as overrate their health).

See *endogenous variable*.

ENDOGENOUS VARIABLE

> Endogenous explanatory variable: An explanatory variable in a multiple regression model that is correlated with the error term, either because of an *omitted variable*, *measurement error*, or *simultaneity*.[1]

An endogenous variable is one that is related to and determined by other variables also in the model. (The definition given above relates to exposure variables, but note that the dependent variable can also be an endogenous variable.) For example, a model looking at whether income influences health will be hampered by the fact that income is also determined in some way by the outcome variable health, indicating the presence of reverse causation (or *simultaneity*), that is, "income is endogenous to health".[11] A typical method in econometrics for dealing with endogenous explanatory variables is to use *instrumental variables*.[1] Hausman[12] proposed a test for endogeneity when results from an OLS model are significantly different from estimates of a two-stage least squares model.

In econometrics, the definition of an endogenous variable is more formal, where it is correlated with the *error term*.[1] The concept of correlation with the error term is not a building block for definitions in epidemiology. Consider the regression model of income as the exposure and health as the outcome, and education as an omitted (or confounding) variable. Education will be associated with the income variable and the error term in the simple income–health regression model, which is analogous to saying that education is associated with health independent of income, that is, one of the properties of a confounder.[8] Such correlation with the error term will also be true for a variable determined by the outcome.

When endogeneity is discussed in biostatistics texts with respect to longitudinal data it is mainly a factor of reverse causation.[13 14] An endogenous exposure variable is a predictor of the outcome at time *t* and is also predicted by the outcome at time *t*–1. This can be controlled for by adding time-lagged variables to the model.[14]

ERROR TERM

> Error term: The variable in a simple or multiple regression model that contains unobserved factors that affect the dependent variable. The error term may also include measurement errors in the observed dependent or independent variables.[1]

The error term in a regression equation accounts for the variation between the observed outcome and that predicted by the model. Several things may contribute to the error term, including excluded (or mis-measured) confounders; *omitted variables* (including confounders and mediators); *measurement error* of any of the variables in the model; and "true" randomness or chance.[7 15] Reference to the error term underpins much of econometrics, and understanding how variables relate to the error term is fundamental to understanding the language of econometricians. In longitudinal data analysis, the error term is decomposed into two components: a time-invariant term α_i and a time-varying idiosyncratic term u_{it}; idiosyncratic because it can change over time.[1]

EXOGENOUS VARIABLES

> Exogenous explanatory variable: An explanatory variable that is uncorrelated with the error term.[1]

Variables which are not determined by any of the other variables in the model are called "exogenous". Exogenous variables can influence and cause change in *endogenous variables* but they are not influenced by them. Any changes to an exogenous variable are due to forces outside of those in the model.[14]

Consider a model looking at whether income influences health which finds or assumes that "lottery wins are exogenous to income".[11] This means that lottery winnings influence income, but influence no other variables in the model (including, most importantly, health) other than via the income variable itself. If this is so, lottery winnings can be used, as an exogenous *instrumental variable* for income, to test the theory that income influences health. This is important because income itself is likely to be an *endogenous variable*, which introduces bias into the analysis of the income–health relationship.

HEALTH CAUSATION OR HEALTH SELECTION

See *simultaneity*.

INDIVIDUAL EFFECT OR INDIVIDUAL HETEROGENEITY

See *unobserved heterogeneity*.

INSTRUMENTAL VARIABLE

> Instrumental variable (IV): In an equation with an endogenous explanatory variable, an IV is a variable that does not appear in the equation, is uncorrelated with the error in the equation, and is (partially) correlated with the endogenous explanatory variable.[1]

Instrumental variables are frequently used in econometrics, particularly when use of other regression analytic techniques are considered flawed because the assumptions underlying these models have been violated. The idea behind instrumental variables (IVs) is that there is a perfect *exogenous variable* that is correlated with the endogenous exposure variable of interest (*x*), but which has no effect on the outcome variable other than

through *x*. To be useful, the instrumental variable must also have a reasonable effect size through *x*. The IV is used to remove the endogeneity from *x*, or extract the non-endogenous information from *x*, by assuming that the only path through which the IV affects the outcome is through *x* and scaling the relationship between the outcome and the IV by the relationship between *x* and the IV. In this way, instrumental variables can remedy *endogeneity* problems and are often used by economists to try to solve problems of *measurement error, omitted variable bias* and *simultaneity* and find the true relationship between the outcome and endogenous exposure variables.[16] However, in health research, it is difficult to find an instrument which is not associated either directly or indirectly with health status. In the example above (under *exogenous variables*), in which lottery wins may be used as an instrument for income on health, there is no certainty that winning the lottery by itself would not affect health (particularly mental health) status.

INTERMEDIATE OR INTERMEDIARY VARIABLES

In epidemiology, an intermediate variable is "any factor that represents a step in the causal chain between the exposure and disease [that] should not be treated as an extraneous confounding factor, but instead requires special treatment."[8]

This definition from Rothman and Greenland[8] explains that an intermediate factor is not a *confounder*, because the direction of association between an intermediate factor and the exposure is actually the opposite to that of a confounder. In causal terms, a confounder is expected to "cause" the exposure; but the exposure will "cause" the intermediate variable.

For example, if income is the exposure variable and health is the outcome, smoking could be postulated to lie on the causal pathway, or be an intermediate variable between income and health. People with more income are less likely to smoke. Therefore, smoking would not be included in the initial model, because some of the association between income and health would then be contained in the smoking variable coefficient. If we are interested in finding out how much of the association between income and health is mediated by smoking, then smoking would be included in an additional model and the change in the regression coefficient for income would be determined.[8] However, this method of determining direct and indirect effects is known to be potentially biased.[17–19]

Economists often discuss this differently, rarely using the terms intermediate or mediating variable, but might say that smoking is endogenous to income (another way of saying the exposure variable causes the intermediate variable) and would end up treating the analyses in much the same way as epidemiologists.

It is possible for a variable to act as a confounder and an intermediate variable. For example, it could be argued that net worth is a confounder of the income–health association (as being wealthy influences employment opportunities and hence one's salary) and that it is on the causal pathway from income to health. In this case, the problems with the analyses become more difficult,[17] requiring (in the absence of randomised trial data) longitudinal data with analyses such as marginal structural models.[20–22]

MEASUREMENT ERROR

> Measurement error: The difference between an observed variable and the variable that belongs in a multiple regression equation.[1]

Measurement error is the random or systematic error arising during data collection of variables. The measured variable *x* is measured with error ε, which is the distance of *x* from the "true" value of *x*. Measurement error can produce bias such as attenuating the estimators of exposure variables, but may also have more complex effects.[23] Longitudinal data analyses such as fixed effects models can actually augment the problem of measurement error in mis-measured explanatory variables that change little over time. *Instrumental variables* can be used by econometricians to address the problem of measurement error in exposure variables.

MEDIATING VARIABLES OR MEDIATION

See *intermediate or intermediary variables.*

MULTICOLLINEARITY

> Multicollinearity: A term that refers to correlation among the independent variables in a multiple regression model; it is usually invoked when some correlations are "large," but an actual magnitude test is not well defined.[1]

Collinearity occurs in multiple regression models when two (or more—multicollinearity) exposure variables are included in the model but are so similar to each other that they are essentially measuring at least part of the same thing. Collinearity creates problems with interpreting the analysis by affecting the standard errors of the variables (as it increases variance)[24] and by causing biased estimates for one or both of the collinear terms (sometimes dramatically so).[25] If there is perfect collinearity between two variables then one should be dropped from the model. However, dropping exposure variables to reduce multicollinearity may lead to bias in the model from dropping useful information.

It might be argued in some cases that it is appropriate to include (collinear) variables in a model even when they are highly correlated, because the theoretical basis for including the variables is strong and the results of the model are consistent with expectations.[10] In such cases, models with and without the variables would be tested and compared. Such consideration of alternative models tends to be described as an issue of (mis)-specification in econometrics. However, in epidemiology this tends to be described in terms of mediating and *intermediary variables*. Thus, an epidemiologist might argue for including an intermediate variable in a model (and testing it against a model without the variable) because including that variable might elucidate direct and indirect pathways between the exposure and outcome variables. An econometrician might do the same thing but describe this in terms of finding the correct model specification.

OMITTED VARIABLES AND OMITTED VARIABLE BIAS

> Omitted Variables: One or more variables, which we would like to control for, have been omitted in estimating a regression model.[1]
> Omitted Variable Bias: The bias that arises in the [ordinary least squares] estimators when a relevant variable is omitted from the regression.[1]

Omitted variables are important covariates that are excluded from the analysis, usually because data on these variables is unavailable or because the model is misspecified. In epidemiological language, omitted variables would mostly be unmeasured or unknown confounders.[26] However, it is critical to note that some known *intermediary variables* meet the definition of,

and are often considered as in practice, a type of omitted variable. Thus, omitted variables are not just unmeasured or unknown confounders. An omitted variable would cause bias only if it were correlated with an included exposure variable and the outcome variable.

REVERSE CAUSALITY (OR REVERSE CAUSATION)

See *simultaneity*.

SIMULTANEITY

The epidemiologic term that is equivalent to ''simultaneity'' is ''reverse causality''.

> Simultaneity: A term that means at least one explanatory variable in a multiple linear regression equation model is determined jointly with the dependent variable.[1]

A classic example used to demonstrate simultaneity is to consider the outcome variable (y) as a city's murder rate and the exposure variable (x) of interest as the size of the police force, then to realise that x is partly determined by y.[7] When x is partly determined by y, x is generally also correlated with the *error term*,[7] so simultaneity meets the definition of *endogeneity*. Where simultaneity exists, the outcome variable is said to be *endogenous* and OLS regression in this situation would lead to biased estimators.[10] In epidemiologic terms, simultaneity is synonymous with reverse causality. When health is the outcome variable of interest, this is known as health selection or health causation (and sometimes also as social selection or social drift). Economists classically take the position that health status has a greater influence on socioeconomic position than the reverse (ie health selection is the stronger pathway—the healthy become wealthy and get better jobs while the unhealthy become poorer and unemployed). Epidemiologists tend to argue that socioeconomic position is the main driver of health outcomes (ie improving income and education levels leads to improved health).

SOCIAL DRIFT OR SOCIAL SELECTION

See *simultaneity*.

UNOBSERVED EFFECT OR UNOBSERUED HETEROGENEITY

> Unobserved Effect: In a panel data model, an unobserved variable in the error term that does not change over time.[1]
> Heterogeneity bias: The bias in OLS due to omitted heterogeneity (or omitted variables).[1]

Unobserved heterogeneity, or unobserved effect, is the term used by economists to describe the unobserved individual characteristics of people that cannot be measured. For example, people can have things about them that make them more likely to be unhealthy and more likely to have a low income (or more likely to be healthy and wealthy), such as an undiagnosed chronic illness, ability, motivation or a pessimistic personality, which are difficult to measure or quantify. The exclusion of these means that any relationship that is found between the exposure variables and outcome variable may in fact be biased (''heterogeneity bias''), because these (omitted) variables may have their own varied effects on that relationship. In observational research, if this unobserved heterogeneity cannot be controlled for in some way, it will never be certain whether they may have biased the results.

Unobserved heterogeneity thus is a type of unmeasured *confounding*, and can lead to *omitted variable bias*.[1]

Unobserved heterogeneity is also known as the unobserved effect, a type of latent variable, and may be called the individual effect or individual heterogeneity if the unit of analysis is the individual person.[1]

Unobserved heterogeneity may be time-varying but, if it is assumed that it is constant over time, longitudinal data and fixed effects models can be used to control for between individual differences or heterogeneity by analysing only what happens to the same individuals over time.[13] If unobserved heterogeneity is thought to exist and to be correlated with the measured exposure variables then an OLS model can never give unbiased results.

Acknowledgements: Thanks to Steven Stillman, Ichiro Kawachi and Michael Hayward for comments on a final draft of this paper.

Competing interests: None.

REFERENCES

1. **Wooldridge JM.** *Introductory econometrics: a modern approach*. 3rd edn. Mason: Thomson South-Western, 2006.
2. **Deaton A.** Policy implications of the gradient of health and wealth. An economist asks, would redistributing income improve population health? *Health Aff (Millwood)* 2002;**21**:13–30.
3. **Contoyannis PJAM.** Socio-economic status, health and lifestyle. *J Health Econ* 2004;**23**:965–95.
4. **Smith JP.** Healthy bodies and thick wallets: the dual relation between health and economic status. *J Econ Perspect* 1999;**13**:145–66.
5. **Greenland S,** Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiology* 1999;**10**:37–48.
6. **VanderWeele TJ,** Robins JM. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *Am J Epidemiol* 2007;**166**:1096–1104.
7. **Wooldridge JM.** *Econometric analysis of cross section and panel data*. Cambridge (MA): MIT Press, 2002.
8. **Rothman KJ,** Greenland S, eds. *Modern epidemiology*. 2nd edn. Philadelphia: Lippincott-Raven, 1998.
9. **Zimmerman FJ,** Katon W. Socioeconomic status, depression disparities, and financial strain: what lies behind the income-depression relationship? *Health Econ* 2005;**14**:1197–215.
10. **Verbeek M.** *A guide to modern econometrics*. 2nd edn. Chichester: John Wiley & Sons, 2004.
11. **Ettner SL.** New evidence on the relationship between income and health. *J Health Econ* 1996;**15**:67–85.
12. **Hausman JA.** Specification tests in econometrics. *Econometrica* 1978;**46**:1251–71.
13. **Singer JD,** Willett JB. *Applied longitudinal data analysis*. Oxford: Oxford University Press, 2003.
14. **Diggle PJ,** Heagerty PJ, Liang K-Y, *et al*. *Analysis of longitudinal data*. Oxford: Oxford University Press, 2002.
15. **Upford G,** Cook I. *Oxford dictionary of statistics*. 2nd edn. Oxford: Oxford University Press, 2006.
16. **Angrist JD,** Krueger AB. Instrumental variables and the search for identification: from supply and demand to natural experiments. *J Econ Perspect* 2001;**15**:69–85.
17. **Robins JM,** Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992;**3**:143–55.
18. **Cole SR,** Hernan MA. Fallibility in estimating direct effects. *Int J Epidemiol* 2002;**31**:163–5.
19. **Blakely T.** Estimating direct and indirect effects: fallible in theory, but in the real world? *Int J Epidemiol* 2002;**31**:166–7.
20. **Robins JM,** Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;**11**:550–60.
21. **Bodnar LM,** Davidian M, Siega-Riz AM, *et al*. Marginal structural models for analyzing causal effects of time-dependent treatments: an application in perinatal epidemiology. *Am J Epidemiol* 2004;**159**:926–34.
22. **Howards PP,** Schisterman EF, Heagerty PJ. Potential confounding by exposure history and prior outcomes. An example from perinatal epidemiology. *Epidemiology* 2007;**18**:544–51.
23. **Bound J,** Brown C, Mathiowetz N, *et al*. Measurement error in survey data. In: Heckman J, Leamer EE, eds. *Handbook of econometrics*. Amsterdam: Elsevier, 2001:3705–843.
24. **Studenmund AH.** *Using econometrics. A practical guide*. 5th edn. Boston: Pearson Education, Inc., 2006.
25. **Kirkwood BR,** Sterne JAC. *Essential medical statistics*. 2nd edn. Malden (MA): Blackwell Science Ltd, 2003.
26. **Glymour MM.** Natural experiments and instrumental variable analyses in social epidemiology. In: Oakes JM, Kaufman JS, eds. *Methods in social epidemiology*. San Francisco: Jossey-Bass, 2006:429–60.