Working Paper Series
CSER WP No. 0005

Total, Direct, and Indirect Effects in Logit Models

Af Kristian Bernt Karlson, Anders Holm, and Richard Breen,

December 23, 2010

Centre for Strategic Educational Research
DPU, Aarhus University

Total, Direct, and Indirect Effects in Logit Models

Richard Breen*, Kristian Bernt Karlson**, and Anders Holm***
This version: December 23, 2010

Running head: Dealing with Selection Bias in Educational Transition Models

Abstract:
It has long been believed that the decomposition of the total effect of one variable on another into direct and indirect effects, while feasible in linear models, is not possible in non-linear probability models such as the logit and probit. In this paper we present a new and simple method that resolves this issue for single equation models and extends almost all the decomposition features of linear models to binary non-linear probability models such as the logit and probit. Drawing on the derivations in Karlson, Holm, and Breen (2011), we demonstrate that the method can also be used to decompose average partial effects, as defined by Wooldridge (2002). We present the method graphically and illustrate it using the National Educational Longitudinal Study of 1988.

* Center for Research on Inequality and the Life Course, Department of Sociology, Yale University, email: richard.breen@yale.edu.
** SFI – The Danish National Centre for Social Research and the Danish School of Education, Center for Research in Compulsory Schooling, Aarhus University, email: kbk@dpu.dk.
*** Center for Strategic Educational Research, Danish School of Education, Aarhus University, email: aholm@dpu.dk.

Total 6092 words (including all text, notes, and references) Running head: Total, Direct, and Indirect Effects in Logit Models5 tables

Total, Direct, and Indirect Effects in Logit Models

Introduction

Regression methods are indispensible tools for empirical sociological research. Many sociological applications compare regression coefficients of the same variable across models with different control variables. In linear models, the difference in these coefficients measures the extent to which the effect is mediated, confounded, or explained by the control variables. For example, stratification researchers may be interested in the extent to which racial differences in income are attributable to the uneven distribution of educational attainments across races. This kind of analysis was coined elaboration by Lazarsfeld (1955), and it is related to the well-known linear path analysis popularized by Duncan (1966; see also Bollen 1987). In linear models, the effect of an explanatory variable, *x*, on an outcome, *y*, may be decomposed into two parts, one mediated by a control variable, *z*, another unmediated by *z*. The part mediated by *z* is called the indirect effect, while the part unmediated by *z* is called the direct effect (cf. Alwin and Hauser 1975). The sum of the indirect and direct effects is called the total effect, equal to the effect of *x* on *y* when the control variable is omitted.

The decomposition into direct and indirect effects is a property of linear (OLS) models; total effects in logit and other non-linear binary probability models cannot be decomposed into the simple sum of direct and indirect effects (Fienberg 1977). Given a dichotomous dependent variable, *y*, the logit coefficient for *x* omitting the control

variable, *z*, will not equal the sum of the direct and indirect (via *z*) effects of *x* on *y*. This is because, in these models, the regression coefficients and the error variance are not separately identified; rather, the model returns coefficient estimates equal to the ratio of the true regression coefficient divided by a scale parameter that depends on the error standard deviation (e.g. Amemiya 1975; Winship and Mare 1983). Because the error variance differs across models these ratios do not decompose in the desired way.

In this paper we present a method that solves this problem and enables researchers to decompose total effects in logit models into the sum of direct and indirect effects.[1] The method is a by-product of the generalized approach developed by Karlson, Holm, and Breen (2011), and so it extends almost all of the decomposition features of linear models to logit models. It applies to all binary non-linear probability models, such as the logit, probit, complementary log-log and so on, as well as to cumulative probability models such as the ordered logit and ordered probit, and to the multinomial logit, although our exposition focuses on the binary logit model for the sake of clarity. We proceed as follows. First, we present the decomposition technique and show that the method may be used on average partial effects as defined by Wooldridge (2002). These effect measures provide researchers with more interpretable decompositions than those obtained with logit coefficients. Second, we give a statistical test for the indirect effect. Finally, we present examples to show how the method works for both continuous and discrete variables of interest.

Total, direct, and indirect effects in a logit model

[1] We use the word 'effect' in the sense in which it is commonly used in much social science: we do not discuss the assumptions that would be required to consider these effects causal (see Sobel 2008; VanderWeele 2010).

In this section we begin with a description and graphical illustration of total, direct, and indirect effects in a linear model, and then proceed to the binary logit model. Then we show how a total logit coefficient may be decomposed into its direct and indirect parts. Greater detail on the relevant mathematical derivations may be found in Karlson, Holm, and Breen (2011) on which we draw heavily. Our notation follows Blalock (1979).

The linear model

Let *y* be some continuous outcome of interest (e.g., respondent's income), let *x* be a continuous variable whose effect we want to decompose or "explain" (e.g., parent's income), and let *z* be a continuous control variable that potentially mediates the *x*-*y* relationship (e.g., respondent's educational attainment measured in years). We center all variables on their respective means and so we do not need to include intercepts in our models. Define the two following models:

$$y = \beta_{yx}x + e \tag{1}$$

$$y = \beta_{yx\cdot z}x + \beta_{yz\cdot x}z + v, \tag{2}$$

where β_{yx} is the gross effect of *x* on *y*, $\beta_{yx\cdot z}$ is the net effect of *x* on *y* given *z*, and $\beta_{yz\cdot x}$ is the net effect of *z* on *y* given *x*. *e* and *v* are random error terms. The difference between the beta-coefficients in the two models expresses the extent to which the *x*-*y* relationship is mediated, confounded, or explained by *z*:

$$\delta = \beta_{yx} - \beta_{yx\cdot z}. \tag{3}$$

The difference in (3) may also be expressed in other terms. Define the following linear model relating *x* to *z*:

$$z = \theta_{zx}x + w, \tag{4}$$

where θ_{zx} captures the effect of *x* on *z*, and *w* is a random error term. Using the properties of linear models and following Clogg, Petkova, and Haritou (1995a), it is easy to show that

$$\delta = \beta_{yx} - \beta_{yx \cdot z} = \theta_{zx} \cdot \beta_{yz \cdot x}. \quad (5)$$

We prove (5) in the Appendix.

Following Duncan (1966) we can decompose the total effect of *x* on *y* into a direct effect net of *z* and an indirect effect mediated by *z*:

Direct: $\beta_{yx \cdot z}$ (6a)

Indirect: $\theta_{zx} \cdot \beta_{yz \cdot x}$ (6b)

Total: $\beta_{yx} = \beta_{yx \cdot z} + \theta_{zx} \cdot \beta_{yz \cdot x}$. (6c)

The gross effect of *x* on *y* is the simple sum of the net effect of *x* on *y* given *z* and the product of the effect for *x* on *z* and the net effect of *z* on *y* given *x*. Readers may note that the expression of the total effect is equivalent to the omitted variable bias formula known from econometrics. The only difference is that in our example the control variable, *z*, is observed. Figure 1 illustrates the system defined by Equations (2) and (4).[2] We see that the indirect effect is the effect of *x* on *y* running through *z*, while the direct effect is the residual effect of *x* on *y* (net of *z*).

The binary logit model

The foregoing holds for linear (OLS) models and continuous outcomes. However, sociologists often work with discrete outcome variables. For example, educational researchers study binary educational decisions, demographers study death,

[2] Note that Figure 1 illustrates the system as fully recursive system in which *z* is an intervening variable. *z* may, however, also be placed "behind" *x* in the system or as a variable on the same recursive level as *x*. We use the illustration in Figure 1 because it depicts how the indirect effect via *z* is calculated.

organizational researchers study promotion, and political sociologists study voting behavior. Therefore sociologists often prefer to use non-linear probability models for discrete or categorical outcomes such as the logit model. However, as we will show, equation (5) generally does not hold for logit models.

Let $y*$ be a continuous latent variable representing the propensity of occurrence of some outcome (e.g., propensity to complete college), let x be an explanatory variable of interest (e.g., parental income), and let z be a control variable (e.g., respondent's academic ability). Readers unacquainted with the latent variable formulation of logit models may see $y*$ as a hypothetical notion or construct that we use for interpreting logit coefficients below.[3] We center these variables on their respective means to avoid including intercepts in the following models. We define an underlying, latent linear model in which the latent propensity is a function of x and z:

$$y* = \beta_{yx.z}x + \beta_{yz.x} + u \text{, where } sd(u) = \sigma_u \tag{8}$$

where e is a random term and σ_e is the residual standard deviation. The model in (8) corresponds to the model in (2), except that y^* unobserved and we therefore cannot estimate $\beta_{yx.z}$ and $\beta_{yz.x}$. However, we do observe a dichotomized version of $y*$, namely y, such that

$$\begin{aligned} &y = 1 \text{ if } y^* > \tau \\ &y = 0 \text{ if otherwise.} \end{aligned} \tag{9}$$

[3] However, although we motivated our exposition by presenting $y*$ as hypothetical this need not always be the case. The latent variable might exist but not be fully observed, as when we only know whether a person's income falls within a given range.

where τ is a threshold, normally set to zero.[4] The expected outcome of this binary indicator is the probability of choosing $y=1$, i.e., $E(y=1)=\Pr(y=1)$. We can rewrite the error term in (8) such that $u=\sigma_e\omega$, where ω is a standard logistic random variable, with mean zero and variance $\pi^2/3$ and σ_e is the scale parameter of the logistic distribution, yielding a variance of $\sigma_u^2=\sigma_e^2\pi^2/3$ for the error term in (10) (Amemiya 1975; Cramer 2003).[5] The scale parameter allows the variance of the error to differ from that of the standardized logistic distribution. Following the derivations in Karlson, Holm, and Breen (2011), we write the logit model corresponding to (8):

$$\Pr(y=1)=\frac{\exp(b_{yx.z}x+b_{yz.x}z)}{1+\exp(b_{yx.z}x+b_{yz.x}z)} \Leftrightarrow$$

$$\operatorname{logit}(\Pr(y=1))=b_{yx.z}x+b_{yz.x}z=\frac{\beta_{yx.z}}{\sigma_e}+\frac{\beta_{yz.x}}{\sigma_e} \qquad (10)$$

Equation (10) makes it clear that the logit coefficients (the *b*'s) are equal to the coefficients from the underlying linear model in (8) divided by the scale parameter of that same model:

$$b_{yx.z}=\frac{\beta_{yx.z}}{\sigma_e};b_{yz.x}=\frac{\beta_{yz.x}}{\sigma_e} \qquad (11)$$

Because logit coefficients are equal to the ratio between two inherently unknown parameters – the underlying coefficient and a function of the underlying residual standard deviation – they are said to be identified only up to scale (cf. Cameron and Heckman 1998). The expressions in (11) also make it clear why we cannot compare the

[4] The categorical formulation of the logit-model known from introductory text books (e.g., Hosmer and Lemeshow 1989) provides another way of interpreting the logit coefficients. However, both formulations return identical results. For a text book description of the two different formulations we refer to Powers and Xie (2000).

[5] Had we assumed ω to be a standard normal random variable and σ_e the scale parameter of the normal distribution, we would have obtained the probit model.

coefficient of *x* from a logit model excluding *z* with the corresponding coefficient from a logit model including *z*. Whenever *z* has an effect on *y* (i.e., $b_{yz\cdot x} \neq 0$), a model without *z* will have a larger residual standard deviation than a model with *z* because the latter model will explain more variation in the latent outcome. As a consequence we cannot separate the extent to which the change in the effect of *x* is due to confounding or to rescaling, as was noted by Winship and Mare (1984). Thus the equalities in (5) for linear models do not hold for logit models.

However, following the framework developed in Karlson, Holm, and Breen (2011), we may make a decomposition of a total logit coefficient into its direct and indirect parts. We define a linear model as in (4) relating *x* to *z*:

$$z = \theta_{zx} x + w. \tag{12}$$

and consider the logit model for the effect of *x* on *y* without controlling for *z*:

$$\text{logit}(\Pr(y=1)) = b_{yx} x = \frac{\beta_{yx} x}{\tilde{\sigma}_e} \tag{13a}$$

Which parallels the underlying linear model:

$$y^* = \beta_{yx} x + \tilde{\sigma}_e \omega \tag{13b}$$

with ω having a standard logistic distribution. The coefficient in (13b) tells us the total effect of *x* on *y**, as we noted, decomposes into a direct and indirect effect. Now we define the following effects for the logit model,

Direct: $$b_{yx.z} = \frac{\beta_{yx.z}}{\sigma_e} \tag{14a}$$

Indirect: $$b_{yz.x} \cdot \theta_{zx} = \frac{\beta_{yz.x}\theta_{zx}}{\sigma_e} \tag{14b}$$

Total: $$b_{yx.z} + b_{yz.x} \cdot \theta_{zx} = \frac{\beta_{yx.z} + \beta_{yz.x}\theta_{zx}}{\sigma_e} \tag{14c}$$

The Appendix proves that (14b) holds. (14a) is a logit coefficient representing the effect of *x* on *y* controlled for *z* and (14b) is a logit coefficient representing the effect of *x* on *y* that is mediated by *z* given by the product of the linear regression coefficient of *x* on *z* and the logit coefficient relating *z* to *y* net of *x*. These quantities can be obtained from equations (10) and (12). (14c) is a logit coefficient representing the sum of the direct (14a) and indirect (14b) effects but this is not equal to the logit coefficient in (13a) because there the true coefficient, β_{yx}, is scaled by $\tilde{\sigma}_e$ whereas the direct and indirect effects are scaled by σ_e. To make the total effect and the direct and indirect effects compatible, we rescale the total effect by σ_e instead of $\tilde{\sigma}_e$. Borrowing the terminology of Clogg, Petkova, and Haritou (1995a), we say that we rescale by the scale factor from the full model defined by (8) and (10), which we consider the 'true' model in this case.[6]

Karlson, Holm and Breen (2011) show that this rescaling can be accomplished by replacing (13a) with the following logit model:

$$\text{logit}(\Pr(y=1)) = b_{yx.\tilde{z}}x + b_{y\tilde{z}.x}\tilde{z} \qquad (15)$$

Here, $\tilde{z}$ is the residualized *z*, that is, the residual from a linear regression of *z* on *x* (that is, $\tilde{z} = w$). By construction, $\tilde{z}$ is orthogonal to *x*. Karlson, Holm and Breen (2011) prove that

$$b_{yx.\tilde{z}} = \frac{\beta_{yx}}{\sigma_e} \qquad (16)$$

[6] By "true model" we do not refer to some deeper philosophical meaning of this word, but simply define it as the model we base our inferences on. See Karlson, Holm, and Breen (2011) for a formalized exposition.

and it follows at once that (16) is equal to (14c), that is, the sum of the direct and indirect effects.[7]

It is important to notice that because these are logit coefficients we do not recover estimates of the effects from the underlying linear model: rather we have decomposed the logit coefficients into direct and indirect effects defined up to scale—the scale, in this case, being that of the 'true' underlying model. However, sometimes researchers may want to assess the relative magnitude of the direct and indirect effects relative to the total effect. For example, one strand of social mobility research is interested in the extent to which the association between social class (*x*) and a binary educational decision (*y*) is mediated by academic performance (*z*) (Erikson et al. 2005; Karlson and Holm 2011). In the framework developed by Boudon (1974), the part mediated by academic skills is called the primary effect, while the unmediated part is called the secondary effect, thereby giving the decomposition a theoretical interpretation. For this kind of decomposition we suggest the following percentage decomposition:

$$\frac{b_{yz.x}\cdot\theta_{zx}}{b_{yx.z}+b_{yz.x}\cdot\theta_{zx}}\cdot 100\% = \frac{\frac{\beta_{yz.x}\cdot\theta_{zx}}{\sigma_e}}{\frac{\beta_{yx.z}+\beta_{yz.x}\cdot\theta_{zx}}{\sigma_e}}\cdot 100\% = \frac{\beta_{yz.x}\cdot\theta_{zx}}{\beta_{yx.z}+\beta_{yz.x}\cdot\theta_{zx}}\cdot 100\%, \qquad (17)$$

which expresses the extent to which the *x*-*y**-relationship in a logit model is mediated, confounded, or "explained" by *z*. Because the direct and indirect effects sum to the total effect, it holds that the part not mediated by *z*, i.e., the direct part, is defined as: Direct =

[7] It follows because, as we already showed, $\beta_{yx} = \beta_{yx.z} + \theta_{zx}\beta_{yz.x}$ and so $\frac{\beta_{yx} = \beta_{yx.z} + \theta_{zx}\beta_{yx.z}}{\sigma_e}$ and therefore $b_{yx} = b_{yz.x} + \theta_{zx}b_{yz.x}$.

100%-Indirect. Note also that (17) does not involve a scaling parameter, and therefore expresses the relationship between the coefficients from the underlying linear models: in other words, it is a scale-free measure. We refer to Karlson, Holm, and Breen (2011) for other measures that assess the relative contributions of direct and indirect effects.

Extensions

So far we have provided a simple decomposition of a total logit coefficient into its direct and indirect parts and provided a simple percentage measure with which researchers may assess the relative magnitude of direct and indirect effects. Thus far we have considered only one control variable, but in some instances we may want to consider several indirect paths by which *x* affects *y*. Because the method developed by Karlson, Holm, and Breen (2011) extends almost all decomposition features of linear models to logit models, it is straightforward to replace a single z with a vector of control variables, z_j, where $j = 1, 2, \ldots, J$, and where J denotes the total number of variables in z_j. Now we may define an underlying linear model including z_j as

$$y^* = \beta_{yx.z1,\ldots zJ}x + \sum_j \beta_{yz(j).x} z_j + t, \text{ with } sd(t) = \sigma_t \text{ and } \sigma_t = \sigma_k \cdot (\pi / \sqrt{3}) \quad (18a)$$

And the corresponding logit:

$$\text{logit}(\Pr(y=1)) = b_{yx.z1,\ldots,zJ}x + \sum b_{yz(j).z} z_j = \frac{\beta_{yx.z1,\ldots,zJ}}{\sigma_k} x + \sum_j \frac{\beta_{yz(j).x}}{\sigma_k} z_j \quad (18b)$$

Similar to (12), we estimate J linear regression models

$$z_j = \theta_{z(j)x} + w_j \quad (19)$$

which provide us with J coefficients of the effect of x on each control variable. Each indirect effect is given by:

Indirect: $$b_{yz(j).x} \cdot \theta_{z(j)x} = \frac{\beta_{yz(j).x} \cdot \theta_{z(j)x}}{\sigma_k} \quad (20a)$$

And the total effect by

Total: $$b_{yx.z1,\ldots,zJ} + \sum_j b_{yz_j.x}\theta_{z_j x} = \frac{\beta_{yx.z1,\ldots,zJ}}{\sigma_k} + \sum_j \frac{\beta_{yz(j).x} \cdot \theta_{z(j)x}}{\sigma_k} = b_{yx.\tilde{z}1,\ldots,\tilde{z}J}\,. \quad (20b)$$

That $b_{yx.\tilde{z}1,\ldots,\tilde{z}J} = \frac{\beta_{yx.\tilde{z}1,\ldots,\tilde{z}J}}{\sigma_k}$ follows by analogy with (16). The full proof can be found in Karlson, Holm and Breen (2011).

In some situations researchers will be interested in controlling the decomposition of the *x*-*y**-relationship for potentially confounding variables. Following Sobel (1998) we name these variables *concomitants*. These variables allow controlling for confounding influence *on the decomposition*, that is, on the estimates of direct and indirect effects. Such control is unaffected by the scale identification of logit coefficients, because our decomposition method assures that the total, direct, and indirect effects are measured on the same scale. Let w_i denote the i'th concomitant, $i = 1, 2, \ldots, I$, where I denotes the number of concomitants. We may control for the potential confounding influence of these concomitants *on the decomposition*. Assume, for simplicity, that we have a single control variable, *z*. We now define an underlying linear model w_i as

$$\begin{aligned} &y^* = \beta_{yx.z,w_1\ldots w_I}\, x + \beta_{yz.x,w_1\ldots w_I}\, z + \sum_i \beta_{yw_i.x,z} w_i + s, \\ &\text{with } sd(s) = \sigma_s \text{ and } \sigma_s = \sigma_l \cdot (\pi\ /\sqrt{3}) \end{aligned} \quad (21a)$$

and the corresponding logit model

$$\begin{aligned} &\text{logit}(\Pr(y=1)) = \\ &b_{yx.z,w_1\ldots w_I}\, x + b_{yz.x,w_1\ldots w_I}\, z + b_{yw_i.x,z} w = \\ &\frac{\beta_{yx.z,w_1\ldots w_I}}{\sigma_l} x + \frac{\beta_{yz.x,w_1\ldots w_I}}{\sigma_l} z + \sum_i \frac{\beta_{yw_i.x,z}}{\sigma_l} w_i. \end{aligned} \quad (21b)$$

where $\sigma_l < \sigma_e$, because the added concomitants, insofar they explain variation in the latent propensity, reduce the residual variation. Now, using the equation in (12) and the decomposition in (14), we may decompose the total effect *net of concomitants* into its direct and indirect parts, where the latter is the part mediated by *z*.

$$\text{Direct:} \qquad b_{yx.z,w_1 \ldots w_I} = \frac{\beta_{yx.z,w_1 \ldots w_I}}{\sigma_l} \tag{22a}$$

$$\text{Indirect:} \qquad b_{yz.x,w_1 \ldots w_I} \cdot \theta_{zx} = \frac{\beta_{yz.x,w_1 \ldots w_I}\theta_{zx}}{\sigma_l} \tag{22b}$$

$$\text{Total:} \qquad b_{yx.z,w_1 \ldots w_I} + b_{yz.x,w_1 \ldots w_I} \cdot \theta_{zx} = \frac{\beta_{yx.z,w_1 \ldots w_I} + \beta_{yz.x,w_1 \ldots w_I}\theta_{zx}}{\sigma_l}. \tag{22c}$$

Thus, replacing the logit coefficients in (10) with those in (21b) gives us a decomposition that is purged of the confounding effects of concomitants. The total, direct, and indirect effects are all measured on the same scale, and using the quantities in (22) for the percentage measure in (17) will therefore also be unaffected by scale parameters. Karlson and Holm (2011) present an example from educational stratification research where inclusion of concomitants in the decomposition may have a substantive interpretation. Thus, including concomitant variables measuring potentially confounding attributes ensures that the estimates of direct and indirect effects are not distorted by these attributes.

Up to this point we have assumed that the mediating variable, z, is continuous (notice that *x* could have been continuous or dichotomous). What happens to the decomposition when the observed mediating variable is binary? In the linear case, where y* is continuous, we have:

$$y^* = \gamma_{yx.z^*}x + \gamma_{yz^*.x}z^* + \sigma_v v \tag{23a}$$

Where *v* is a standardized error term and *z** is the dichotomous mediating variable. In general $\gamma_{yx.z^*} \neq \beta_{yx.z}$ and $\gamma_{yz^*.x} \neq \beta_{yz.x}$. Nevertheless, given the linear model

$$z^* = \varphi_{z^*x} x + m \tag{23b}$$

where m is an error term, it remains the case that

$$\beta_{yx} = \gamma_{yx.z^*} + \varphi_{z^*x}\gamma_{yz^*.x} \tag{23c}$$

That is to say, the total effect of *x* on *y* decomposes into a direct and indirect effect, given that the effect of *x* on *z** is estimated using a linear probability model and not a logit or other non-linear probability model.

Given *y*, a binary realization of *y**, we estimate the logit:

$$\text{logit}(\Pr(y=1)) = c_{yx.z} x + c_{yz^*.x} z^* = \frac{\gamma_{yx.z^*} x + \gamma_{yz^*.x} z^*}{\sigma_v} \tag{24}$$

Then the decomposition of the total effect into direct and indirect components is:

$$\frac{\gamma_{yx}}{\sigma_v} = c_{yx.\tilde{z}^*} = \frac{\gamma_{yx.z^*} + \varphi_{z^*x}\gamma_{yz^*.x}}{\sigma_v} \tag{25}$$

Where $c_{yx.\tilde{z}^*}$ is the logit coefficient for *x* in the model which controls for the residualized *z**.

MacKinnon and Dwyer (1993) present a method for decomposing effects in logit and probit models based on the y-standardization technique of Winship and Mare (1984). Briefly, this entails dividing the coefficients for *x* in each equation (in this case, equations 10 and 13a) by the estimated standard deviation of the predicted latent outcome, $\hat{y}$, for that model. The calculated coefficients are thus y-standardized, because they compensate for the rescaling of the "non-standardized" coefficients.

However, the method relies on the variance of the predicted logit index and, as Karlson, Holm and Breen (2011) demonstrate, whenever the prediction is skewed, the variance is a poor measure of dispersion, and y-standardization consequently fails as a method for comparing coefficients across nested models.

Using average partial effects for decompositions

The method we have presented can also be applied to average partial effects (APEs: Wooldridge 2002: 22-4). One advantage of APEs over logit or probit coefficients is that they are measured on the probability scale and are therefore intuitive and more easily understood than, say, partial log odds-ratios.

In logit and probit models, the marginal effect, ME, of *x* is the derivative of the predicted probability with respect to *x*, given by (when *x* is continuous and differentiable):

$$\frac{d\hat{p}}{dx} = \hat{p}(1-\hat{p})b = \hat{p}(1-\hat{p})\frac{\beta}{\sigma} = \frac{\hat{p}(1-\hat{p})}{\sigma}\beta, \qquad (26)$$

where $\hat{p} = \Pr(y=1 \mid x)$ is the predicted probability given *x* and $b = \frac{\beta}{\sigma}$ is the logit coefficient of *x*. The APE is then the average value of this derivative over the whole population. That is, the APE is defined as

$$\frac{1}{N}\sum_{i=1}^{N}\frac{d\hat{p}_i}{dx_i} = \frac{1}{N}\sum_{i=1}^{N}\frac{\hat{p}_i(1-\hat{p}_i)}{\sigma}\beta \qquad (27)$$

If the sample is drawn randomly from the population, the APE estimates the average marginal effect of *x* in the population.

Figure 2 illustrates how APEs can be used in decompositions of a total effect into direct and indirect effects. As in Equations (14a), (14b), and (14c), we express the direct, indirect, and total effect in the following way (dropping the *i* subscript for convenience):

$$\text{Direct: } APE(b_{yx.z}) = \frac{1}{N}\sum_{i=1}^{N}\frac{p(1-p)}{\sigma_e}\beta_{yx.z} \tag{28a}$$

$$\text{Indirect: } APE(b_{yz.x})\cdot\theta_{zx} = \frac{1}{N}\sum_{i=1}^{N}\frac{p(1-p)}{\sigma_e}\beta_{yz.x}\theta_{zx} \tag{28b}$$

$$\text{Total: } APE(b_{yx.z}) + APE(b_{yz.x})\cdot\theta_{zx} = \frac{1}{N}\sum_{i=1}^{N}\frac{p(1-p)}{\sigma_e}\beta_{yx.z} + \frac{1}{N}\sum_{i=1}^{N}\frac{p(1-p)}{\sigma_e}\beta_{yz.x}\theta_{zx}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\frac{p(1-p)}{\sigma_e}(\beta_{yx.z} + \beta_{yz.x}\theta_{zx}) \tag{28c}$$

Here θ is, as previously, the coefficient from the linear regression of *z* on *x*; β is the partial underlying regression coefficient, controlling for *z*, of *y* on *x*; γ is the partial logit regression coefficient, controlling for *x*, of *y* on *z*; and σ_e is the scale parameter residual standard deviation from the 'true' model. The total effect is equal to the APE of $\beta_{yx.\tilde{z}}$ as this is defined in (15) (see Karlson, Holm, and Breen 2011).

A statistical test

Drawing on the results in Sobel (1982), Karlson, Holm, and Breen (2011) develop a statistical test of the indirect effect using the delta method. Let $\Sigma_{b\theta}$ be the variance-

covariance matrix of the coefficients $b_{yz.x}$ and θ_{zx}, and define $b = b_{yz.x}$, $\theta = \theta_{zx}$ and $f = b_{yz.x}\theta_{zx}$. Then the asymptotic standard error of the indirect effect is given by

$$s.e(b_{yx.\tilde{z}} - b_{yx.z}) = \sqrt{\mathbf{a}'\frac{1}{N}\mathbf{\Sigma_{b\theta}}\mathbf{a}} \tag{29}$$

where N is the sample size and ***a*** is the vector of partial derivatives of the indirect effect, with respect to the parameters, b and θ: $\frac{\partial f}{\partial b}, \frac{\partial f}{\partial \theta}$.

Equations (10) and (12) form a recursive system of simultaneous equations, and so $b_{yz.x}$ and θ_{zx} are asymptotically independent, and thus $\Sigma_{b\theta}$ is a diagonal matrix whose entries are the asymptotic variances of the two coefficients (Sobel 1982: 294-5). We divide this by N to obtain the asymptotic standard errors.

This method extends easily to the case in which there are J *z* variables and thus J indirect effects. Now we define $f = \sum_j b_{yz(j).x}\theta_{z(j)x} = \mathbf{b}'\mathbf{\theta}$ and **b** and θ are column vectors and **a** is the 2J ×1 column vector whose entries are

$$\frac{\partial f}{\partial b_1}, \frac{\partial f}{\partial b_2}, ..., \frac{\partial f}{\partial b_J}, \frac{\partial f}{\partial \theta_1}, \frac{\partial f}{\partial \theta_2}, ..., \frac{\partial f}{\partial \theta_J}$$

Where the subscripts indicate the elements of the vectors **b** and θ. $\Sigma_{b\theta}$ is a diagonal matrix of dimension 2J and once again, (29) follows directly.

Given the standard error of the indirect effect we can test its significance in the usual way using the test statistic, Z which, in large samples, will be normally distributed:

$$Z = \frac{\sqrt{N}(b_{yx.\tilde{z}} - b_{yx.z})}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}_{\mathbf{b}\theta}\mathbf{a}}} \tag{30}$$

Examples

In this section we turn to two examples based on the National Educational Longitudinal Survey of 1988 (NELS). NELS is a nationally representative survey of 8th grade students in the US in 1988 who were followed until the year 2000, giving us the opportunity to study educational progress. We examine how much of the effect of parental socio-economic status (SES) on four year college graduation (COL) by year 2000 is explained or mediated by student academic ability (ABIL) and level of educational aspiration (LEA).[8] We standardize SES, ABIL, and LEA to have mean zero and variance of unity. Because we expect ability and aspirations to be positively correlated with parental SES and college graduation (e.g., Boudon 1974, Keller and Zavalloni 1964), we expect that both ability and aspirations mediate the effect of parental SES on college graduation. We also investigate which of ability and aspirations is the larger mediator. Because we suspect the decomposition to be affected by potentially confounding variables, we also include concomitants, gender (MALE), race (RACE), and intact family (INTACT). The final sample comprises 9,820 individuals, and Table 1 contains the descriptive statistics.[9] We calculate the decompositions using the Stata command *khb* (Kohler, Karlson, and Holm 2011), which implements the method developed by Karlson, Holm and Breen (2011).

[8] Within educational stratification research such empirical decompositions of family social status effects on educational decisions have received considerable attention, because they link to a theoretical model developed in a classic work on inequality of educational opportunity by Raymond Boudon (1974) and its generalization by Breen and Goldthorpe (1997) (see Erikson et al. 2005; Morgan 2010).

[9] Because we use the NELS Public Use File, the original sample comprises around 12,144 individuals. Because this example acts as an illustration of our method, we do not discuss the nonresponse patterns and the possible biases they may entail.

-- TABLE 1 HERE --

We structure the analysis in four steps. First, we decompose the effect of SES on COL using ABIL. Second, we add LEA to the decomposition and evaluate which variable, ABIL or LEA, has the larger indirect effect. Third, we add three concomitants, MALE, RACE, and INTACT to the decomposition to control for possibly confounding variables. Fourth, we report the results in terms of average partial effects, giving the decomposition a more substantive interpretation. Because the results may be sensitive to model choice, we report them for both logit and probit models.

Table 2 reports the results of a decomposition of SES on COL with ABIL as the mediator. Using the formulas in (14), we decompose, in logits (probits) the total effect of 1.348 (0.781) into a direct part, 0.914 (0.524), and an indirect part, 0.434 (0.257). Using the formula for the z-statistic in (30), we see that all effects are highly statistically significant. We also see that the indirect effect is around half the magnitude of the total effect. In relative terms, the indirect effects accounts for 32.2% of the total effect in the logit model and 32.9% in the probit model. In the second row from the bottom of Table 2 we label this the confounding percentage. This is very similar for the logit and probit, indicating that our decomposition is not sensitive to the choice of a normal or logistic error distribution. In the final row we report the naive confounding percentage, which is what we would have obtained had we simply compared the coefficients across models with and without ABIL. This is 25.3% for the logit model and 26.8% for the probit model, indicating that a naive comparison of effects would underestimate the true amount of confounding net of rescaling.

-- TABLE 2 HERE --

In Table 3 we add LEA to the decomposition and break down the indirect effect due to both ABIL and LEA into its respective components. We see that all effects are highly statistically significant. Because the logit and probit return near-identical results, we focus only on the results based on the former. Looking at the relative measures of the indirect effect, we see that, compared to Table 2, the confounding percentage has increased from 32.2 to 56.6%. However, more of the effect of SES is mediated by LEA than by ABIL, LEA accounting for 37.5% of the total effect, ABIL for 19.1%. The confounding percentage for ABIL is considerably smaller than the 32.2% reported in Table 2. Thus, including LEA in the decomposition reduces the contribution of ABIL to the total effect of about 13 percentage points, and this is because LEA is positively correlated with SES, ABIL, and COL. We also see that the naïve use of the logit would underestimate the confounding percentage by about 16 percentage points (41 compared with 57%).

-- TABLE 3 HERE --

In Table 4 we add three concomitants, MALE, RACE, and INTACT, which we suspect may confound the decomposition. Using the formulae 22a to c, these concomitants are included in all models used for the decomposition, thereby holding constant their possible influence on the results. We see that the results are virtually identical to those

reported in Table 3. These findings suggest that the decomposition presented in Table 3 is unaffected by the influence of the concomitants.

-- TABLE 4 HERE --

In Table 5 we report average partial effects (APE) of the results in Table 4, using formulae 28a to c. Because the standard error of the indirect effect is unknown, we only report the APEs and once again we focus on the results from the logit model. We see that the total effect is 0.224, which means that for a standard deviation change in SES, the probability of graduating college increases by 22.4 percentage points. Decomposing this effect returns a direct effect of 9.6 percentage points, and an indirect of 12.8 percentage points. Breaking down the indirect effect to its two components, we find that the indirect effect via ABIL is 4.3 percentage points, and 8.5 percentage points via LEA. Thus, the effect of SES on COL running via LEA is substantially larger than the one running through ABIL. We note that the confounding percentages in Table 5 equal those in Table 4, as is evident from (28c). However, the naïve confounding percentage in the final column differs between the two tables. In Table 4, the naïve percentage conflates confounding and rescaling, while the counterpart in Table 5 conflates confounding with the distributional sensitivity of the APE. As we would expect, the naïve confounding percentage is much smaller for the APE than for the logit. APE underestimates the true percentage by about 3.5 percentage points compared with the underestimate from the logit of 16 percentage points.

Discussion

In this paper we have reported a series of new findings about mediating or confounding relationships in non-linear probability models such as logit and probit models. Drawing on Karlson, Holm, and Breen (2011), we developed a method that decomposes logit or probit coefficients into total, direct, and indirect effects. We provided a statistical test and developed several extensions of the method; in particular we applied it to average partial effects, giving researchers an effect measure on the probability scale which may be more interpretable than logit and probit coefficients. We illustrated our method using data from the National Longitudinal Survey of Youth 1988 and found that it fares much better than naïve decompositions of logit and probit coefficients. Because naïve decompositions conflate rescaling and confounding, they tend to underestimate the degree of confounding. Average partial effects also underestimate the degree of confounding, though not by as much.

The method presented in this paper extends the decomposability properties of linear models to logit and probit models and can be applied when the variable of interest and the mediating variables are continuous or discrete. We can also include a set of concomitants which may confound the decomposition of interest. Perhaps most usefully, the method can be applied very easily using the Stata routine *khb* (Kohler, Karlson, and Holm 2011) based on Karlson, Holm and Breen (2011). As we noted earlier, the method we present can be extended to other models such as the ordered logit and probit and also to the multinomial logit and all these cases can be dealt with using *khb*. Whether the method extends to all models of the family of generalized linear models is a topic for future research.

References

Amemiya, Takeshi. 1975. "Qualitative Response Models." *Annals of Economic and Social Measurement* 4:363-388.

Alwin, Duane and Robert M. Hauser. 1975. "The Decomposition of Effects in Path Analysis". *American Sociological Review* 40: 37-47.

Blalock, Hubert M. 1979. *Social Statistics*, 2nd ed. New York: McGraw-Hill.

Bollen, Kenneth A. 1987. "Total, direct and indirect effects in structural equation models." *Sociological Methodology* 17: 37-69.

Boudon, R. 1974. *Education, Opportunity and Social Inequality*. New York: Wiley.

Breen, Richard and John H. Goldthorpe.1997. "Explaining Educational Differentials: Towards a Formal Rational Action Theory." *Rationality and Society* 9: 275-305.

Cameron, Stephen V. and James J. Heckman. 1998. "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males". *Journal of Political Economy* 106:262-333.

Clogg, Clifford C., Eva Petkova, and Adamantios Haritou. 1995a. "Statistical Methods for Comparing Regression Coefficients Between Models." *The American Journal of Sociology* 100:1261-1293.

------. 1995b. Reply to Allison: More on Comparing Regression Coefficients. *The American Journal of Sociology*, 100, 1305-1312.

Cramer, J.S. 2003. *Logit Models*. *From Economics and Other Fields.* Cambridge: Cambridge University Press.

Duncan, Otis D. 1966. "Path Analysis: Sociological Examples." *The American Journal of Sociology* 72:1-16.

Erikson, Robert, John H. Goldthorpe, Michelle Jackson, Meir Yaish, and D.R. Cox. 2005. "On class differentials in educational attainment." *Proceedings of the National Academy of Science (PNAS) of the USA*, 102, 9730-9733.

Fienberg, Stephen E. 1977. *The Analysis of Cross-classified Categorical Data.* Cambridge, Massachusetts: MIT Press.

Hosmer, David W. and Stanley Lemeshow. 1989. *Applied Logistic Regression*. New York: Wiley.

Jackson, Michelle, Robert Erikson, John H. Goldthorpe, and Meir Yaish. 2007. Primary and Secondary Effects in Class Differentials in Educational Attainment: The Transition to A-Level Courses in England and Wales. *Acta Sociologica*, 50, 211-229.

Karlson, Kristian B. and Anders Holm. 2011. "Decomposing primary and secondary effects: A new decomposition method." Accepted for publication in *Research in Stratification and Social Mobility*.

Karlson, Kristian B., Anders Holm, and Richard Breen. 2011. "Comparing Regression Coefficients Between Models using Logit and Probit: A New Method." *Under review*. [Available from authors].

Keller, S. and Zavalloni, M. 1964. "Ambition and Social Class: A Respecification." *Social Forces* 43:58-70.

Kohler, Ulrich, Kristian B. Karlson, and Anders Holm. 2011. "Decomposition of total effects into direct and indirect effects using the KHB-method." *Working Paper* on new Stata command *khb*. [Available from authors].

Lazarsfeld, Paul F. 1955. "The Interpretation of Statistical Relations as a Research Operation." Pp. 115-125 in *The Language of Social Research,* edited by P.F. Lazarsfeld and M. Rosenberg. Glencoe, Illinois: The Free Press.

Long, J.S. 1997. *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks: Sage.

MacKinnon, David P. and James H. Dwyer. 1993. "Estimating Mediated Effects in Prevention Studies." *Evaluation Review* 17: 144-58.

Morgan, Stephen L. 2010. "Models of College Entry and the Challenges of Estimating Primary and Secondary Effects." *Working Paper.* [Obtained from author].

Powers, Daniel A. and Yu Xie. 2000. *Statistical Methods for Categorical Data Analysis*. San Diego: Academic Press.

Sobel, Michael. 2008. "Identification of Causal Parameters in Randomized Studies with Mediating Variables." *Journal of Educational and Behavioral Statistics* 33: 230-51.

VanderWeele, Tyler J. 2010. "Marginal Structural Models for the Estimation of Direct and Indirect Effects." *Epidemiology* 20: 18-26.

Winship, Christopher and Robert D. Mare. 1983. "Structural Equations and Path Analysis for Discrete Data." *The American Journal of Sociology* 89:54-110.

------. 1984. "Regression Models with Ordinal Variables." *American Sociological Review* 49:512-525.

Wooldridge, Jeffrey M. 2002. *Econometric analysis of cross section and panel data*. Massachusetts: MIT Press.

Yatchew, Adonis and Zvi Griliches. 1985. "Specification Error in Probit Models." *The*

Review of Economics and Statistics 67:134-139.

Appendix

For the linear model, we want to prove that

$$\beta^* - \beta = \gamma\theta . \qquad \text{(A1)}$$

where β^* is the effect of *x* on *y* in a model excluding a confounder, *z*, β is the counterpart in a model including *z*, γ is the effect of *z* on *y* in a model including *z*, and θ is the effect of *x* on *z* from a linear regression. Applying basic principles of OLS to the model with *y* as dependent variable, we may write the coefficients of *x* as

$$\beta^* = r_{yx}\frac{\sigma_y}{\sigma_x} \quad \text{and} \quad \beta = \frac{(r_{yx} - r_{xz}r_{yz})}{(1 - r_{xz}^2)}\frac{\sigma_y}{\sigma_x} \qquad \text{(A2)},$$

where r denotes the correlation coefficient. Clogg *et al.* (1995b) show that the difference between these two coefficients equals

$$\beta^* - \beta = \frac{r_{xz}(r_{yz} - r_{xz}r_{yx})}{(1 - r_{xz}^2)}\frac{\sigma_y}{\sigma_x} .$$

Now, the coefficients γ and θ can be written as

$$\gamma = \frac{(r_{yz} - r_{xz}r_{yx})}{(1 - r_{xz}^2)}\frac{\sigma_y}{\sigma_z} \quad \text{and} \quad \theta = r_{zx}\frac{\sigma_z}{\sigma_x} .$$

The product of these two coefficients is

$$\beta_{yz\cdot x}\theta_{zx} = \frac{(r_{yz} - r_{xz}r_{yx})}{(1 - r_{xz}^2)}\frac{\sigma_y}{\sigma_z}r_{xz}\frac{\sigma_z}{\sigma_x} = \frac{r_{xz}(r_{yz} - r_{xz}r_{yx})}{(1 - r_{xz}^2)}\frac{\sigma_y}{\sigma_x}, \qquad \text{(A3)}$$

which equals (A2). We have consequently proven the equality in (A1).

For the logit model we want to prove that

$$\frac{\beta^* - \beta}{\sigma_e} = \frac{\gamma\theta}{\sigma_e} . \qquad \text{(A4)}$$

However, multiplying both sides by σ_e leaves us with

$$\beta^* - \beta = \gamma\theta,$$

which we proved for the linear model. We have consequently proven the equality in (A4).

TABLES

TABLE 1. Variable descriptive. N = 9,820.

	Mean	SD
COL	0.36	-
SES	0	1
ABIL	0	1
LEA	0	1
MALE	0.47	-
RACE		
White (reference)	0.69	-
Hispanic	0.12	-
Black	0.09	-
Other	0.10	-
INTACT	0.90	-

TABLE 2. Decomposition of total effect of SES on COL into direct and indirect effect via ABIL.

	LOGIT		**PROBIT**	
	Coef.	z	Coef.	z
Coefficients				
Total effect	1.348	42.06	0.781	45.23
Direct effect	0.914	28.90	0.524	29.57
Indirect effect	0.434	26.06	0.257	26.79
Relative measures				
Confounding percentage	32.2	-	32.9	-
Naive conf. percentage	25.3		26.8	

TABLE 3. Decomposition of total effect of SES on COL into direct and indirect effect via ABIL and LEA.

	LOGIT		**PROBIT**	
	Coef.	z	Coef.	z
Coefficients				
Total effect	1.657	42.83	0.939	46.33
Direct effect	0.718	21.48	0.421	22.31
Indirect effect	0.938	29.08	0.518	30.67
via ABIL	0.317	18.87	0.192	19.78
via LEA	0.621	22.55	0.326	23.58
Relative measures				
Confounding percentage	56.6	-	55.2	-
via ABIL	19.1	-	20.4	-
via LEA	37.5	-	34.7	-
Naive conf. percentage	41.3	-	41.2	-

TABLE 4. Decomposition of total effect of SES on COL into direct and indirect effect via ABIL and LEA, controlling for concomitants MALE, RACE, and INTACT

	LOGIT		PROBIT	
	Coef.	z	Coef.	z
Coefficients				
Total effect	1.634	41.31	0.927	44.3
Direct effect	0.702	20.43	0.413	21.25
Indirect effect	0.932	28.33	0.514	29.9
via ABIL	0.312	17.95	0.189	18.94
via LEA	0.620	22.20	0.325	23.21
Relative measures				
Confounding percentage	57.0	-	55.4	-
via ABIL	19.1	-	20.4	-
via LEA	38.0	-	35.0	-
Naive conf. percentage	40.9	-	40.7	-

TABLE 5. APE decomposition of total effect of SES on COL into direct and indirect effect via ABIL and LEA, controlling for concomitants MALE, RACE, and INTACT

	LOGIT	PROBIT
	APE	APE
Coefficients		
Total effect	0.2242	0.2205
Direct effect	0.0963	0.0983
Indirect effect	0.1279	0.1223
via ABIL	0.0428	0.0451
via LEA	0.0851	0.0772
Relative measures		
Confounding percentage	57.0	55.4
via ABIL	19.1	20.4
via LEA	38.0	35.0
Naive conf. percentage	53.6	52.5

FIGURES

Figure 1 Path decomposition into direct and indirect effects

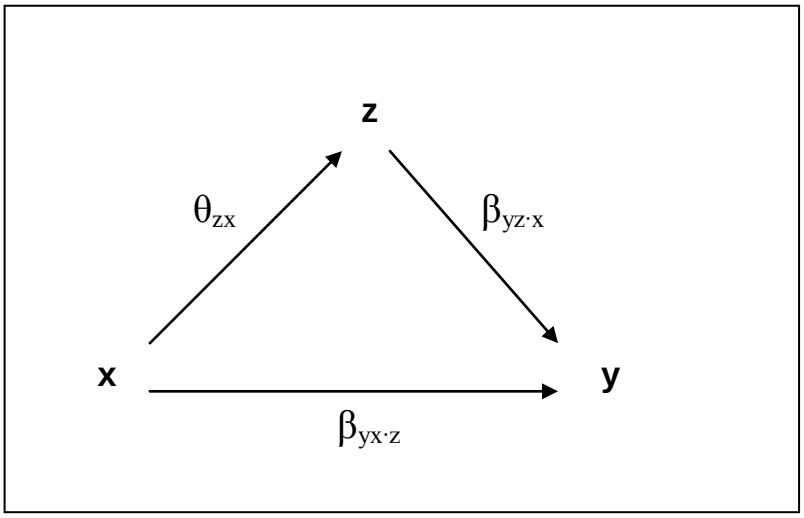
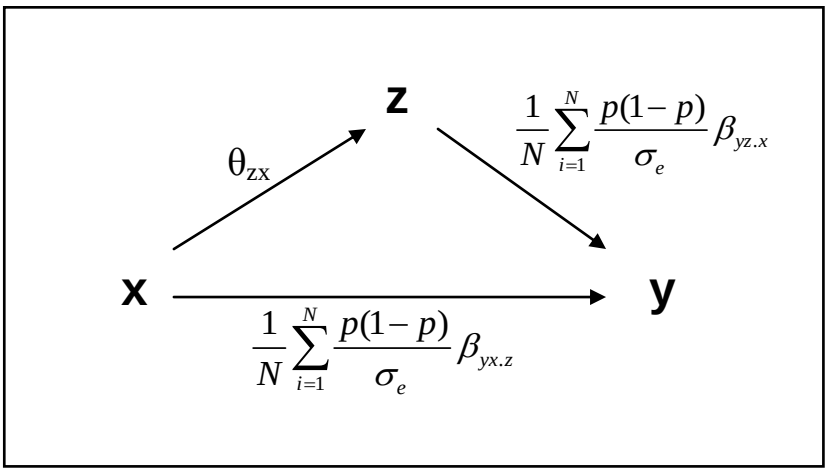


Figure 2 A simple path model illustrating a fully recursive system using APEs



CSER
CENTRE FOR STRATEGIC RESEARCH IN EDUCATION