1

Effect Heterogeneity and Bias in Main-Effects-Only Regression Models

FELIX ELWERT AND CHRISTOPHER WINSHIP

Introduction

The overwhelming majority of OLS regression models estimated in the social sciences, and in sociology in particular, enter all independent variables as main effects. Few regression models contain many, if any, interaction terms. Most social scientists would probably agree that the assumption of constant effects that is embedded in main-effects-only regression models is theoretically implausible. Instead, they would maintain that regression effects are historically and contextually contingent; that effects vary across individuals, between groups, over time, and across space. In other words, social scientists doubt constant effects and believe in effect heterogeneity.

But why, if social scientists believe in effect heterogeneity, are they willing to substantively interpret main-effects-only regression models? The answer—not that it's been discussed explicitly—lies in the implicit assumption that the main-effects coefficients in linear regression represent straightforward averages of heterogeneous individual-level causal effects.

The belief in the averaging property of linear regression has previously been challenged. Angrist [1998] investigated OLS regression models that were correctly specified in all conventional respects except that effect heterogeneity in the main treatment of interest remained unmodeled. Angrist showed that the regression coefficient for this treatment variable gives a rather peculiar type of average—a conditional variance weighted average of the heterogeneous individual-level treatment effects in the sample. If the weights differ greatly across sample members, the coefficient on the treatment variable in an otherwise well-specified model may differ considerably from the arithmetic mean of the individual-level effects among sample members.

In this paper, we raise a new concern about main-effects-only regression models. Instead of considering models in which heterogeneity remains unmodeled in only one effect, we consider standard linear path models in which unmodeled heterogeneity is potentially pervasive.

Using simple examples, we show that unmodeled effect heterogeneity in more than one structural parameter may mask confounding and selection bias, and thus lead to biased estimates. In our simulations, this heterogeneity is indexed by latent (unobserved) group membership. We believe that this setup represents a fairly realistic scenario—one in which the analysts has no choice but to resort to a main-effects-only regression model because she cannot include the desired interaction terms since group-membership is unobserved. Drawing on Judea Pearl's theory of directed acyclic graphs (DAG) [1995, 2009] and VanderWeele and Robins [2007], we then show that the specific biases we report can be predicted from an analysis of the appropriate DAG. This paper is intended as a serious warning to applied regression modelers to beware of unmodeled effect heterogeneity, as it may lead to gross misinterpretation of conventional path models.

We start with a brief discussion of conventional attitudes toward effect heterogeneity in the social sciences and in sociology in particular, formalize the notion of effect heterogeneity, and briefly review results of related work. In the core sections of the paper, we use simulations to demonstrate the failure of main-effects-only regression models to recover average causal effects in certain very basic three-variable path models where unmodeled effect heterogeneity is present in more than one structural parameter. Using DAGs, we explain which constellations of unmodeled effect heterogeneity will bias conventional regression estimates. We conclude with a summary of findings.

The Presumed Averaging Property of Main-Effects-Only Regression Models

The great majority of empirical work in the social sciences relies on the assumption of constant coefficients to estimate OLS regression models that contain nothing but main effect terms for all variables considered.[1] Of course, most researchers do not believe that real-life social processes follow the constant-coefficient ideal of conventional regression. For example, they aver that the effect of martial conflict on children's self-esteem is larger for boys than for girls [Amato and Booth 1997]; or that the death of a spouse increases mortality more for white widows than for African American widows [Elwert and Christakis 2006]. When pressed, social scientists would probably agree that the causal effect of almost any treatment on almost any outcome likely varies from group to group, and from person to person.

But if researchers are such firm believers in effect heterogeneity, why is the constant-coefficients regression model so firmly entrenched in empirical practice? The answer lies in the widespread belief that the coefficients of linear regression models estimate averages of heterogeneous parameters—average causal effects—representing the average of the individual-level causal effects across sample members. This (presumed) averaging property of standard regression models is important for empirical practice for at least

[1] Whether a model requires an interaction depends on the functional form of the dependent and/or independent variables. For example, a model with no interactions in which the independent variables are entered in log form, would require a whole series of interactions in order to approximate this function if the independent variables where entered in nonlog form.

three reasons. First, sample sizes in the social sciences are often too small to investigate effect heterogeneity by including interaction terms between the treatment and more than a few common effect modifiers (such as sex, race, education, income, or place of residence); second, the variables needed to explicitly model heterogeneity may well not have been measured; third, and most importantly, the complete list of effect modifiers along which the causal effect of treatment on the outcome varies is typically unknown (indeed, unknowable) to the analyst in any specific application. Analysts thus rely on faith that their failure to anticipate and incorporate all dimensions of effect heterogeneity into regression analysis simply shifts the interpretation of regression coefficients from individual-level causal effects to average causal effects, without imperiling the causal nature of the estimate.

Defining Effect Heterogeneity

We start by developing our analysis of the consequences of causal heterogeneity within the counterfactual (potential outcomes) model. For a continuous treatment $T \in (-\infty,\infty)$, let $T = t$ denote some specific treatment value and $T = 0$ the control condition. $Y(t)_i$ is the potential outcome of individual i for treatment $T = t$, and $Y(0)_i$ is the potential outcome of individual i for the control condition. For a particular individual, generally only one value of $Y(t)_i$ will be observed. The *individual-level causal effect* (ICE) of treatment level $T = t$ compared to $T = 0$ is then defined as: $\delta_i = Y(t)_i - Y(0)_i$.

Since δ_i is generally not directly estimable, researchers typically attempt estimating the *average causal effect* (ACE) for some sample or population:

$$\bar{\delta} = \sum_{i=1}^{N} \delta_i / N$$

We say that the effect of treatment T is *heterogeneous* if: $\delta_i \neq \bar{\delta}$ for at least one i.

In other words, effect heterogeneity exists if the causal effect of the treatment differs across individuals. The basic question of this paper is whether a regression estimate for the causal effect of the treatment can be interpreted as an average causal effect if effect heterogeneity is present.

Regression Estimates as Conditional Variance Weighted Average Causal Effects[2]

The ability of regression to recover average causal effects under effect heterogeneity has previously been challenged by Angrist [1998]. Here, we briefly sketch the main result. For a binary treatment, T=0,1, Angrist analyzed situations where the effect of treatment varied across strata defined by the values of the covariates X in the model but the OLS regression estimated was misspecified to include only a main effect term and no interactions between treatment and X. Angrist showed that the regression estimate for the main effect of treatment can be expressed as a weighted average of stratum-specific treatment effects, albeit one that is difficult to interpret. For each stratum defined by fixed values of X, the numerator of the OLS estimator has the form: $\delta_x W_x P(X=x)$, where δ_x is the

[2] This presentation follows Angrist 1998 and Angrist and Pischke 2009.

stratum-specific causal effect and P(X=x) is the relative size of the stratum in the sample. The weight, W_x, is a function of the propensity score, P_x=P(T=1 | X), associated with the stratum, $W_x = P_x$ (1- P_x), which equals the stratum-specific variance of treatment. This variance, and hence the weight, is largest if P_x is .5 and smaller as P_x goes to 0 or 1.

If the treatment effect is constant across strata, these weights make good sense. OLS gives the minimum variance linear unbiased estimator of the model parameters under homoscedasticity assuming correct specification of the model. Thus in a model without interactions between treatment and covariates X the OLS estimator gives the most weight to strata with the smallest variance for the estimated within-stratum treatment effect, which, not considering the size of the strata, are those strata with the largest treatment variance, i.e. with the P_x that are closest to .5. However, if effects are heterogeneous across strata, this weighting scheme makes little substantive sense: in order to compute the average causal effect, $\bar{\delta}$, as defined above, we would want to give the same weight to every individual in the sample. As a variance-weighted estimator, however, regression estimates under conditions of unmodeled effect heterogeneity will not converge to the (unweighted) average treatment effect.

Path Models with Pervasive Effect Heterogeneity

Whereas Angrist analyzed a misspecified regression equation that incorrectly assumed no treatment-covariate interaction for a *single* treatment variable, we investigate the ability of a main-effects-only regression model to recover unbiased average causal effects in simple path models with effect heterogeneity across *multiple* parameters.

Setup: To illustrate how misleading the belief in the averaging power of the constant-coefficient model can be in practice, we present simulations of basic linear path models of the form:

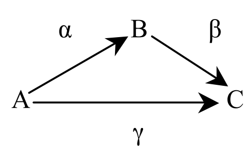

Figure 1: A simple linear path model

where we have repressed the usual uncorrelated error terms. To introduce effect heterogeneity, let G = 0, 1 index membership in a latent group and permit the possibility that the three structural parameters α, β, and γ vary across (but not within) levels of G. The above path model can then be represented by two linear equations: $B = A\alpha_G + \varepsilon_B$ and $C = A\gamma_G + B\beta_G + \varepsilon_C$. In our simulations, we assume that A~N(0,1) and ε_B, and ε_C are iid N(0,1), and hence all variables are normally distributed. From these equations, we next simulate populations of N=100,000 observations, with P(G=1) =P(G=0)= 1/2. We start with a population in which all three parameters are constant across the two subgroups defined by G, and then systematically introduce effect heterogeneity by successively permitting the structural parameters to vary by group, yielding one population for each of the $2^3 = 8$ possible combinations of constant/varying parameters. To fix ideas, we choose the following group-specific parameter values, shown in Table 1:

Table 1: Group-specific structural parameters for simulations

	α_G	β_G	γ_G
Group:			
G=0	0.4	0.5	0.6
G=1	1.2	2.5	1.4
Average	0.8	1.5	1.0

For simulations in which one or more parameters do not vary by group, we set the constant parameter(s) to the average of the group specific parameters, e.g. $\alpha = (\alpha_0 + \alpha_1)/2$. Finally, we estimate a conventional linear regression model for the effects of A and B on C using the conventional default specification, in which all variables enter as main effects only, $C=A\gamma + B\beta + \varepsilon$.

Recall that G is latent and therefore cannot be included in the model. The parameter, γ refers to the direct effect of A on C holding B constant, and β refers to the total effect of B on C.[3] In much sociological and social science research, this main-effects regression model is intended to recover average structural (causal) effects, and is commonly believed to be well suited for the purpose.

Results: Table 2 shows the regression estimates for the main effect parameters across the eight scenarios of effect heterogeneity. We see that the main effects regression model correctly recovers the desired (average) parameters, γ=1 and β=1.5 if none of the parameters vary across groups (column 1), or if only one of the three parameters varies (columns 2-4).

Other constellations of effect heterogeneity, however, produce biased estimates. If α_G and β_G (column 5); or α_G and γ_G (column 6); or α_G, β_G, and γ_G (column 8) vary across groups, the main-effects-only regression model fails to recover the true (average) parameter values known to underlie the simulations. For our specific parameter values, the estimated (average) effect of B on C in these troubled scenarios is always too high, and the estimated average direct effect of A on C is either too high or too low. Indeed, if we set γ=0 but let α_G and β_G vary across groups, the main-effects-only regression model would indicate the presence of a direct effect of A on C even though it is known by design that no such direct effect exists (not shown).

Failure of the regression model to recover the known path parameters is not merely a function of the number of paths that vary. Although none of the scenarios in which fewer than two parameters vary yield incorrect estimates, and the scenario in which all three parameters vary is clearly biased, results differ for the three scenarios in which exactly two parameters vary. In two of these scenarios (columns 5 and 6), regression fails to

[3] The notion of direct and indirect effects is receiving deserved scrutiny in important recent work by Robins and Greenland 1992; Pearl 2001; Robins 2003; Frangakis and Rubin 2002; Sobel 2008; and VanderWeele 2008.

recover the desired (average) parameters, while in the third scenario (column 7), regression does recover the correct average parameters.

Table 2: OLS regression estimates for the main effects of A and B on C across eight different combinations of effect heterogeneity in α, β,and/or γ

	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)	
Heterogeneity in:	-		α		β		γ		α, β		α, γ		β, γ		α, β, γ	
Group:	G0	G1	G0	G1	G0	G1	G0	G1	G0	G1	G0	G1	G0	G1	G0	G1
α	0.8		0.4	1.2	0.8		0.8		0.4	1.2	0.4	1.2	0.8		0.4	1.2
β	1.5		1.5		0.5	2.5	1.5		0.5	2.5	1.5		0.5	2.5	0.5	2.5
γ	1.0		1.0		1.0		0.6	1.4	1.0		0.6	1.4	0.6	1.4	0.6	1.4
Pooled regression estimate for:																
β	1.50		1.50		1.50		1.50		**1.77**		**1.64**		1.50		**1.91**	
γ	1.00		1.00		1.00		1.00		**1.17**		**0.89**		1.00		**1.07**	

Note: Bold estimates are biased for the true (average) parameters. Results from independent simulations of N=100,000 for each scenario, using (group-specific) parameters listed above. See text for details.

In sum, the naïve main-effects-only linear regression model recovers the correct average parameter values only under certain conditions of limited effect heterogeneity, and it fails to recover the true average effects in certain other scenarios, including the scenario we consider most plausible in the majority of sociological applications, i.e., where all three parameters vary between groups. If group membership is latent—because group membership is unknown to or unmeasured by the analyst— and thus unmodeled, linear regression generally will fail to recover the true average effects.

DAGs to the Rescue

These results spell trouble for empirical practice in sociology. Judea Pearl's work on causality and directed acyclic graphs (DAGs) [1995, 2000, 2009] offers an elegant and powerful approach to understanding the problem. The critical insight for the present discussion is that effect heterogeneity, rather than being a nuisance that is easily averaged away, encodes structural information that analysts ignore at their peril.

Pearl's DAGs are nonparametric path models that encode causal dependence between variables: an arrow between two variables indicates that the second variable is causally dependent on the first (for detailed formal expositions of DAGs, see Pearl 1995, 2009; for less technical introductions see Robins 2001; Greenland, Pearl and Robins 1999 in epidemiology, and Morgan and Winship 2007 in sociology). For example, the DAG in Figure 2 indicates that Z is a function of X and Y, Z= f(X,Y,ε_Z), where ε_Z is an unobserved error term independent of (X,Y):

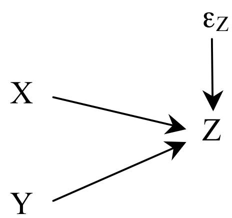

Figure 2: DAG illustrating direct effect modification of the effect of Y on Z in X

The advantage of using non-parametric DAGs is that f( ) can be any function. For example, the DAG in Figure 2 is consistent with a linear structural equation in which X only modifies (i.e. introduces heterogeneity into) the effect of Y on Z, $Z=Y\xi + YX\psi + \varepsilon_Z$.[4] In the language of VanderWeele and Robins [2007], who provide the most extensive treatment of effect heterogeneity using DAGs to date, one may call X a "direct effect modifier" of the effect of Y on Z. The point is that a variable that modifies the effect of Y on Z is causally associated with Z, as represented by the arrow from X to Z.

Returning to our simulation, one realizes that the social science path model of Figure 1, although a useful tool for informally illustrating the data generation process, does not, generally, provide a sufficiently rigorous description of the causal structure underlying the simulations. Figure 1, although truthfully representing the separate data generating mechanism for each group and each individual in the simulated population, is not generally the correct DAG for the pooled population containing groups G = 0 and G = 1 for all of the heterogeneity scenarios considered above. Specifically, in order to turn the social science path model of Figure 1 into a DAG, one would have to integrate G into the picture. How this is to be done depends on the structure of heterogeneity. If only β_G (the effect of B on C) and/or γ_G (the direct effect of A on C holding B constant) varied with G, then one would add an arrow from G into C. If α_G (the effect of A on B) varied with G, then one would add an arrow from G into B.

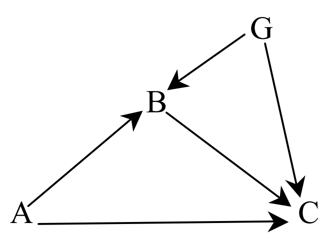

Figure 3: DAG consistent with effect modification of the effects of A on B, and B on C and/or A on C, in G

[4] It is also consistent with an equation that adds a main effect of X. For the purposes of this paper it does not matter whether the main effect is present.

The DAG in Figure 3 thus represents the DAG for those simulated scenarios in which α_G as well as either β_G or γ_G, or both, vary with G (columns 5, 6, and 8). Interpreted in terms of a linear path model, this DAG is consistent with the following two structural equations: $B = A\alpha_0 + AG\alpha_1 + \varepsilon_B$ and $C = A\gamma_0 + AG\gamma_1 + B\beta_0 + BG\beta_1 + \varepsilon_C$, where the iid errors, ε, have been omitted from the DAG and are assumed to be uncorrelated.[5]

In our analysis, mimicking the reality of limited observational data with weak substantive theory, we have assumed that A, B, and C are observed, but that G is not observed. It is immediately apparent that the presence of G in Figure 3 creates two problems: first, G is a confounder for the effect of B on C. Second, B is a "collider" [Pearl 2009] on the undirected path from A to C via B and G. Together, these two problems explain the failure of the main-effects-only regression model to recover the true parameters in panels 5, 6, and 8: in order to recover the effect of B on C, β, we need to condition on the confounders A and G. But G is latent so it cannot be conditioned on. At the same time, conditioning on the collider B in the regression opens a "backdoor path" from A to C via B and G (when G is not conditioned on), which induces selection bias in the estimate for the direct effect of A on C, γ [Pearl 2000; Hernán et al 2004]. Hence, both coefficients in the main-effects-only regression model will be biased for the true (average) parameters.

By contrast, if G modifies neither β nor γ, then the DAG would not contain an arrow from G into C; and if G does not modify α then the DAG would not contain an arrow from G into B. Either way, if either one (or both) of the arrows emanating from G are missing, then G is not a confounder for the effect of B on C, and conditioning on B will not open a backdoor path from A to C. If so, the main effects regression model is unbiased and will recover the true parameters, and correctly average them where appropriate, as seen in panels 1-4 and 7.

In sum, Pearl's DAGs neatly display the structural information encoded in effect heterogeneity [VanderWeele and Robins 2007]. Consequently, Pearl's DAGs immediately draw attention to problems of confounding and selection bias that may occur when more than one effect in a causal system varies across sample members. Analyzing the appropriate DAG, the failure of main-effects-only regression models to recover average structural parameters in certain constellations of effect heterogeneity becomes predictable.

Conclusion

This paper considered a conventional structural model of a kind commonly used in the social sciences and explored its performance under various basic scenarios of effect heterogeneity. Simulations show that the standard social science strategy of dealing with effect heterogeneity—by ignoring it—is prone to failure. In certain situations, the main-

---

[5] By construction of the example, we assume that A is randomized and thus marginally independent of G. Note, however, that even though G is mean independent of B and C (no main effect of G on either B or C), G is not marginally independent of B or C because $var(B|G=1) \neq var(B|G=0)$ and $var(C|G=1) \neq var(C|G=0)$. Adding main effects of G on B and C would not change the arguments presented here.

effects-only regression model will recover the desired quantities, but in others it will not. We believe that effect heterogeneity in all arrows of a path model is plausible in many, if not most, substantive applications. Since the sources of heterogeneity are often not theorized, known, or measured, social scientists continue routinely to estimate main-effects-only regression models in hopes of recovering average causal effects. Our examples demonstrate that the belief in the averaging powers of main-effects-only regression models may be misplaced if heterogeneity is pervasive, as estimates can be mildly or wildly off the mark. Judea Pearl's DAGs provide a straightforward explanation for these difficulties. DAGs remind analysts that effect heterogeneity encodes structural information about confounding and selection bias that requires consideration when designing statistical strategies for recovering the desired average causal effects.

Acknowledgments: We thank Jamie Robins for detailed comments on a draft version of this paper, and Stephen Morgan, Hyun Sik Kim, and Elizabeth Wrigley-Field for advice. Genevieve Butler provided editorial assistance.

References

Amato, Paul R., and Alan Booth. (1997). *A Generation at Risk: Growing Up in an Era of Family Upheaval*. Cambridge, MA: Harvard University Press.

Angrist, Joshua D. (1998). "Estimating the Labor Market Impact on Voluntary Military Service Using Social Security Date on Military Applicants." *Econometrica 66*: 249-88.

Angrist, Joshua D. and Jörn-Steffen Pischke. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Elwert, Felix, and Nicholas A. Christakis. (2006). "Widowhood and Race." *American Sociological Review 71*: 16-41.

Frangakis, Constantine E., and Donald B. Rubin. (2002). "Principal Stratification in Causal Inference." *Biometrics 58*: 21–29.

Greenland, Sander, Judea Pearl, and James M. Robbins. (1999). "Causal Diagrams for Epidemiologic Research." *Epidemiology 10*: 37-48.

Hernán, Miguel A., Sonia Hernández-Diaz, and James M. Robins. (2004). "A Structural Approach to Section Bias." *Epidemiology 155* (2): 174-184.

Morgan, Stephen L. and Christopher Winship. (2007). *Counterfactuals and Causal Inference: Methods and Principles of Social Research.* Cambridge: Cambridge University Press.

Pearl, Judea. (1995). "Causal Diagrams for Empirical Research." *Biometrika 82* (4): 669-710.

Pearl, Judea. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pearl, Judea. (2009). *Causality: Models, Reasoning, and Inference*. Second Edition. Cambridge: Cambridge University Press.

Pearl, Judea. (2001). "Direct and Indirect Effects." In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann, 411-420.

Robins, James M. (2001). "Data, Design, and Background Knowledge in Etiologic Inference," *Epidemiology 11* (3): 313-320.

Robins, James M. (2003). "Semantics of Causal DAG Models and the Identification of Direct and Indirect Effects." In: *Highly Structured Stochastic Systems*, P. Green, N. Hjort and S. Richardson, Eds. Oxford: Oxford University Press.

Robins, James M, and Sander Greenland. (1992). "Identifiability and Exchangeability for Direct and Indirect Effects." *Epidemiology 3*:143-155.

Sobel, Michael. (2008). "Identification of Causal Parameters in Randomized Studies with Mediating Variables," *Journal of Educational and Behavioral Statistics 33* (2): 230-251.

VanderWeele, Tyler J. (2008). "Simple Relations Between Principal Stratification and Direct and Indirect Effects." *Statistics and Probability Letters 78*: 2957-2962.

VanderWeele, Tyler J. and James M. Robins. (2007). "Four Types of Effect Modification: A Classification Based on Directed Acyclic Graphs." *Epidemiology 18* (5): 561-568.