5

EXTENSIONS

We turn next to some more technical statistical concerns often raised regarding interaction-term usage in regression analysis. The first issue regards separate-sample versus pooled-sample estimation of interactive effects. The second issue concerns estimation and interpretation of interaction terms in nonlinear models, including qualitative dependent-variable models like logit and probit models of binary outcomes. The third issue concerns modeling and estimating stochastically (rather than determinately) interactive relationships.[1]

Separate-Sample versus Pooled-Sample Estimation of Interactive Effects

Researchers often explore the interactive effects of nominal (binary, categorical, etc.) variables by splitting their samples according to these categories and estimating the same model separately in each subsample.[2] In behavioral research, for example, scholars may analyze interactive hypotheses that individual characteristics structure the impact of other variables by estimating the same model in subsamples separated by race,

[1] See Franzese (2005) for further, formal discussion of the first and third issues.

[2] Indeed, sometimes even ordinal or cardinal variables are separated into high(er) and low(er) categories for sample splitting in this manner. In addition to the considerations to be discussed in this section, this will typically entail inefficiency as the gradations of ordinal or cardinal information are discarded in the conversion to nominal categorization, although the practice may be justifiable in some cases on other grounds.

gender, and so on. A researcher might, for instance, estimate the effect of socioeconomic status on political participation separately in samples of male and female respondents to explore whether socioeconomic status affects the propensity to vote differently by gender. In comparative or international politics too, researchers might estimate the same model separately by country or region to explore whether national or regional contexts condition the effects of key variables. A political economist might, for instance, estimate a model of electoral cycles in monetary policy separately in subsamples of fixed- and flexible-exchange-rate country times. Similar subsample estimation strategies populate all subfields of political science and other social sciences.

Such subsample estimation (1) produces valid estimates of the (conditional) effects of the other variables at these different values of the "moderating" variable, (2) commendably recognizes the conditionality of the underlying arguments, and (3) can (perhaps with some effort) reproduce any of the efficiency and other desirable statistical properties of the alternative strategy of pooling with (nominal) interactions. However, these subsample procedures also isolate, at least presentationally, one variable as the moderator in what is logically a symmetric process—if x moderates the effect of z on y, then z moderates the effect of x on y and vice versa—thereby obscuring the converse. More fundamentally, these procedures do not facilitate statistical comparison of the effects of "moderated" or "moderating" variables; that is, one cannot as easily determine whether any differences in estimated effects across subsamples are statistically significant or as easily determine the (conditional) effects of the variable being treated as the moderating variable as one can in the pooling strategy.

An alternative approach is to estimate a model that keeps the subsamples together and that includes interaction terms of all of the other covariates, including the constant, with the variable being treated as the moderator; this is sometimes called a "fully dummy-interactive" model. The two approaches (separate sample versus fully dummy interactive pooled sample) extract almost identical sets of information from the data, but pooled-sample estimation extracts slightly more, potentially more efficiently, and more easily allows statistical testing of the full set of typical interactive hypotheses. That is, any desirable statistical properties that one can achieve by one strategy can, perhaps with considerable effort, be achieved by the other (see, e.g., Jusko and Shively 2005). However, we believe that the pooled interactive strategy lends itself more easily to obtaining these desirable qualities and, in some cases, also to presenting and interpreting results. Hence, we suggest that separate-sample estimation be reserved for exploratory and sensitivity and robustness consideration stages of analysis. We recommend pooled-sample approaches for final analysis and presentation. In either case, however, we note that theory should dictate the use of fully interactive (or separate-subsample) versus selectively interactive models. We do not advocate that fully interactive models or separate-sample models be used as a substitute for theoretically informed specifications. However, if a researcher is intent on "splitting the samples," then estimation using a fully interactive pooled model is a better alternative to separate-sample estimation.

As an example, a researcher, wishing to explore gender differences, *g*, in the effect of socioeconomic status and other independent variables, **X**, on propensity to vote, *y*, separates the sample into males and females and estimates

$$\textit{Sample } g = \textit{Male}: \mathbf{y}_{\mathrm{m}} = \mathbf{X}\boldsymbol{\beta}_{\mathrm{m}} + \mathbf{u}_{\mathrm{m}} \tag{37}$$

$$\textit{Sample } g = \textit{Female}: \mathbf{y}_{\mathrm{f}} = \mathbf{X}\boldsymbol{\beta}_{\mathrm{f}} + \mathbf{u}_{\mathrm{f}} \tag{38}$$

Let M (F) be the number of observations in the male (female) sample. Let k index the columns of **X** (e.g., $\mathbf{x}_{gk}$ represents the kth independent variable for the gender g sample; β_{gk} is the coefficient on that kth independent variable for that gender g sample) and let K be the number of independent variables (excluding the constant). To obtain distinct coefficient estimates by gender, the researcher has several options.

Most easily, the researcher could estimate models (37) and (38) separately, once per subsample. Or, he or she could pool the data into one sample and reconfigure the **X** matrix by manually creating separate $\mathbf{X}_{\mathrm{m}}$ and $\mathbf{X}_{\mathrm{f}}$ variables for each column of **X**, where $\mathbf{X}_{\mathrm{m}}$ replaces each female respondent's **X** value with zero and $\mathbf{X}_{\mathrm{f}}$ does so for male respondents. This allows distinct coefficients on $\mathbf{X}_{\mathrm{m}}$ and $\mathbf{X}_{\mathrm{f}}$ and, if the constant (intercept) is also separated into $\mathbf{X}_{\mathrm{m}}$ and $\mathbf{X}_{\mathrm{f}}$ in this way, will produce exactly the same coefficient estimates as separate-sample estimation does. Identically to this manual procedure, the researcher could simply create an indicator variable for $g_m = \textit{Male}$ and another indicator for $g_f = \textit{Female}$ and include these two indicators in place of the intercept and the interaction of each of these indicators with all of the other independent variables in place of those independent variables. Each $g_m\mathbf{X}$ and $g_f\mathbf{X}$ here will equal the $\mathbf{X}_{\mathrm{m}}$ and $\mathbf{X}_{\mathrm{f}}$ from the manual procedure just described, and so this also produces exactly the same coefficient estimates as the separate-sample estimation. Finally, the researcher could simply create one gender indicator, say, the female g_f, and include in the pooled-sample estimation all of the **X** independent variables (including the constant), unmodified, plus that g_f

indicator times each of these **X** variables (including the constant, which product just reproduces g_f). This, too, would produce the same substantive estimates for the model as separate-sample estimation, but the coefficients would now refer to different aspects of that substance. The coefficient on each variable x_k (including the intercept) in this last option would refer to the effect on y of that variable among males, whereas those coefficients on each x_k *plus the coefficient on the corresponding interaction term,* $g_f x_k$, would refer to the effect on y of that x_k among females. And the coefficient on $g_f x_k$ would refer to the difference in the effect of that x_k among females and the effect of that x_k among males. If all of these approaches produce the same substantive results from their estimates, why might researchers prefer one or the other of them?

In our review, researchers rarely offer reasons for presenting separate subsample estimations of interactive effects. Perhaps some do not realize that pooled-sample alternatives using interaction terms exist and, as we show next, are at least equivalent on all grounds except, perhaps, convenience. Others may note more explicitly that, lacking a priori hypotheses about what differences in the effects of the various x_k to expect across their subsamples, they wish simply to explore inductively what some possible candidates for interactive effects might be, and they find separate-sample estimation a convenient and easily interpreted means of conducting such exploration. The more technically savvy might even suggest that they did not wish to impose or estimate any distributional features of the residual term across subsamples, which would be necessary to validate statistical comparison of subsample coefficient estimates in pooled estimation.

In the separate-sample approach, researchers estimate one equation for males:

$$\begin{bmatrix} y_{1m} \\ \vdots \\ y_{Mm} \end{bmatrix} = \begin{bmatrix} 1 & X_{m11} & \cdots & X_{mK1} \\ & & & \\ \vdots & \vdots & & \vdots \\ 1 & X_{m1M} & \cdots & X_{mKM} \end{bmatrix} \begin{bmatrix} \beta_{m0} \\ \beta_{m1} \\ \\ \beta_{mK} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1m} \\ \vdots \\ \varepsilon_{Mm} \end{bmatrix} \tag{39}$$

and the exactly analogous equation for females. Table 26 provides OLS regression results from conducting this split-sample analysis (using our very simple *Support for Social Welfare* example). Typically, researchers will estimate these equations separately in each subsample and "eyeball" the results for differences in estimated β, which, assuming no other interactions, reflect directly the effect of the associated x in that subsample. This provides the often-cited ease of interpretation in separate-sample estimation. However, the second or third of the pooled-sample options described earlier (i.e., creating distinct $\mathbf{X}_f$ and $\mathbf{X}_m$ variables manually or by dummy-variable interaction) exactly replicates these separate-subsample coefficient estimates. If researchers prefer this sort of interpretability, pooled-sample estimation can also produce it. Presentationally, too, one can just as easily display two columns of coefficient estimates from one pooled-sample equation as from two separate-sample estimations. Therefore, direct interpretability of effects by subsample cannot adjudicate between pooled-sample and separate-sample approaches since one can present the same results in the same fashion regardless of whether those results derived from pooled-sample or separate-sample estimation.

Underlying any separate-sample estimation in the first place is at least the hunch that the effects of some independent variables differ across the categories distinguished by the subsamples. Thus, certainly, anyone conducting such analysis will wish to compare coefficient estimates across such subsamples. In table 26, a researcher might eyeball the differences in the estimated coefficient for *Republican* in the sample for males, $\hat{\beta}_R = -0.2205$, and in the sample for females, $\hat{\beta}_R = -0.1368$, and conclude (often by some unspoken or, worse, arbitrary standard) that these coefficients look "different enough." *If* classical OLS assumptions apply in

TABLE 26. OLS Regression Results, *Support for Social Welfare,* Pooled and Split Samples

	Pooled Sample Coefficient (standard error) *p*-Value	Males Only Coefficient (standard error) *p*-Value	Females Only Coefficient (standard error) *p*-Value
Female	−0.0031 (0.0144) *0.828*	—	—
Republican	−0.2205 (0.0155) *0.000*	−0.2205 (0.0154) *0.000*	−0.1368 (0.0148) *0.000*
Female × *Republican*	0.0837 (0.0214) *0.000*	—	—
Intercept	0.7451 (0.0110) *0.000*	0.7451 (0.0109) *0.000*	0.7420 (0.0094) *0.000*
N (df)	1,077 (1,073)	498 (496)	579 (577)
Adjusted R^2	0.223	0.290	0.128
$P > F$	0.000	0.000	0.000

Note: Cell entries are the estimated coefficient, with standard error in parentheses, and two-sided p-level (probability $|T| > t$) referring to the null hypothesis that $\beta = 0$ in italics.

each subsample (the OLS $\hat{\boldsymbol{\beta}}$ are the best linear unbiased estimates [BLUE]), then the researcher could test the statistical significance of any differences in parameters estimated separately across subsamples by difference tests of each $\hat{\beta}_f$ and corresponding $\hat{\beta}_m$:[3]

$$H_0\colon \beta_f = \beta_m \quad \text{or} \quad \beta_f - \beta_m = 0$$

Conducting the standard t-test of this null hypothesis:

$$\frac{(\hat{\beta}_f - \hat{\beta}_m) - 0}{s.e.\ (\hat{\beta}_f - \hat{\beta}_m)} = \frac{(\hat{\beta}_f - \hat{\beta}_m)}{\sqrt{\widehat{V(\hat{\beta}_f)} + \widehat{V(\hat{\beta}_m)} - 2\widehat{C(\hat{\beta}_f,\hat{\beta}_m)}}} = \frac{(\hat{\beta}_f - \hat{\beta}_m)}{\sqrt{\widehat{V(\hat{\beta}_f)} + \widehat{V(\hat{\beta}_m)}}} \tag{40}$$

The equality of the last expression to the previous two follows in this case, as it would not generally, because $\hat{\beta}_f$ and $\hat{\beta}_m$ will not covary due to the orthogonality of the gender indicators. Using our example, we would thus calculate $\left((\hat{\beta}_f - \hat{\beta}_m) - 0\right)/s.e.(\hat{\beta}_f - \hat{\beta}_m) = \left(-0.1368 - (-0.2205)\right)/\sqrt{(0.0148)^2 + (0.0154)^2} = 0.0837/0.0214 \approx 3.92$. The resulting t-test on this value suggests $p < 0.0001$: these estimated coefficients do appear to be statistically distinguishable from each other.

Few researchers in our review of the literature actually conducted this test; at best, they offered some reference to the *individual* standard errors of the two coefficient estimates in question. The subsample coefficient estimates will be independent by construction (the orthogonality of the indicator variables assures this), but the simple sum of the standard errors of the two coefficients is not the correct standard error for the estimated difference. The standard error of the estimated difference between the two coefficients is the square root of the sum of the estimated variances of the two coefficients. To conduct this comparison across subsamples of estimated effects, the reader should square the reported standard-error estimates, sum those variances, and square-root that sum.

Pooled-sample estimation allows a more directly interpretable formulation if the goal is to test whether effects differ across subsamples. Namely, with the right-hand side of the model specified as **X** and the nominal indicator(s) times **X**, the coefficient(s) on the interaction terms directly reveal the difference in effects across subsamples and the standard t-tests of those interaction-term coefficients directly reveal the statistical significance of those differences in effects.[4] A researcher seeking to deter-

[3] Researchers may also conduct the joint-hypothesis test that all of the coefficients are equal across subsamples, $H_0\colon \hat{\boldsymbol{\beta}}_{\mathrm{m}} = \hat{\boldsymbol{\beta}}_{\mathrm{f}}$, with a standard F-test: $(\hat{\boldsymbol{\beta}}_{\mathrm{m}} - \hat{\boldsymbol{\beta}}_{\mathrm{f}})'[\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}}_{\mathrm{m}} - \hat{\boldsymbol{\beta}}_{\mathrm{f}})}]^{-1}(\hat{\boldsymbol{\beta}}_{\mathrm{m}} - \hat{\boldsymbol{\beta}}_{\mathrm{f}}) \sim F_{k,n-k}$.

[4] Likewise, the standard F-test on the set of interaction terms tests whether the set of effects of **X** jointly differ across subsamples; see note 3.

mine whether the effect of *Republican* differs across females and males would need to calculate $\hat{\beta}_{R,females} - \hat{\beta}_{R,males} = 0.0837$ by subtracting the respective estimated coefficients acquired through separate-sample estimation. The pooled-sample estimation already provides this information, in the estimated coefficient, $\hat{\beta}_{FR} = 0.0837$. Further, instead of calculating the estimated standard error $s.e.(\hat{\beta}_f - \hat{\beta}_m)$ based on the two separate samples, per equation (40), the researcher can determine whether the difference in the effect of *Republican* between females and males is statistically distinguishable from zero by simply conducting a *t*-test using the results from the pooled-sample estimation: divide the estimated coefficient $\hat{\beta}_{FR}$ by its estimated standard error: $0.0837/0.0214 \approx 3.92$.

Thus, pooled-sample estimation offers two ways of presenting the same substantive results. One way replicates the same interpretability of coefficients as effects in subsamples afforded by separate-sample estimation. Another affords direct interpretation of coefficients as the estimated difference between effects across subsamples, as well as the standard *t*-tests or *F*-tests on those coefficients as revealing the statistical significance of that estimated difference. Pooled-sample estimation streamlines the process of testing the substantive hypotheses that researchers often seek to examine.

Moreover, pooling not only produces identical effect estimates as those obtained from separate samples, but it also (under classical linear regression model [CLRM] assumptions) constrains the variance of residuals, s^2, to be equal for the two samples and not to covary across subsamples. Separate-sample estimation makes no such assumptions; thus, pooled-sample estimation borrows strength from the other subsample(s) to obtain better (i.e., more efficient) standard error estimates, although only correctly so if these assumptions are true. Formally, these features are seen most directly for the case where **X** is arranged in block diagonal, either manually or by dummy-variable interactions:

$$\underset{(M+F)\times 1}{\begin{bmatrix} y_{m1} \\ \vdots \\ y_{mM} \\ y_{f1} \\ \vdots \\ y_{mF} \end{bmatrix}} = \underset{(M+F)\times(2K+2)}{\begin{bmatrix} 1 & X_{m11} & \cdots & X_{mK1} & 0 & \cdots & & 0 \\ \vdots & & & & & & & \\ & & & \vdots & \vdots & & \ddots & \\ 1 & X_{m1M} & & X_{mKM} & 0 & \cdots & & 0 \\ 0 & \cdots & & 0 & 1 & X_{f11} & \cdots & X_{fK1} \\ & 0 & & \vdots & & & & \\ \vdots & & \ddots & & \vdots & & & \vdots \\ 0 & \cdots & & 0 & 1 & X_{f1F} & & X_{fKF} \end{bmatrix}} \underset{2K+2}{\begin{bmatrix} \beta_{m0} \\ \beta_{m1} \\ \vdots \\ \beta_{mK} \\ \beta_{f0} \\ \beta_{f1} \\ \vdots \\ \beta_{fK} \end{bmatrix}} + \begin{bmatrix} \varepsilon_{m1} \\ \vdots \\ \varepsilon_{mM} \\ \varepsilon_{f1} \\ \varepsilon_{fF} \end{bmatrix} \quad (41)$$

Recall that $\hat{\boldsymbol{\Sigma}} = s^2\ (\mathbf{X}'\mathbf{X})^{-1}$. Since the $\mathbf{X}$ matrix here is block diagonal, the inverse will also be block diagonal, and the elements for males of $(\mathbf{X}'\mathbf{X})^{-1}$ and $\mathbf{X}'\mathbf{y}$, which comprise the coefficient estimate for males, $\hat{\boldsymbol{\beta}}_{\mathrm{m}} = (\mathbf{X}_{\mathrm{m}}'\mathbf{X}_{\mathrm{m}})^{-1}\,\mathbf{X}_{\mathrm{m}}'\mathbf{y}_{\mathrm{m}}$, are identical to what they would have been with the samples separated. The statistical test for the equality of the male and female coefficient estimates is then just the standard F-test on the equality of sets of two parameters ($\beta_f = \beta_m$). Note, though, that the single s^2 estimated here naturally differs from the two, s_m^2 and s_f^2, estimated separately in the subsample estimates. Pooled OLS assumes that s^2 is the same across the two samples. That one s^2 estimate, which is the average squared residual, sums squared residuals across the entire sample and divides by $N - j$ with the N reflecting the entire sample ($M + F$) and j reflecting all of the coefficients in the pooled estimation, including the constant. Separate-sample estimation produces a separate estimate of s^2 for each subsample (e.g., s_m^2 and s_f^2). Each separate-sample estimation sums only the squared residuals from its subsample and divides only by the number of observations in its subsample, minus the number of coefficients in the subsample estimation, $N_s - j_s$. The subsample estimates are inefficient. In other words, we obtain better estimates of s^2 and, with them, of estimated variance-covariances of the coefficient estimates in pooled-sample than in separate-sample estimation—if, indeed, the residual variances are equal across subsamples.[5] In this case, the inefficiency manifests as one of the s_m^2 and s_f^2 being larger than it needs to be and the other smaller than it should be. More generally, some of the s_i^2 will be larger than they need to be and others smaller than they should be. To explore whether such a common error-variance assumption is warranted, we can test whether heteroskedasticity instead prevails. If the data insist that heteroskedasticity exists, then one can model that variance (or variance-covariance) structure and employ weighted (or feasible generalized) least squares in the pooled sample.

Other model restrictions, such as constraining some coefficients to be equal across subsamples while allowing others to vary, are also easier to implement in pooled-sample estimation and will also, if true, enhance coefficient and standard-error estimates' efficiency. For example, we may posit, or theory may establish, that some x affects males' and females' voting propensities equally (or equally and oppositely, or otherwise relatedly in some deterministic manner). In some contexts, accounting or other mathematical identities may even require certain relations between

[5] In this case, the efficiency gains imply that estimated standard errors will be more accurate, not necessarily lower. As pooling borrows strength from the other subsamples to improve standard-error estimates, generally one (some) estimated effect(s) will be lower and (some) other(s) higher.

particular coefficients. Rather than estimate both of these effects separately, as separate-sample estimation all but requires,[6] one could in pooled-sample estimation simply refrain from including those dummy-variable interactions (or reverse the sign of those variables in the male or female sample, or analogously impose the constraints directly for other cases). As with a common-variance assumption, such cross-subsample restrictions can be tested, rather than assumed and imposed without testing, and again more conveniently in pooled-sample than in separate-subsample estimation. If the data insist that coefficients differ, this is easily allowed.

Thus, in short, compared to separate-sample estimation: (1) pooled-sample estimation can yield identical or superior interpretability; (2) it can encourage statistical comparison of effects over mere eyeballing; and (3) it may improve efficiency (precision) of estimation more easily if any efficiency-enhancing cross-subsample coefficient or error variance-covariance constraints are warranted. Therefore, if theory dictates that the effects of all variables should be dependent upon some *x*, we generally recommend that researchers present pooled-sample estimates as their final analysis—and report on the statistical certainty of any differences in effects they deem substantively important—even if they find conducting preliminary exploratory analysis in separate subsamples more convenient. We reiterate that while fully interactive pooled-sample estimation is preferable to separate-sample estimation, neither substitutes for a theoretically motivated model that identifies persuasively why the effect of some (set of) variable(s) should depend on *x*.[7]

Nonlinear Models

To this point, we have limited our discussion to interactive terms in linear models. However, we must also address interactions in nonlinear

[6] To our knowledge, only some relatively complicated iterative procedure, like MCMC (Markov chain Monte Carlo), could succeed in imposing that some $\hat{\beta}_m = \hat{\beta}_f$ across separate-subsample estimations, for example, and correctly gauge the statistical uncertainty of that single coefficient estimate.

[7] In multicategory cases, one can include **X**, indicators for all the categories except one, and all the interactions of the former with the latter, in which case the excluded category becomes the suppressed reference group that serves as the baseline for comparison. Standard *t*-tests would in this case refer to whether the effect in the category in question differs significantly from that base case for that category's indicator. Alternatively, one could block-diagonalize **X**, and then the coefficients would refer directly to the estimated effects of each *x* in each category, whereas tests of significance of any differences in estimated effects would require additional steps. In either case, one can interpret these interactive effects by calculating differences in predicted probabilities or derivatives (treating the derivatives of noncontinuous indicators as approximations).

models, which would include most models of qualitative dependent variables, given their prevalence in social science. For nonlinear models that include explicit linear-interactive terms among their right-hand-side variables, much of the discussion in preceding sections and chapters continues to apply. However, a further complication arises regarding the effect of *x* on *y* when right-hand-side variables are nonlinearly related to *y* by construction in the model. In logit or probit models of binary outcomes, for example, the effect of a variable *x* on *y* depends on the values of (all) the other variables **z** automatically due to the imposed nonlinear structure of the model. Thus, nonlinear models express conditional (i.e., interactive) relationships of the independent to dependent variables by construction, although they may also contain additional explicitly modeled linear interactions among the right-hand-side arguments of those nonlinear functions. The issue, then, arises regarding the proper interpretation of the effects of variables upon which a conditional relationship has been imposed, or assumed by construction, by virtue of the particular model specification employed.

Logit and probit models of dichotomous outcomes, for example, both (1) impose conditional relationships of *x* to *y* by construction and (2) use a sigmoidal (i.e., S-shaped) functional form implying specific character to those interactions. In these sigmoidal functional forms, the effects of changes in one variable on *y* are larger when the predicted probabilities are closer to the midpoint. Noting this, Nagler (1991), for example, critiques the claim of Wolfinger and Rosenstone (1980) that registration requirements discourage turnout to a greater extent among low education groups. He argues that this larger effect derives from the functional form assumed a priori and not necessarily from an explicit or direct interaction between education and registration requirements, for instance, that the less educated find surmounting registration requirements more difficult. The logit form by itself implies that education interacts with registration requirements and vice versa only because of and only through the other variable's effect on the overall propensity to vote. Insofar as being less educated puts one nearer a 0.5 probability of voting and being more educated puts one further from that point, registration requirements will have greater effect on the less educated's propensity to vote for that reason alone. Nagler tests whether education and registration requirements additionally interact explicitly to move a respondent along the S curve by including a specific linear interaction between education and registration requirement in the argument to a logit function. He also estimates logit coefficients on strict versus lax registration requirements separately in samples split by education (a strategy discussed earlier in this chapter). He finds little support for Wolfinger and Rosenstone's conclusion.[8]

The notion that multiple explicit interactions determine one's dependent variable suggests explicit modeling of those interactions, in as precise a fashion as theoretically possible. The defense for the specific form of interactivity implicit in logit, probit, and related models is, in fact, explicit and theoretical in this way. First, the logit and probit functional form implies a particular and very specific set of interactions to produce S shapes. That such S shapes should describe the relations of independent to dependent variables is substantively and theoretically derived from the proposition that inducing probabilities to increase or to decrease becomes increasingly difficult, that is, requires larger movements in independent variables, as probabilities near one or zero (see also note 8). If the researcher wishes to infer beyond the specific forms of interactions that produce these S shapes, we concur with Nagler (1991) that he or she must model those further interactions explicitly.

We now discuss in more detail the interpretation of effects in two commonly used nonlinear models: probit and logit. For example, suppose some nonlinear function, $F(\cdot)$, often called a "link function," is used to relate a binary outcome, y, with $\mathbf{x}'\boldsymbol{\beta}$, where y refers to a binary dependent variable, $\mathbf{x}'$ refers to a row vector of $k + 1$ regressors, and $\boldsymbol{\beta}$ refers

[8] Similarly, Frant (1991) reviews the research of Berry and Berry (1990) on state lottery policy adoptions. Frant argues that Berry and Berry draw their conclusions about the interaction between motivation, obstacles to innovation, and resources to overcome obstacles to innovation from the assumption inherent in the probit specification they employ. Berry and Berry (1991), however, disagree. They believe that their theory suggests that they estimate a probit model with no interactions or a linear probability model with a number of multiplicative terms. However, they prefer the probit model because the complexly interactive theory driving their model would require "so large a number of multiplicative terms as to render the model useless for empirical analysis because of extreme colinearity" (578). To argue that the complexly interactive nature of one's theory debars explicit modeling of it is a very weak defense by itself for applying an arbitrary specific functional form (probit) to allow all the independent variables to interact according to that specific functional form rather than explicitly to derive the form of these complex interactions from the theory. As we suggest and Frant (1991) notes, a stronger argument in defense would have been to demonstrate directly and explicitly that the theory implied specifically a set of interactions like those entailed inherently in a probit model, which indeed seems possible in this case. To generalize the example to a form common in many contexts, an argument might involve some overcoming of resistance from a broad set of conditions (explanatory factors) being necessary to produce an outcome. It might also then invoke some notion of a tipping point set by some values of this set of conditions and possibly even consider the outcome to become increasingly "overdetermined" as the factors all push for the outcome. Such an argument, which seems similar to Berry and Berry's, would indeed imply an S-shaped relation, such as logit or probit, between the explanatory factors and the outcome. Alternative sources or types of interactions, however, would not be inherent in sigmoid functions lacking those further, explicit interactions.

to a column vector of coefficients. In such a case, one could model the probability that y takes the value one as $p(y = 1) \equiv p = F(\mathbf{x}'\boldsymbol{\beta})$.

In the probit case, $p = \Phi(\mathbf{x}'\boldsymbol{\beta})$, where Φ is the cumulative standard-normal distribution. Cumulative normal distributions are S shaped, and so ever larger increases or decreases in $\mathbf{x}'\boldsymbol{\beta}$ are required to increase or decrease the probability $y = 1$ as this probability draws closer to one or zero. In the logit case, $p = \Lambda(\mathbf{x}'\boldsymbol{\beta})$, where $\Lambda(\cdot)$ is the logit function: $\Lambda(\mathbf{x}'\boldsymbol{\beta}) = e^{\mathbf{x}'\boldsymbol{\beta}}/(1 + e^{\mathbf{x}'\boldsymbol{\beta}})$ or, equivalently, $\Lambda(\mathbf{x}'\boldsymbol{\beta}) = [1 + e^{-\mathbf{x}'\boldsymbol{\beta}}]^{-1}$. (Several other useful formulations of the logit function also exist.)

We begin with a simple probability model that omits explicit interaction terms:

$$p = F(\mathbf{x}'\boldsymbol{\beta}) = F(\beta_0 + \beta_x x + \beta_z z + \cdots + \beta_k w)$$

As always, the marginal effects of a variable x on p can be calculated by taking the first derivative of this function.[9] Note here the use of the chain rule in differentiating the function:

$$\partial p/\partial x = [\partial p/\partial F(\mathbf{x}'\boldsymbol{\beta})][\partial F(\mathbf{x}'\boldsymbol{\beta})/\partial x]$$

$$\partial p/\partial x = [\partial p/\partial F(\mathbf{x}'\boldsymbol{\beta})][\partial F(\mathbf{x}'\boldsymbol{\beta})/\partial x]$$

In the probit case, using the same model that omits explicit interaction terms, this is simply

$$\begin{aligned} p &= \Phi(\mathbf{x}'\boldsymbol{\beta}) = \Phi(\beta_0 + \beta_x x + \beta_z z + \cdots + \beta_k w) \\ &\equiv \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} e^{-(\frac{1}{2})t^2}\, dt \quad \text{where } t \equiv \beta_0 + \beta_x x + \beta_z z + \cdots + \beta_k w \end{aligned}$$

$$\partial p/\partial x = [\partial \Phi(\mathbf{x}'\boldsymbol{\beta})/\partial x] = \phi(\mathbf{x}'\boldsymbol{\beta})\beta_x = \frac{1}{\sqrt{2\pi}} e^{-(\frac{1}{2})(\mathbf{x}'\boldsymbol{\beta})^2} \times \beta_x$$

where $\phi(\mathbf{x}'\boldsymbol{\beta})$ is the standard-normal probability density function evaluated at $\mathbf{x}'\boldsymbol{\beta}$.[10] Thus, as is central to the theoretical proposition of an S-shaped relationship, the magnitude of effects of x on the probability that $y = 1$ is largest at $p = 0.5$ (at $\mathbf{x}'\boldsymbol{\beta} = 0$) and becomes smaller, approaching zero, as that probability goes to one or zero (as $\mathbf{x}'\boldsymbol{\beta}$ approaches in-

[9] Note the distinction here between conceptualizing effects of a one-unit change in x literally computed (i.e., $\hat{p}_c - \hat{p}_a$) versus marginal effects, that is, effects of an infinitesimal change in x, $\partial y/\partial x$. Generally, the former is recommended for discrete variables and the latter for continuous variables. (See Greene 2003 for elaboration.)

[10] The derivative of any cumulative probability distribution function (cdf), F, is the corresponding probability density function (pdf), f, and so the derivative of Φ, the cdf of the standard normal, is ϕ, the pdf of the standard normal.

finity or negative infinity). One sees also that the effect of each x depends on itself and all of the other variables, since all the covariates and their coefficients appear in the $\phi(\mathbf{x}'\boldsymbol{\beta})$ that multiplies the coefficient on x to determine the effect of x.

In the logit case, again for this model omitting explicit interaction terms, this is simply

$$\begin{aligned} p &= \Lambda(\mathbf{x}'\boldsymbol{\beta}) = \left(1 + e^{-\mathbf{x}'\boldsymbol{\beta}}\right)^{-1} = \left(1 + e^{-(\beta_0+\beta_x x+\beta_z z+\cdots+\beta_k w)}\right)^{-1} \\ \partial p/\partial x &= \Lambda(\mathbf{x}'\boldsymbol{\beta})(1 - \Lambda(\mathbf{x}'\boldsymbol{\beta}))(\beta_x) \end{aligned} \tag{42}$$

In the specific model of equation (42), this would be

$$\begin{aligned} \frac{\partial p}{\partial x} &= \left(\frac{e^{\beta_0+\beta_x x+\beta_z z+\cdots+\beta_k w}}{\left(1 + e^{\beta_0+\beta_x x+\beta_z z+\cdots+\beta_k w}\right)^2}\right)\beta_x \\ &= \Lambda(\beta_0 + \beta_x x + \beta_z z + \cdots + \beta_k w) \\ &\quad \times [1 - \Lambda(\beta_0 + \beta_x x + \beta_z z + \cdots + \beta_k w)]\beta_x \end{aligned}$$

Obviously, as with probit, the effect of x depends on the values of x, $z, \ldots, w$ as well as the estimated coefficients for $\beta_0, \ldots, \beta_k$. We can also see, again as with probit, that the largest magnitude effects of x occur at $p = 0.5$, which occurs at $\mathbf{x}'\boldsymbol{\beta} = 0$, and that these effects become progressively smaller in magnitude as p approaches one or zero, which occurs as $\mathbf{x}'\boldsymbol{\beta}$ approaches positive or negative infinity, producing that familiar S shape again (although a slightly different S shape than probit produces).

When an explicit linear-interaction term (e.g., between x and z) is included in the $\mathbf{x}'\boldsymbol{\beta}$ part of the model, the effects of x continue to depend on the values of the other variables via the nonlinear form, specifically the S shape, of the model as before. In addition, movements along this S shape induced by movements in x depend directly on the value of z as well:

$$p = \Phi(\mathbf{x}'\boldsymbol{\beta}) = \int_{-\infty}^{\mathbf{x}'\boldsymbol{\beta}} \frac{1}{\sqrt{2\pi}} e^{-(½)t^2}\, dt \quad \text{where } t \equiv \begin{cases} \beta_0 + \beta_x x + \beta_z z \\ + \beta_{xz} xz + \cdots + \beta_k w \end{cases} \tag{43}$$

$$p = \Lambda(\mathbf{x}'\boldsymbol{\beta}) = \frac{e^{\beta_0+\beta_x x+\beta_z z+\beta_{xz} xz+\cdots+\beta_k w}}{1 + e^{\beta_0+\beta_x x+\beta_z z+\beta_{xz} xz+\cdots+\beta_k w}} \tag{44}$$

For illustration, we discuss a simple empirical example predicting turnout, using data from the 2004 National Election Studies. The dependent variable, *Voted*, is binary: 1 if the respondent voted; 0 if not. We model turnout as a function of two individual-level characteristics: education, ranging from one to seventeen years of *Schooling*, and strength of partisanship, *StrPID*, an ordinal measure equaling 0 for independents and 1 for leaning, 2 for weak, and 3 for strong partisans.[11] We interact education and strength of partisanship to explore whether education explicitly conditions the effect of strength of partisanship and vice versa. A researcher might argue that education and strength of partisanship each bring resources and motivation that reinforce each other in reducing the costs or increasing the benefits of voting, such that increases in one variable will boost the impact of the other in generating the net benefit to the individual of voting that relates nonlinearly (specifically: sigmoidally) to that individual's propensity to vote. Alternatively, the researcher might suspect the opposite: that educational and partisan resources and motivations undermine each other, such that increases in one variable contribute less to the net benefit of voting when the other is high than when it is low. Notice how these propositions argue something further than that the effect of one variable is higher or lower when the other is lower or higher *because both augment (detriment) the propensity to vote and so each has less effect when the other already leans the individual far toward or away from voting.* This last possibility is what the S-shaped function relating education and partisanship to vote propensity already assumes. Formally, we specify the following model (a fully specified model of turnout would, of course, include several additional covariates):

$$\begin{aligned} Voted = F(&\beta_0 + \beta_{Sch} Schooling + \beta_{Str} StrPID + \beta_{Sch\times Str} Schooling \\ &\times StrPID + \varepsilon) \end{aligned}$$

The logit and probit estimates appear in table 27.

The effects of x can be calculated using the derivative method or the method of differences in predicted probabilities. For the first-derivative approach, interpretation of a model with an explicit interaction in addition to its implicit ones would again require application of the chain rule. For logit:

$$\begin{aligned} \frac{\partial p}{\partial x} &= \frac{\partial p}{\partial \mathbf{x}'\boldsymbol{\beta}} \frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial x} = -(1 + e^{-\mathbf{x}'\boldsymbol{\beta}})^{-2} e^{-\mathbf{x}'\boldsymbol{\beta}}(-1)\frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial x} \\ &= \left(\frac{e^{\mathbf{x}'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}'\boldsymbol{\beta}}}\right)\left(\frac{1}{1 + e^{\mathbf{x}'\boldsymbol{\beta}}}\right)(\beta_x + \beta_{xz} z) \\ &= [\Lambda(\mathbf{x}'\boldsymbol{\beta})][1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})]\frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial x} = p(1 - p)(\beta_x + \beta_{xz} z) \qquad (45) \end{aligned}$$

[11] Here, as is common in such cases, we are treating the ordinal information on partisan leanings recorded by this measure as interval (or effectively interval, plus only some unimportant and unproblematic noise) by giving it simple linear coefficients in $\mathbf{x}'\boldsymbol{\beta}$.

This is the same expression as before except that now the effect of x depends not only on the other $\mathbf{x}$ through $[\Lambda(\mathbf{x}'\boldsymbol{\beta})][1 - \Lambda(\mathbf{x}'\boldsymbol{\beta})]$ but also and again on the value of z in the manner implied by the linear interaction of x and z contained in $\mathbf{x}$. Thus, z modifies the effect of x on p not only by its role in the calculation of $\Lambda(\mathbf{x}'\boldsymbol{\beta})$, where it enters in the $+\beta_z z + \beta_{xz} xz$ terms, but also in the final term, $\partial\mathbf{x}'\boldsymbol{\beta}/\partial x$, where it enters in the expression $\partial\mathbf{x}'\boldsymbol{\beta}/\partial x = \beta_x + \beta_{xz} z$. The former role is that imposed by the assumed sigmoidal relationship from independent to dependent variables; the latter role is imposed by the explicit interaction term as z conditions the effect of x on movement along that S shape.

Similarly, for the probit model, when there is an explicit interaction between x and z:

$$\frac{\partial p}{\partial x} = \frac{\partial p}{\partial \mathbf{x}'\boldsymbol{\beta}} \frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial x} = \phi(\mathbf{x}'\boldsymbol{\beta})\frac{\partial \mathbf{x}'\boldsymbol{\beta}}{\partial x} = \phi(\mathbf{x}'\boldsymbol{\beta})(\beta_x + \beta_{xz} z)$$

$$= \frac{1}{\sqrt{2\pi}} e^{-(½)(\mathbf{x}'\boldsymbol{\beta})^2} \times (\beta_x + \beta_{xz} z) \qquad (46)$$

In our example, the marginal effects of *Schooling* would be calculated at specific values of *Schooling* along varying values of *StrPID*, given as $\partial\hat{p}/\partial x = \hat{p}(1 - \hat{p})(\hat{\beta}_x + \hat{\beta}_{xz} z)$ in the logit case and $\partial\hat{p}/\partial x =$

TABLE 27. **Logit and Probit Regression Results, *Turnout***

	Logit Estimates Coefficient (standard error) *p*-Value	Probit Estimates Coefficient (standard error) *p*-Value
Years of Schooling	0.310	0.190
	(0.065)	(0.037)
	0.000	*0.000*
Strength of Partisanship	0.904	0.607
	(0.445)	(0.251)
	0.042	*0.015*
Years of Schooling × *Strength of Partisanship*	−0.021	−0.019
	(0.034)	(0.019)
	0.536	*0.321*
Intercept	−3.842	−2.340
	(0.852)	(0.489)
	0.000	*0.000*
$N(df)$	1,065 (1,062)	1,065 (1,062)
$\ln L$	−476.26	−476.04
$P > \chi^2$	0.000	0.000

Note: Cell entries are the estimated coefficient, with standard error in parentheses, and two-sided p-level (probability $|T| > t$) referring to the null hypothesis that $\beta = 0$ in italics.

$\phi(\mathbf{x}'\hat{\boldsymbol{\beta}})(\hat{\beta}_x + \hat{\beta}_{xz}z)$ in the probit case. Table 28 and table 29 provide the marginal effects of *Schooling* and *StrPID*, respectively, holding *Schooling* and *StrPID* at substantively interesting values. A sample calculation of the marginal effect of *Schooling*, when *Schooling* = 12 and *StrPID* = 3, using the logit results, is

$$\frac{\partial\hat{p}}{\partial Sch} = \left(\frac{e^{-3.84+0.31\times12+0.904\times3-0.021\times12\times3}}{1+e^{-3.84+0.31\times12+0.904\times3-0.021\times12\times3}}\right)\left(1-\frac{e^{-3.84+0.31\times12+0.904\times3-0.021\times12\times3}}{1+e^{-3.84+0.31\times12+0.904\times3-0.021\times12\times3}}\right)$$

$$\times\ (0.31 + -0.021 \times 3)$$

$$= (0.861)(1 - 0.861)(0.31 + -0.021 \times 3) \approx 0.029$$

Alternatively, one could calculate the predicted probabilities, $\hat{p}$, with appropriate confidence intervals. The intuition behind calculating the predicted probabilities in a nonlinear model is exactly the same as that behind calculating predicted values of y in a linear model. The nonlinear model merely requires an additional step, in projecting the linear index (i.e., the sum of the coefficients times their covariates) through the nonlinear model onto probability space (in the cases of logit and probit). For example, suppose we estimated the following relationship:

$$p = F(\beta_0 + \beta_x x + \beta_z z + \beta_{xz}xz)$$

Denote the predicted probabilities $\hat{F} = F(\mathbf{x}'\hat{\boldsymbol{\beta}})$, with the linear index, $\mathbf{x}'\hat{\boldsymbol{\beta}}$, computed in identical fashion to the linear-regression case:

$$\mathbf{x}'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_x x + \hat{\beta}_z z + \hat{\beta}_{xz}xz$$

TABLE 28. Marginal Effects of *Schooling,* Using Logit Results

	Independents	Leaning Partisans	Weak Partisans	Strong Partisans
Years of Schooling = 9	0.059 (0.007)	0.070 (0.008)	0.065 (0.010)	0.046 (0.016)
Years of Schooling = 12	0.077 (0.016)	0.067 (0.010)	0.048 (0.008)	0.029 (0.008)
Years of Schooling = 15	0.066 (0.011)	0.046 (0.005)	0.029 (0.003)	0.016 (0.003)

Note: Cell entries are the estimated marginal effect, with standard error in parentheses.

TABLE 29. Marginal Effects of *Strength of Partisanship,* Using Logit Results

	Independents	Leaning Partisans	Weak Partisans	Strong Partisans
Years of Schooling = 9	0.137 (0.017)	0.173 (0.035)	0.172 (0.035)	0.134 (0.018)
Years of Schooling = 12	0.162 (0.020)	0.152 (0.022)	0.117 (0.014)	0.078 (0.006)
Years of Schooling = 15	0.125 (0.032)	0.093 (0.022)	0.063 (0.011)	0.039 (0.004)

Note: Cell entries are the estimated marginal effect, with standard error in parentheses.

After calculation of the linear index, the researcher must use the link function, $F(\mathbf{x}'\hat{\boldsymbol{\beta}})$ (here, the logit $\Lambda(\mathbf{x}'\hat{\boldsymbol{\beta}})$ or probit $\Phi(\mathbf{x}'\hat{\boldsymbol{\beta}})$), to convert the linear index into probability space. In either case, the predicted probabilities would be calculated at various values of x (say, between x_a and x_c), holding z at some substantively meaningful and logically relevant value (e.g., its sample mean, $\bar{z}$) and of course allowing xz to vary from $x_a\bar{z}$ to $x_c\bar{z}$.

Thus, to calculate the effect on the predicted probability of a discrete change in x, say, from x_a and x_c, one would simply first compute the linear index at x_a and x_c:

$$(\mathbf{x}'\hat{\boldsymbol{\beta}})_a = \hat{\beta}_0 + \hat{\beta}_x x_a + \hat{\beta}_z \bar{z} + \hat{\beta}_{xz} x_a \bar{z};$$

$$(\mathbf{x}'\hat{\boldsymbol{\beta}})_c = \hat{\beta}_0 + \hat{\beta}_x x_c + \hat{\beta}_z \bar{z} + \hat{\beta}_{xz} x_c \bar{z}$$

Then one would project each linear index into probability space; for the logit case:

$$\hat{p}_a = \frac{e^{(\mathbf{x}'\hat{\boldsymbol{\beta}})_a}}{1 + e^{(\mathbf{x}'\hat{\boldsymbol{\beta}})_a}}; \qquad \hat{p}_c = \frac{e^{(\mathbf{x}'\hat{\boldsymbol{\beta}})_c}}{1 + e^{(\mathbf{x}'\hat{\boldsymbol{\beta}})_c}}$$

And then one simply computes the difference between the two probabilities: $\hat{p}_c - \hat{p}_a$. For probit, the process is identical except that one uses $\Phi(\mathbf{x}'\hat{\boldsymbol{\beta}})_a$ instead of $[1 + e^{-(\mathbf{x}'\hat{\boldsymbol{\beta}})_a}]^{-1}$, that is, the cumulative standard normal rather than the logit, as the link function.

We reiterate our strong recommendation that researchers compute and report measures of uncertainty around marginal effects and predicted probabilities. Standard errors for marginal effects can be computed by the delta method, as described in most statistics texts, for example, Greene (2003, 70).[12] The variance of any nonlinear function of parameter estimates, such as an estimated marginal effect like $\partial\hat{p}/\partial x$, is approximated asymptotically as a linear function of the estimated variance-covariance matrix of the parameter estimates, here $\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}$, and the derivative of the function with respect to $\hat{\boldsymbol{\beta}}$:[13]

[12] For confidence intervals around predicted levels, $\hat{p}$, a simpler expedient of calculating confidence intervals for the linear $\mathbf{x}'\hat{\boldsymbol{\beta}}$ and then translating those bounds to probability space using the link function will also suffice and, indeed, would have the advantage of constraining the confidence interval bounds to lie between zero and one, which the delta method's linearization strategy would not. That expedient would seem unavailable for confidence intervals around marginal effects and differences, however.

[13] The derivative of a function with respect to a vector of its arguments is called a gradient and denoted $\nabla_{\hat{\boldsymbol{\beta}}}$, but we eschew this terminology and notation as probably less familiar to many readers.

$$\widehat{V\left(\frac{\partial \hat{p}}{\partial x}\right)} \approx \left[\frac{\partial\left(\frac{\partial \hat{p}}{\partial x}\right)}{\partial \hat{\boldsymbol{\beta}}'}\right]\left[\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}\right]\left[\frac{\partial\left(\frac{\partial \hat{p}}{\partial x}\right)}{\partial \hat{\boldsymbol{\beta}}'}\right]' \tag{47}$$

We now apply this to the logit case, where $\hat{p} = (1 + e^{-\mathbf{x}'\hat{\boldsymbol{\beta}}})^{-1}$, $\partial\hat{p}/\partial x = \hat{p}(1 - \hat{p})\ \partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial x$.[14]

Next, using the product rule[15] to solve $[\partial(\partial\hat{p}/\partial x)/\partial\hat{\boldsymbol{\beta}}']$:

$$\begin{aligned}\partial(\partial\hat{p}/\partial x)/\partial\hat{\boldsymbol{\beta}}' &= [\partial(\partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial x)/\partial\hat{\boldsymbol{\beta}}'](\hat{p}(1 - \hat{p})) + [\partial\hat{p}/\partial\hat{\boldsymbol{\beta}}'] \\ &\quad \times ((1 - \hat{p})(\partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial x)) + [\partial(1 - \hat{p})/\partial\hat{\boldsymbol{\beta}}'](\hat{p}(\partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial x))\end{aligned}$$

Reexpressing terms, given that $\partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial x = \hat{\beta}_x + \hat{\beta}_{xz}z$: $\partial(\partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial x)/\partial\hat{\boldsymbol{\beta}}' = \partial(\hat{\beta}_x + \hat{\beta}_{xz}z)/\partial\hat{\boldsymbol{\beta}}'$. Let $\hat{\beta}_x + \hat{\beta}_{xz}z = \mathbf{r}'\hat{\boldsymbol{\beta}}$, where $\mathbf{r}' = [1\ \ 0\ \ z\ \ 0]$, assuming the estimated coefficients are arranged as $\hat{\boldsymbol{\beta}}' = [\hat{\beta}_x\ \hat{\beta}_z\ \hat{\beta}_{xz}\ \hat{\beta}_0]$, in that order. Differentiating: $\partial(\partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial x)/\partial\hat{\boldsymbol{\beta}}' = \partial\mathbf{r}'\hat{\boldsymbol{\beta}}/\partial\hat{\boldsymbol{\beta}}' = \mathbf{r}'$.

For the next term, $\partial\hat{p}/\partial\hat{\boldsymbol{\beta}}'$:

$$\partial\hat{p}/\partial\hat{\boldsymbol{\beta}}' = (\partial\hat{p}/\partial(\mathbf{x}'\hat{\boldsymbol{\beta}}))(\partial(\mathbf{x}'\hat{\boldsymbol{\beta}})/\partial\hat{\boldsymbol{\beta}}') = \frac{e^{-\mathbf{x}'\hat{\boldsymbol{\beta}}}}{(1 + e^{-\mathbf{x}'\hat{\boldsymbol{\beta}}})^2}\,\mathbf{x}' = \hat{p}(1 - \hat{p})\mathbf{x}'$$

And for the next term, $\partial(1 - \hat{p})/\partial\hat{\boldsymbol{\beta}}'$:

$$\begin{aligned}\frac{\partial(1 - \hat{p})}{\partial\hat{\boldsymbol{\beta}}'} &= \frac{\partial(1 - \hat{p})}{\partial(\mathbf{x}'\hat{\boldsymbol{\beta}})}\,\frac{\partial(\mathbf{x}'\hat{\boldsymbol{\beta}})}{\partial\hat{\boldsymbol{\beta}}'} = \frac{\partial\left(1 - (1 + e^{-\mathbf{x}'\hat{\boldsymbol{\beta}}})^{-1}\right)}{\partial(\mathbf{x}'\hat{\boldsymbol{\beta}})}\,\frac{\partial(\mathbf{x}'\hat{\boldsymbol{\beta}})}{\partial\hat{\boldsymbol{\beta}}'} \\ &= \frac{-e^{-\mathbf{x}'\hat{\boldsymbol{\beta}}}}{(1 + e^{-\mathbf{x}'\hat{\boldsymbol{\beta}}})^2}\mathbf{x}' = -\left(\hat{p}(1 - \hat{p})\mathbf{x}'\right)\end{aligned}$$

Substituting:

$$\begin{aligned}\partial(\partial\hat{p}/\partial x)/\partial\hat{\boldsymbol{\beta}}' &= \mathbf{r}'(\hat{p}(1 - \hat{p})) + [\hat{p}(1 - \hat{p})\mathbf{x}']((1 - \hat{p})(\partial\mathbf{x}'\boldsymbol{\beta}/\partial x)) \\ &\quad + [-(\hat{p}(1 - \hat{p})\mathbf{x}')](\hat{p}(\partial\mathbf{x}'\boldsymbol{\beta}/\partial x))\end{aligned}$$

Substituting $\partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial x = \hat{\beta}_x + \hat{\beta}_{xz}z$:

$$\partial(\partial\hat{p}/\partial x)/\partial\hat{\boldsymbol{\beta}}' = (\hat{p}(1 - \hat{p}))(\mathbf{r}' + (1 - 2\hat{p})(\hat{\beta}_x + \hat{\beta}_{xz}z)\mathbf{x}')$$

Then substituting into equation (47):

$$\begin{aligned}\widehat{V(\partial\hat{p}/\partial x)} &\approx (\hat{p}(1 - \hat{p}))((\mathbf{r}' + (1 - 2\hat{p})(\hat{\beta}_x + \hat{\beta}_{xz}z)\mathbf{x}'))\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}(\hat{p}(1 - \hat{p})) \\ &\quad \times ((\mathbf{r}' + (1 - 2\hat{p})(\hat{\beta}_x + \hat{\beta}_{xz}z)\mathbf{x}'))' \\ &= (\hat{p}(1 - \hat{p}))^2(\mathbf{r}' + (1 - 2\hat{p})(\hat{\beta}_x + \hat{\beta}_{xz}z)\mathbf{x}')\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})} \\ &\quad \times (\mathbf{r} + (1 - 2\hat{p})(\hat{\beta}_x + \hat{\beta}_{xz}z)\mathbf{x})\end{aligned}$$

[14] In the simple case that contains no explicit interaction, $\partial\hat{p}/\partial x = \hat{p}(1 - \hat{p})\ (\partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial \mathrm{x}) = \hat{p}(1 - \hat{p})\hat{\beta}_x$. When x interacts with another variable, z, as in equation (44), then $\partial\hat{p}/\partial x = \hat{p}(1 - \hat{p})(\hat{\beta}_x + \hat{\beta}_{xz}z)$.

[15] Recall that $\partial(f(x)g(x))/\partial x = \partial f(x)/\partial x\ g(x) + \partial g(x)/\partial x\ f(x)$.

Note that $(1 - 2\hat{p})(\hat{\beta}_x + \hat{\beta}_{xz}z)$ is a scalar for a given set of values of x, z, and xz that scales the values in vector $\mathbf{x}'$. Let s_L be the value of the scaling value in the logit: $s_L = (1 - 2\hat{p})(\hat{\beta}_x + \hat{\beta}_{xz}z)$:

$$\begin{aligned}\widehat{V(\partial\hat{p}/\partial x)} &\approx (\hat{p}(1 - \hat{p}))^2(\mathbf{r}' + s_L\mathbf{x}')\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}(\mathbf{r} + s_L\mathbf{x})\\ &= (\hat{p}(1 - \hat{p}))^2\left(\mathbf{r}'\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}\mathbf{r} + 2s_L\mathbf{x}'\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}\mathbf{r} + s_L^2\mathbf{x}'\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}\mathbf{x}\right)\end{aligned}$$

Using our empirical example, we can calculate the estimated variance around the estimated marginal effect of *Schooling*, when *Schooling* = 12 and *StrPID* = 3. In this example, $\mathbf{x}' = [12\ \ 3\ \ 36\ \ 1]$; the value at which *Schooling* is held is located in the first column; the value at which *StrPID* is held is in the second column; the interaction term's value appears in the third column; and a 1 is located in the last column, to represent the intercept. We established previously that $(\hat{p} \mid Sch = 12, Str = 3) = 0.861$. Because we are taking $\partial\hat{p}/\partial x$ with respect to *Schooling*, and because the value of *StrPID* is 3, $\mathbf{r}' = [1\ \ 0\ \ 3\ \ 0]$. As with linear regression, the estimated variance-covariance matrix of the estimated logit or probit coefficients can be easily called by a postestimation command. In this case,

$$\mathbf{V}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 0.004 & 0.024 & -0.002 & -0.055\\ 0.024 & 0.198 & -0.015 & -0.323\\ -0.002 & -0.015 & 0.001 & 0.024\\ -0.055 & -0.323 & 0.024 & 0.726 \end{bmatrix},$$

a 4 × 4 matrix that lists the estimated coefficient variances and covariances in the order in which they appear in the regression results and corresponding with the order in which values are arrayed in $\mathbf{x}'$. Substituting the set values in $\mathbf{x}'$, the values in $\mathbf{r}'$, and the estimated values for $\hat{p}$ and $\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}$:

$$\begin{aligned}\widehat{V(\partial\hat{p}/\partial x)} \approx{}& (0.861(1 - 0.861))^2\\ &\times\left(\begin{aligned} &[1\ \ 0\ \ 3\ \ 0]\begin{bmatrix} 0.004 & 0.024 & -0.002 & -0.055\\ 0.024 & 0.198 & -0.015 & -0.323\\ -0.002 & -0.015 & 0.001 & 0.024\\ -0.055 & -0.323 & 0.024 & 0.726 \end{bmatrix}\begin{bmatrix}1\\0\\3\\0\end{bmatrix}\\ &+2s_L[12\ \ 3\ \ 36\ \ 1]\begin{bmatrix} 0.004 & 0.024 & -0.002 & -0.055\\ 0.024 & 0.198 & -0.015 & -0.323\\ -0.002 & -0.015 & 0.001 & 0.024\\ -0.055 & -0.323 & 0.024 & 0.726 \end{bmatrix}\begin{bmatrix}1\\0\\3\\0\end{bmatrix}\\ &+s_L^2[12\ \ 3\ \ 36\ \ 1]\begin{bmatrix} 0.004 & 0.024 & -0.002 & -0.055\\ 0.024 & 0.198 & -0.015 & -0.323\\ -0.002 & -0.015 & 0.001 & 0.024\\ -0.055 & -0.323 & 0.024 & 0.726 \end{bmatrix}\begin{bmatrix}12\\3\\36\\1\end{bmatrix}\end{aligned}\right)\end{aligned}$$

where $s_L = (1 - 2 \times 0.861)(0.31 - 0.021 \times 3)$. A standard statistical package or a spreadsheet program can easily perform these calculations.

Similarly for the probit case, standard errors around marginal effects are calculated following equation (47); specifying $\hat{p} = \Phi(\mathbf{x}'\hat{\boldsymbol{\beta}})$, we have $\partial\hat{p}/\partial x = \phi(\mathbf{x}'\hat{\boldsymbol{\beta}})\partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial x$. Using the product rule, $\partial(\partial\hat{p}/\partial x)/\partial\hat{\boldsymbol{\beta}}' = [\partial(\partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial x)/\partial\hat{\boldsymbol{\beta}}']\phi(\mathbf{x}'\hat{\boldsymbol{\beta}}) + (\partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial x)[\partial\phi(\mathbf{x}'\hat{\boldsymbol{\beta}})/\partial\hat{\boldsymbol{\beta}}']$. Reexpressing the first term in brackets: $\partial(\partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial x)/\partial\hat{\boldsymbol{\beta}}' = \mathbf{r}'$. The second term in brackets is $\partial\phi(\mathbf{x}'\hat{\boldsymbol{\beta}})/\partial\hat{\boldsymbol{\beta}}' = (\partial\phi(\mathbf{x}'\hat{\boldsymbol{\beta}})/\partial(\mathbf{x}'\hat{\boldsymbol{\beta}}))(\partial(\mathbf{x}'\hat{\boldsymbol{\beta}})/\partial\hat{\boldsymbol{\beta}}') = (1/\sqrt{2}\pi \;\; e^{-(½)(\mathbf{x}'\hat{\boldsymbol{\beta}})^2})(-\mathbf{x}'\hat{\boldsymbol{\beta}})(\partial(\mathbf{x}'\hat{\boldsymbol{\beta}})/\partial\hat{\boldsymbol{\beta}}') = -(\phi(\mathbf{x}'\hat{\boldsymbol{\beta}}))(\mathbf{x}'\hat{\boldsymbol{\beta}})\mathbf{x}'$. Substituting into $\partial(\partial\hat{p}/\partial x)/\partial\hat{\boldsymbol{\beta}}'$: $\partial(\partial\hat{p}/\partial x)/\partial\hat{\boldsymbol{\beta}}' = [\mathbf{r}']\phi(\mathbf{x}'\hat{\boldsymbol{\beta}}) - (\partial\mathbf{x}'\hat{\boldsymbol{\beta}}/\partial x)(\phi(\mathbf{x}'\hat{\boldsymbol{\beta}}))(\mathbf{x}'\hat{\boldsymbol{\beta}})\mathbf{x}' = \phi(\mathbf{x}'\hat{\boldsymbol{\beta}})(\mathbf{r}' - (\hat{\beta}_x + \hat{\beta}_{xz}z)(\mathbf{x}'\hat{\boldsymbol{\beta}})\mathbf{x}')$. Then substituting into equation (47):

$$\begin{aligned}\widehat{V(\partial\hat{p}/\partial x)} &\approx [(\phi(\mathbf{x}'\hat{\boldsymbol{\beta}}))(\mathbf{r}' - \mathbf{x}'\hat{\boldsymbol{\beta}}(\hat{\beta}_x + \hat{\beta}_{xz}z)\mathbf{x}')]\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}[(\phi(\mathbf{x}'\hat{\boldsymbol{\beta}})) \\ &\quad\times (\mathbf{r}' - \mathbf{x}'\hat{\boldsymbol{\beta}}(\hat{\beta}_x + \hat{\beta}_{xz}z)\mathbf{x}')]' \\ &= (\phi(\mathbf{x}'\hat{\boldsymbol{\beta}}))^2\, (\mathbf{r}' - \mathbf{x}'\hat{\boldsymbol{\beta}}(\hat{\beta}_x + \hat{\beta}_{xz}z)\mathbf{x}')\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}[(\mathbf{r} - \hat{\boldsymbol{\beta}}'\mathbf{x}(\hat{\beta}_x + \hat{\beta}_{xz}z)\mathbf{x})]\end{aligned}$$

Again, note that $\mathbf{x}'\hat{\boldsymbol{\beta}}(\hat{\beta}_x + \hat{\beta}_{xz}z)$ is a scalar for a given set of values of x, z, and xz. Let $s_P = \mathbf{x}'\hat{\boldsymbol{\beta}}(\hat{\beta}_x + \hat{\beta}_{xz}z)$. Substituting:

$$\begin{aligned}\widehat{V(\partial\hat{p}/\partial x)} &\approx (\phi(\mathbf{x}'\hat{\boldsymbol{\beta}}))^2\, (\mathbf{r}' - s_P\mathbf{x}')\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}(\mathbf{r} - s_P\mathbf{x}) \\ &= (\phi(\mathbf{x}'\hat{\boldsymbol{\beta}}))^2\, (\mathbf{r}'\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}\mathbf{r} - 2s_P\mathbf{x}'\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}\mathbf{r} + s_P^2\mathbf{x}'\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}\mathbf{x})\end{aligned}$$

For standard errors around predicted probabilities, we can also use the delta method. In the logit case, $\widehat{V(\hat{p})} \approx [\partial\hat{p}/\partial\hat{\boldsymbol{\beta}}]'\,[\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}][\partial\hat{p}/\partial\hat{\boldsymbol{\beta}}] = [\hat{p}(1 - \hat{p})\mathbf{x}'][\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}][\hat{p}(1 - \hat{p})\mathbf{x}] = (\hat{p}(1 - \hat{p}))^2\,\mathbf{x}'[\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}]\mathbf{x}$. That is, square $\hat{p}(1 - \hat{p})$ and multiply the result by the estimated variance-covariance matrix of the estimated coefficients, pre- and postmultiplied by the $\mathbf{x}$ vector specified at the values of interest. In the probit case, $\widehat{V(\hat{p})} \approx [-(\phi(\mathbf{x}'\hat{\boldsymbol{\beta}}))\mathbf{x}]'[\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}][-\,(\phi(\mathbf{x}'\hat{\boldsymbol{\beta}}))\mathbf{x}] = (\phi(\mathbf{x}'\hat{\boldsymbol{\beta}}))^2\,\mathbf{x}'[\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}]\mathbf{x}$. As with linear-regression models, predicted probabilities are most effective presentationally when graphed with confidence intervals. Confidence intervals can be generated using the same formulas: $\hat{p} \pm t_{df,p}\sqrt{\widehat{V(\hat{p})}}$.

Calculation of the standard error for the difference between two predicted probabilities, say, those reflecting the effect of a specific change in x from x_a to x_c, follows the same delta method:

$$\begin{aligned}\widehat{V(\hat{F}_c - \hat{F}_a)} \approx \left[\frac{\partial(\hat{F}_c - \hat{F}_a)}{\partial\hat{\boldsymbol{\beta}}}\right]'\left[\widehat{V(\hat{\boldsymbol{\beta}})}\right]\left[\frac{\partial(\hat{F}_c - \hat{F}_a)}{\partial\hat{\boldsymbol{\beta}}}\right] &= \left[\frac{\partial\hat{F}_c}{\partial\hat{\boldsymbol{\beta}}} - \frac{\partial\hat{F}_a}{\partial\hat{\boldsymbol{\beta}}}\right]'\left[\widehat{V(\hat{\boldsymbol{\beta}})}\right] \\ &\times \left[\frac{\partial\hat{F}_c}{\partial\hat{\boldsymbol{\beta}}} - \frac{\partial\hat{F}_a}{\partial\hat{\boldsymbol{\beta}}}\right] \\ &= [\hat{f}_c\mathbf{x}_c' - \hat{f}_a\mathbf{x}_a']\left[\widehat{V(\hat{\boldsymbol{\beta}})}\right][\hat{f}_c\mathbf{x}_c - \hat{f}_a\mathbf{x}_a]\end{aligned}$$

Here $\hat{F}_a$ and $\hat{F}_c$ are the link function (logit or probit here), and $\hat{f}_a$ and $\hat{f}_c$ are their derivatives with respect to $\mathbf{x}'\hat{\boldsymbol{\beta}}$, ($\hat{p}(1 - \hat{p})$ for logit and $\phi(\mathbf{x}'\hat{\boldsymbol{\beta}})$ for probit). These link functions and derivatives are evaluated at $\mathbf{x}_a$ and $\mathbf{x}_c$, respectively.

Many existing statistical software packages will calculate these standard errors of estimated probabilities for the researcher, and some will even calculate standard errors for derivatives or differences at user-given levels of the variables. Our intention here is to reemphasize the importance of examining effects rather than simply coefficients (or predicted levels), be they estimated in a linear or nonlinear specification, and to provide readers with a sense of the mathematics underlying the calculation of these estimated effects and their corresponding standard errors.

Random-Effects Models and Hierarchical Models

When modeling relationships between a set of covariates, **X**, and a dependent variable, y, scholars make assumptions about the deterministic (i.e., fixed) versus stochastic (i.e., random) nature of those relationships. In the interaction context, for example, scholars might propose that the effects of x and of z on y depend either deterministically or stochastically on the other variable. The burgeoning "random-effects" literature proposes the latter, probabilistic, relationship. (The related multilevel-model or hierarchical-model literature addresses a similar issue, although possibly with different assumptions about the properties of the stochastic aspects of the relationships: see the discussion that follows.)[16]

Let us start thus:

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon \tag{48}$$

As before, the linear-interactive specification of the posited interactive relationships could be

$$\beta_0 = \gamma_0 + \gamma_1 x + \gamma_2 z, \qquad \beta_1 = \delta_1 + \delta_2 z, \qquad \text{and} \quad \beta_2 = \delta_3 + \delta_4 x \tag{49}$$

[16] For more thorough discussion of the issues in this section, see Franzese (2005).

in the deterministic case, suggesting our standard linear-interactive regression model:

$$y = \gamma_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon \tag{50}$$

where $\beta_x = \gamma_1 + \delta_1$, $\beta_z = \gamma_2 + \delta_3$, $\beta_{xz} = \delta_2 + \delta_4$. Notice, however, that this standard model in fact assumes that the effect of x on y varies with z, and the effect of z on y varies with x, *without error.* Likewise, the intercept does not vary across repeated samples. A linear-interactive model with random effects would instead be

$$\beta_0 = \gamma_0 + \gamma_1 x + \gamma_2 z + \varepsilon_0, \qquad \beta_1 = \delta_1 + \delta_2 z + \varepsilon_1, \qquad \text{and}$$
$$\beta_2 = \delta_3 + \delta_4 x + \varepsilon_2 \tag{51}$$

suggesting the following similar-looking linear-interactive regression model:

$$y = \gamma_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon^* \tag{52}$$

but with $\varepsilon^* = \varepsilon + \varepsilon_0 + \varepsilon_1 x + \varepsilon_2 z$.

Thus, the distinction between the deterministically interactive and the stochastically interactive models occurs only in the "error" term; the two models are identical except for the difference between ε and ε^*. In the first case, where the conditioning effects are assumed to be deterministic, OLS would be BLUE, that is, yielding the best (most efficient), linear unbiased estimates (provided the model is also correctly specified in other regards, of course). In the latter case, where effects are assumed stochastic, or probabilistic, one suspects that OLS estimates might not be BLUE. Notice, however, that, assuming all the stochastic terms have mean zero, $E(\varepsilon, \varepsilon_0, \varepsilon_1, \varepsilon_2) = \mathbf{0}$, and do not covary with the regressors, $C(\{\varepsilon, \varepsilon_0, \varepsilon_1, \varepsilon_2\}, \mathbf{x}) = \mathbf{0}$, as commonly done in most regression contexts including random effects/hierarchical modeling, OLS estimation would still yield unbiased and consistent coefficient estimates.[17] On the other hand, the composite residual's variance, $V(\varepsilon^*)$, is not constant (homoskedastic) but differs (heteroskedastic) across observations, even if $V(\varepsilon) \ldots V(\varepsilon_2)$ are each constant, rendering coefficient estimates and standard errors inefficient. Moreover, this nonconstant variance moves with the values of x and z, which implies that the standard-error estimates (but not the coefficient estimates) are biased and inconsistent as well. Thus, even if the error components in the random-effects model have constant variance,

[17] $E(\hat{\boldsymbol{\beta}}_{ols}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \varepsilon^*)) = \boldsymbol{\beta} + E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon^*) = \boldsymbol{\beta} + 0 = \boldsymbol{\beta}$ if each component of ε^* has mean zero and does not covary with $\mathbf{x}$. See Franzese (2005) for a fuller discussion of the proof.

mean zero, and no correlation with regressors, as we would commonly assume, OLS coefficient estimates will be inefficient, and OLS standard-error estimates will be biased, inconsistent, and inefficient. These problems, though potentially serious, are probably small in magnitude in most cases and, anyway, easy to redress by simple techniques with which political scientists are already familiar.

As mentioned before, similar issues arise in the literature on hierarchical, or multilevel, models (see, e.g., Bryk and Raudenbush 2001; Kedar and Shively 2005; Steenbergen and Jones 2002). Often these models propose that some unit-level y_{ij} depends on a contextual-level variable, z_j, varying only across and not within the j contexts, and a unit-level variable, x_{ij}, and furthermore that the effect of the unit-level variable x_{ij} depends (deterministically or stochastically) on z_j:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + \varepsilon_{ij} \qquad (53)\\ \beta_0 &= \gamma_0 (+\varepsilon_{0ij})\\ \beta_1 &= \delta_1 + \delta_2 z_j\ (+\varepsilon_{1j})\\ \beta_2 &= \delta_3 + \delta_4 x_{ij}\ (+\varepsilon_{2ij}) \end{aligned}$$

which implies that one may model y for regression analysis as

$$y = \gamma_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon^* \qquad (54)$$

where $\varepsilon^* = \varepsilon_{ij}\ (+\varepsilon_{0ij} + \varepsilon_{1j} x_{ij} + \varepsilon_{2ij} z_j)$ and the coefficients remain identical to those given previously.

Assuming deterministic conditional relationships so that $\varepsilon^* = \varepsilon_{ij}$, that is, the parenthetical terms are all zero, and assuming that this simple residual is well behaved (mean zero, constant variance, and no correlation with regressors, as usual), OLS is BLUE. If, instead, ε_{ij} exhibits heteroskedasticity and/or correlation across i or j, then OLS coefficient and standard-error estimates would be unbiased and consistent but inefficient in the case that the patterns of these nonconstant variances and/or correlations were themselves uncorrelated with the regressors, their cross-products, and their squares. In the case that these patterns correlated in some fashion with the regressors, their cross-products, or their squares, OLS coefficient estimates would still be unbiased and consistent but inefficient, but OLS standard errors would be biased and inconsistent as well as inefficient in this context. These standard-error inconsistency problems could be redressed in a familiar manner by replacing the OLS formula for estimating the variance-covariance of estimated coefficients with a heteroskedasticity-consistent formula like White's or the appropriate heteroskedasticity-and-correlation-consistent formula, like Newey-West for temporal correlation, Beck-Katz for contemporaneous (spatial) correlation, or "cluster" for the case of common stochastic shocks to all units *i* in each context *j*.

With stochastic dependence such that $\varepsilon^* = \varepsilon_{ij} + \varepsilon_{0ij} + \varepsilon_{1j}x_{ij} + \varepsilon_{2ij}z_j$, on the other hand, OLS coefficient estimates are still unbiased and consistent, but the error term presents us with two issues even in the case of well-behaved ε_{ij}: heteroskedasticity (the composite residual term, ε^*, varies; in fact, it varies depending on some linear combination of x and z) as well as potentially severe autocorrelation (each ε_{1j} will be common to all individuals *i* in context *j*).[18]

Thus, the random-effects and multilevel (hierarchical) cases produce identical problems in OLS, and so the same solutions will apply. Note first that some form of the familiar White or Huber-White consistent variance-covariance estimators, that is, "robust" standard errors, will redress the inconsistency in OLS estimates of the estimated coefficients' variance-covariance, that is, $\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}_{ols}$.

Recall that, given nonspherical disturbances,

$$\begin{aligned}\mathbf{V}(\hat{\boldsymbol{\beta}}) &= E[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'] = E[\{\boldsymbol{\beta} -(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\boldsymbol{\varepsilon})\} \\ &\qquad \times \{\boldsymbol{\beta} - (\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon})\}'] \\ &= E[\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\}\{(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\boldsymbol{\varepsilon}\}'] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,[E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,[V(\boldsymbol{\varepsilon})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad (55)\end{aligned}$$

Under classical linear-regression assumptions, $\boldsymbol{\varepsilon} \sim \mathbf{N}(0, \sigma^2\mathbf{I})$ and $E(\boldsymbol{\varepsilon}'\mathbf{X}) = 0$, and so this reduces to

[18] Some current literature even suggests that OLS is biased in the presence of such multilevel random effects. This is false if *biased* refers to the OLS coefficient estimates. Provided that the context-specific or other components of the composite error term do not correlate with the regressors, OLS coefficient estimates will remain unbiased and consistent, although inefficient. The fact that Z_j and ε_j are both common to all individuals in context *j* implies that the pattern of the nonsphericity in the composite $V(\varepsilon^*)$ relates to a regressor, *Z*, producing biased, inconsistent, and inefficient OLS standard-error estimates, but that does not imply that $C(Z_j,\varepsilon^*)$ is nonzero, which is the condition that would bias OLS coefficient estimates. The "problem" with OLS for hierarchical models therefore resides solely in the inefficiency of OLS coefficient estimates and in the generally poor properties of the OLS estimates of $\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})}$. The problem is similar to that typically induced by strong temporal or spatial correlation: OLS coefficient estimates are unbiased and consistent but inefficient; standard errors are biased, inconsistent, and inefficient. The inefficiency in coefficient estimates can be dramatic if the within-context correlation of individual errors is great, perhaps dramatic enough to render unbiasedness and consistency of little practical comfort, but, even so, the problem is efficiency, not bias or inconsistency.

$$\begin{aligned}\mathbf{V}(\hat{\boldsymbol{\beta}}) &= E[(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[V(\boldsymbol{\varepsilon})](\mathbf{X}'\mathbf{X})^{-1}\mathbf{X} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

With random effects, $\varepsilon^* = \varepsilon + \varepsilon_0 + \varepsilon_1 x + \varepsilon_2 z$; in multilevel data, $\varepsilon^* = \varepsilon_{ij} + \varepsilon_{0ij} + \varepsilon_{1j}x_{ij} + \varepsilon_{2ij}z_j$. Both violate the assumptions of classical linear regression in essentially the same way. In our random-coefficient case:

$$\begin{aligned}E(\boldsymbol{\varepsilon}^*\boldsymbol{\varepsilon}^{*\prime}) &= E(\varepsilon + \varepsilon_0 + \varepsilon_1 x + \varepsilon_2 z)(\varepsilon + \varepsilon_0 + \varepsilon_1 x + \varepsilon_2 z)' \\ &= E\left(\begin{array}{l}\varepsilon\varepsilon' + \varepsilon_0\varepsilon' + \varepsilon_1 x\varepsilon' + \varepsilon_2 z\varepsilon' + \varepsilon\varepsilon_0' + \varepsilon_0\varepsilon_0' + \varepsilon_1 x\varepsilon_0' \\ + \varepsilon_2 z\varepsilon_0' + \varepsilon x'\varepsilon_1' + \varepsilon_0 x'\varepsilon_1' + \varepsilon_1 xx'\varepsilon_1' + \varepsilon_2 zx'\varepsilon_1' \\ + \varepsilon z'\varepsilon_2' + \varepsilon_0 z'\varepsilon_2' + \varepsilon_1 xz'\varepsilon_2' + \varepsilon zz'\varepsilon_2'\end{array}\right) \qquad (56)\end{aligned}$$

Even assuming that $(\varepsilon, \varepsilon_0, \varepsilon_1, \varepsilon_2)$ are independently and identically distributed (i.i.d.) $N(\mathbf{0}, \sigma^2\mathbf{I})$, the variance-covariance matrix for $\hat{\boldsymbol{\beta}}$ in the random coefficient model will be

$$\mathbf{V}(\hat{\boldsymbol{\beta}}_{RC}) = 2\sigma^2 + \mathbf{xx}'\sigma^2 + \mathbf{zz}'\sigma^2 = \sigma^2(2\mathbf{I} + \mathbf{xx}' + \mathbf{zz}') \qquad (57)$$

In the hierarchical model, the basic structure is the same, but the claim that $(\varepsilon, \varepsilon_0, \varepsilon_1, \varepsilon_2)$ would be i.i.d. is less plausible because, among other reasons, context-level variance (ε_{1j}) is unlikely to equal unit-level variances $(\varepsilon_{ij}, \varepsilon_{0ij}, \varepsilon_{2ij})$. It is more plausible to assume that between-level variation differs but within-level variation is constant. If so, the variance-covariance of $\hat{\boldsymbol{\beta}}$ in the hierarchical case is

$$\begin{aligned}\mathbf{V}(\hat{\boldsymbol{\beta}}_{HM}) &= 2\sigma^2_{ind} + \mathbf{xx}'\sigma^2_{context} + \mathbf{zz}'\sigma^2_{ind} \\ &= \sigma^2_{ind}\,(2\mathbf{I} + \mathbf{zz}') + \mathbf{xx}'\sigma^2_{context} \qquad (58)\end{aligned}$$

Notice that the expressions for $\mathbf{V}(\hat{\boldsymbol{\beta}}_{HM})$ in the hierarchical case and for $\mathbf{V}(\hat{\boldsymbol{\beta}}_{RC})$ in the random-coefficient case are almost identical. The only difference is the separation we allow for the variances of components of ε^* in the hierarchical case, because such separation seems substantively sensible, that we do not allow in the random-coefficient case. In either case, the familiar class of robust estimators and/or reasonably familiar versions of feasible generalized least squares (FGLS) will redress OLS problems sufficiently in a relatively straightforward manner.

Recall that White's heteroskedastic-consistent estimator, for example, is

$$\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})} = n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{S}_0(\mathbf{X}'\mathbf{X})^{-1} \quad \text{where } \mathbf{S}_0 = \frac{1}{n}\sum_{i=1}^{n}\mathrm{e}_i^2\mathbf{x}_\mathrm{i}\mathbf{x}_\mathrm{i}'$$

As Greene (2003) writes, White's estimator "implies that, without actually specifying the type of heteroskedasticity, we can still make appropriate inferences based on the results of least squares" (199). More precisely, White's estimator produces *consistent* estimates of the coefficient estimates' variance-covariance matrix in the presence of pure heteroskedasticity (nonconstant variance) whose pattern is somehow related to a pattern in $\mathbf{xx}'$, that is, to some pattern in the regressors, the regressors squared, or the cross-products of the regressors. Thus, in our pure random-coefficient case, White's estimator provides consistency ("robustness") to precisely the heteroskedasticity issue raised because the pattern of nonconstant variance depends on the regressors x and z and heteroskedasticity is the only issue raised. In the hierarchical-model case, we might additionally have concerns about a correlation among residuals due to the common components, ε_{1j}, in the errors of all individuals in context j. The pattern of this induced correlation will likewise relate to the regressors x and z (and their products and cross-products). In this case, a Huber-White heteroskedasticity-*and-clustering*-consistent variance-covariance estimator will produce the appropriately "robust" standard errors.[19]

Such "robust" standard-error estimators leave the inefficient coefficient estimates unchanged and are not efficient in their estimates of coefficient-estimate variance-covariance either. To redress these issues, feasible weighted least squares (FWLS) may be appropriate for the pure heteroskedasticity induced by simple random effects, and FGLS may be appropriate for the heteroskedasticity and correlation induced by the clustering likely in the hierarchical context. Specifically, since the patterns of heteroskedasticity or correlated errors producing the concerns are a simple function of the regressors involved in the interactions, one can conduct FWLS if appropriate and desired following these steps: (1) estimate by OLS; (2) save the OLS residuals; (3) square the OLS residuals; (4) regress the squared residuals on the offending regressors (x and z here); (5) save the predicted values of this auxiliary regression. The researchers would then (6) use the inverse of the square roots of these predicted values as weights for the FWLS reestimation. One may wish instead to regress the *log* of the squared OLS residuals on the offending regressors and save the *exponential* of these fitted values in step (5) to avoid estimating negative variances and then attempting to invert their square roots in step (6). The procedure for implementing FGLS if appropriate and desired is similar, except that both variance and covariance parameters are to be esti-

[19] Again, Franzese (2005) discusses this matter further.

mated in steps (3) and (4) for insertion into the $\widehat{\mathbf{V}(\hat{\boldsymbol{\varepsilon}})}$ whose "square root inverse" is to provide the weighting matrix in step (6).[20]

As evidence in support of the claim that some form of a robust-cluster estimate will suffice in the hierarchical model with random coefficients case, we conducted several Monte Carlo experiments applying OLS, OLS with heteroskedasticity-consistent standard-error estimation, OLS with heteroskedasticity-and-cluster-consistent standard-error estimation, and random-effect-model estimation.[21] In all cases, the data were actually generated using hierarchical-model structures (with several alternative relative variances and covariances of the error components and the right-hand-side variables) and in samples with fifty j units and one hundred observations per unit (to correspond to a rather small survey conducted in each of the fifty U.S. states). All four estimation techniques yielded unbiased coefficient estimates, but the standard-error estimates, not surprisingly, were wrong with OLS and with robust standard-error estimates that ignore within-level autocorrelation (i.e., estimators consistent to heteroskedasticity only) but were nearly as good with the robust-cluster-estimation strategy as with the full random-effects model (the estimates were within 5 percent of each other). Appreciable efficiency gains in coefficient estimates from the hierarchical models relative to the OLS models were also notably absent. Accordingly, the main conclusion of our exercise was that one seemed generally to have little to gain—*in linear models in samples of these dimensions anyway*—from complicated random-coefficients and hierarchical-modeling strategies. OLS with robust variance-covariance estimator strategies (e.g., in STATA, one simply appends ", robust" or ", robust cluster" to the end of the estimation command) seemed generally to suffice. Of course, we would demand much further simulation, across wider and more systematically varying model types and ranges of parameters and sample dimensions, to support this conclusion more wholeheartedly as a general one. In this sample dimension and model context at least, however, simpler strategies work almost indistinguishably from the more

[20] The "square-root inverse" of a matrix with nonzero off-diagonal elements is not a simple inversion of the square root of each of the elements, as it is in the FWLS case where $V(\boldsymbol{\varepsilon})$ is diagonal. However, most statistical software packages will find the square-root inverse of a matrix, and so we need not detain the reader with these computations.

One could also iterate the FWLS or FGLS procedures, and common practice is to do so, even though, statistically, the iterated and one-shot strategies have identical properties.

[21] The variance-covariance matrix for coefficients estimated with the particular robust cluster we implemented (using STATA) is $\widehat{\mathbf{V}(\hat{\boldsymbol{\beta}})} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{S}_{\mathrm{J}}(\mathbf{X}'\mathbf{X})^{-1}$ where $\mathbf{S}_{\mathrm{J}} = \sum_{j=1}^{J}\mathbf{u}_{\mathrm{j}}'\mathbf{u}_{\mathrm{j}}$ and where $\mathbf{u}_{\mathrm{j}} = \sum_{i=1}^{n_j}\mathrm{e}_{ij}\mathbf{x}_{ij}$. We estimated the random effects model using hierarchical linear model (HLM) software.

complex ones, and so we are happy to argue for simplicity in cases like this at any rate. We also note, however, that the properties of these "robust" standard-error estimators deteriorate in smaller samples. For the simple heteroskedasticity-consistent estimator, this seems to occur only in very small samples beginning around $N = 35$. For robust-cluster estimators, two sample-size dimensions are key: total, N, and J, the number of "contexts." Again, very small J, say, below about thirty, and/or N become increasingly problematic.[22]

[22] These sample sizes and dimensions come from consideration of the small-sample adjustments some statisticians have recommended to these robust estimators, multiplying White's by a term involving $N/(N - 1)$ and robust cluster by a term involving $[N/(N - 1)][J/(J - 1)]$. Franzese (2005) discusses these considerations in far greater depth. See also Achen (2005), who correctly stresses the possible reliance upon linearity for many of these results and conclusions.