MODELING AND INTERPRETING INTERACTIVE HYPOTHESES IN REGRESSION ANALYSIS:

A REFRESHER AND SOME PRACTICAL ADVICE

Cindy D. Kam and Robert J. Franzese, Jr.[1]

Submitted to the University of Michigan Press

March 31, 2005

[1] This project evolved from a previous paper: Kam, Cindy D., Robert J. Franzese, Jr., and Amaney Jamal. 1999. "Modeling Interactive Hypotheses and Interpreting Statistical Evidence Regarding Them," presented at the Annual Meetings of the American Political Science Association. We thank Amaney Jamal for her key role in those origins, and we also gratefully acknowledge Joel Simmons for research assistance in updating data from that previous paper for this project.

PREFACE

This pedagogical manuscript addresses the modeling, interpreting, testing, and presentation of interactive propositions in regression analysis. As an instructional text, we intend it to provide guidance on these issues to advanced undergraduates, graduate students, and social science researchers in political science and other disciplines. The manuscript begins by explaining how verbal statements of interactive arguments and hypotheses translate into mathematical empirical models including, and statistical inferences regarding, interactive terms. It then provides advice on estimating, interpreting, and presenting the results from such models. The manuscript provides next an explanation of some existing general practice rules and, lastly, a discussion of more advanced topics including nonlinear models and stochastically interactive models. The manuscript's summary conclusion outlines our general advice for researchers as they formulate, estimate, test, interpret, and present interactive hypotheses in their empirical work.

Modeling and Interpreting Interactive Hypotheses in Regression Analysis:

A Refresher and Some Practical Advice

Cindy D. Kam and Robert J. Franzese, Jr.

I. INTRODUCTION

The complexity of the phenomena political scientists study has led them to explore a wide variety of quantitative methods and tools for empirical analyses. Typically, political scientists are interested in identifying the impact of some variable(s), ***X*,** on some dependent variable, ***Y***. One of the simplest model specifications for such empirical explorations posits some set of independent variables to have linear, additive relationship with a dependent variable; indeed, much of the quantitative analysis in print exemplifies this approach.

One common theoretical claim of additional complexity is that the effect of some variable, ***X*,** on the dependent variable, ***Y***, depends upon a third independent variable(s), ***Z***.[2] For example, the effect of partisanship on legislative voting behavior may depend upon whether the legislation has bipartisan sponsorship or partisan sponsorship. As another example, the effect of some core value, such as egalitarianism, on a citizen's support for a policy outcome may depend on whether elites have framed the issue to evoke that core value. As yet another example, the effect of the structure and relative power of interest groups in society on government policy may depend upon the nature of the electoral system that produces from that structure of interests the representatives who make policy in that polity.

At their core, interactive hypotheses such as these propose that the effect of some

[2] More generally: the effect or effects of some variable or variables, ***X***, depend on some other variable or set of variables, ***Z***. For ease of exposition, the discussion primarily focuses on a single variable, *x*, and a single variable, *z*, as they relate to a single dependent variable, *y*. The general claims can be extended to vectors *X* and *Z*.

variable, *x*, on the dependent variable, *y*, is conditioned by a third variable, *z*. Political scientists often evaluate such hypotheses using the linear interactive, or multiplicative, term.[3]

Interaction terms are hardly new to political science; indeed, their use is now almost common. Given the growing attention to the roles of institutions in politics, and the growing attention to how context (e.g., neighborhood composition, social networks) conditions the influence of individual-level characteristics on political behavior, interactive hypotheses should perhaps become more common still. However, despite occasional constructive pedagogical treatises on interaction usage in the past, a commonly known, accepted, and followed methodology for using and interpreting interaction terms continues to elude much of the field. Partly as a consequence, misinterpretation and substantive and statistical confusion remains rife. Sadly, Friedrich's (1982) summary of the state of affairs could still serve today:

> …while multiplicative terms are widely identified as a way to assess interaction in data, the extant literature is short on advice about how to interpret their results and long on caveats and disclaimers regarding their use (798).

This manuscript seeks to redress this and related persistent needs. Our discussion assumes working knowledge of the linear additive regression model.[4] Section II begins our discussion of modeling and interpreting interactive hypotheses by emphasizing the ways in which interactive terms are essential for testing common and important classes of theories in political science.

[3] Scholars also refer to the interactive term as the multiplicative or product term, or the moderator variable, depending on the discipline. We use interactive term and multiplicative term interchangeably. In the field of psychology, distinctions have been made between mediator and moderator variables (Baron and Kenny 1986). The distinction is similar to that made in other disciplines, including sometimes in political science, between intervening and interactive variables, but this terminology is not consistently applied across, or sometimes even within, discipline. Our discussion applies to *moderator* and *interactive* variables, which Baron and Kenny (1986) define as "a qualitative… or quantitative… variable that affects the direction and/or strength of the relation between an independent or predictor variable and a dependent or criterion variable" (1174).

[4] For a refresher on the linear additive regression model, the interested reader might consult Achen (1982).

In Section III, we offer a generic consideration of the process of writing empirical models to embody interactive hypotheses. We then show which standard statistical tests speak to each hypothesis typically nested within interactive propositions and how they do so. That is, we note the correspondence between certain common *t*- and *F*-tests and specific aspects of the typical complex of hypotheses nested in interactive arguments. We include a suggested generic approach to interpreting the estimation results of interactive models. We also address the presentation of interaction effects. We urge researchers (and editors and publishers) to go beyond merely reporting standard errors for individual coefficients. Instead, we strongly suggest graphical or tabular presentation of results, including effect-line graphs or conditional-coefficient tables, complete with standard errors, confidence intervals, or significance levels of those effects or conditional coefficients. We suspect that the failure of authors to generate (or editors to retain or publishers to print) such tables and graphs underlies much of the lingering misunderstanding regarding interaction terms in our profession. We show how to construct such graphs and tables using statistical software commonly used in political science, in addition to specific mathematical formulae for their elements. Our approach underscores the inescapable importance of understanding the elementary mathematics underlying models that use interactive terms, rather than simply providing a set of commands for the user to enter by rote.

In Section IV, we consider certain general-practice rules for modeling interactions that some previous methodological treatments advise and political scientists often follow. We suggest that some scholars may be misinterpreting these and argue that such general rules should never substitute for a solid understanding of the simple mathematical structure of interaction terms. For example, "centering" the variables to be interacted, as several methods texts advise, alters nothing important statistically and nothing at all substantively. Furthermore, the common admonition that one must include both *x* and *z* if the model contains an *xz* term is an often-advisable philosophy-of-science guideline—as an application of Occam's Razor and as a practical matter, it is usually a much safer adage than its absence—but it is neither logically nor statistically necessary and not always advisable, much less required, in all cases for any question.

Section V discusses some more-technical concerns often expressed regarding interactive models. We briefly review three-way (and multiple) interaction models. We discuss the question of pooled versus separate-sample estimation. We assess whether to estimate interactive effects in separate or pooled samples, showing that the approaches are essentially equivalent but that pooled-sample estimation usually facilitates statistical comparisons even if one might prefer separate-sample estimation in preliminary analyses. Although much of our discussion addresses multiplicative terms exclusively in the context of linear regression models, we realize that a growing proportion of statistical research in political science employs qualitative or limited dependent-variable models or other models beyond linear ones. While most of the discussion regarding linear regression models also holds for nonlinear models, we will discuss in more detail the special case of interactive terms in such models in this section. Finally, we address random-coefficient and hierarchical models. As, e.g., Western (1998) notes, using multiplicative terms alone to capture the dependence on *z* of *x*'s effect on *y* (and *vice versa*) implicitly assumes that the dependence is deterministic. Yet this dependence is surely as stochastic as any other empirical relationship we might posit in social science, so we should perhaps model it as such, which concern many take as dictating the use of random-coefficient models. Others go further to claim that cross-level interaction terms—i.e. those involving variables at a more micro-level (e.g., individual characteristics in a survey) and at a more macro-level (e.g., characteristics of that individual's state of residence)—may induce a bias, which they take as demanding the use of hierarchical linear models or separate-sample estimation. These issues are actually related, and, as we show (perhaps surprisingly), the simple multiplicative term usually sacrifices little relative to these more complicated or tedious approaches. Moreover, steps of intermediate complexity can allay those concerns (not quite fully, but likely sufficiently), and, in fact, most political-science practice has actually been taking one such step all along, if for other reasons. Thus, some of these concerns are, strictly speaking, well founded, but they rarely amount to serious practical problems for social scientists.

Section VI provides a summary of our advice for researchers seeking to formulate, estimate, and test interactive hypotheses in empirical research.

II. INTERACTIONS IN POLITICAL SCIENCE

The interaction term received intense scrutiny, much of it critical, upon its introduction to social science. Althauser (1971) wrote, “It would appear, in short, that including multiplicative terms in regression models is not an appropriate way of assessing the presence of interaction among our independent variables” (466). Zedeck (1971) concurred, “The utility of the moderator variable research is limited by statistical problems, by the limited understanding of the statistical operation of moderators, and by lack of a rapid systematic approach to the identification of moderators” (307).

As Friedrich noted, this early criticism of interactions raised three concerns: difficulty interpreting coefficients, colinearity among independent variables induced by the multiplication of terms, and the nature of measurement of independent variables (whether they be interval, ratio, or nominal scales). These concerns inspired some scholars (Althauser 1971; Zedeck 1971; among others) to object to any interactive-term usage. Others suggested alternative methods to incorporate interactions in models, by rescaling variables to reduce colinearity (Allison 1979;

Cronbach 1987; Dunlap and Kemery 1987; Smith and Sazaki 1979; Tate 1984).

Over twenty years after Friedrich's (1982) seminal article defending interactions, full and accurate understanding of the modeling, interpretation, and presentation of interactive hypotheses still eludes the field, even though including multiplicative terms in linear regressions is now a common method of incorporating conditional relationships into empirical analysis.

For example, in a count of journal articles that appeared from 1996-2001 in three top political science journals, we have found that 54% of articles use some statistical methods (defined as articles reporting any standard errors or hypothesis tests, using linear regression models or nonlinear models). Of these statistical articles, 24% employ interactive terms. This amounts to about 1/8 of all articles published during this time.[5] Despite the appreciable and increasing use of interaction terms in empirical analysis, careful consideration of important classes of theoretical arguments in political science strongly suggest that they nonetheless remain considerably under-utilized. Further, when interactions are deployed in empirical work, several misunderstandings regarding their interpretation still permeate the field.

The widespread and perhaps expanding usage of interactions notwithstanding, we contend that still more empirical work should contain interactions than currently does, given the substance of many political-science arguments. Such interactive arguments arise commonly in every empirical subfield in political science, including the study of political institutions, political behavior, and perhaps especially researchers who study the impact of institutions on political behavior (not to mention political economy, political culture, and all the other substantive areas

[5] Incidentally, these shares likely dramatically understate the mathematical technical nature of the field since our denominator includes pure-theory articles, formal and philosophical, and our nominator excludes formal theory. The share of statistical and formal-theoretical articles in these journals likely approaches 75% of all non-political-philosophy articles.

of study within political science).

Consider, for example, the gist of most institutional arguments, reflecting perhaps the dominant approach to modern, positive[6] political science. In one influential statement of the approach, Hall (1986) states:

> …the institutional analysis of politics… emphasizes institutional relationships, both formal and conventional, that bind the components of the state together and structure its relations with society… [I]nstitutions…refers to the formal rules, compliance procedures, and standard operating practices that structure the relationship between individuals in various units of the polity and economy… Institutional factors play two fundamental roles… [They] affect the degree of power that any one set of actors has over policy outcomes […and they…] influence an actor's definition of his own interests, by establishing his… responsibilities and relationship to other actors… With an institutionalist model we can see policy as more than the sum of countervailing pressure from social groups. That pressure is mediated by an organizational [i.e., institutional] dynamic… (19, emphases added).

Thus, in this approach, and we believe inherently in all institutional approaches, institutions are *interactive* variables that funnel, moderate, or otherwise shape the political processes that translate the societal structure of interests into effective political pressures, and/or those pressures into public-policymaking responses, and/or those policies into outcomes. Across all the methodological and substantive domains of institutional analysis, further examples abound:

> …[political struggles] are mediated by the institutional setting in which they take place (Ikenberry 1988: 222-3, emphases added).
>
> …[1] institutions constrain and refract politics but…are never the sole "cause" of outcomes. Institutional analyses do not deny the broad political forces that animate […class or pluralist conflict, but stress how…] institutions structure these battles and, in so doing, influence their outcomes. [2. They] focus on how [the effects of]

[6] We intend the term *positive* as opposed to *normative* here and do not intend it to connote *formal* necessarily.

> macrostructures such as class are magnified or mitigated by intermediate-level institutions…[they] help us integrate an understanding of general patterns of political history with an explanation of the contingent nature of political and economic development… [3] Institutions may be resistant to change, but their impact on political outcomes can change over time in subtle ways in response to shifts in the broader socioeconomic or political context (Steinmo and Thelen 1992: 3, 11-2, 18, emphases added).
>
> …the idea of structure-induced equilibrium is clearly a move [toward] incorporating institutional features into rational-choice approaches. Structure and procedure combine with preferences to produce outcomes (Shepsle 1989: 137, emphases added).

Other recent examples include research that connects the societal structure of interests to effective political pressure through electoral institutions: most broadly, plurality-majority *versus* proportional representation (e.g., Cox 1997; Lijphart 1994); that studies how governmental institutions, especially those that affect the number and polarization of key policymakers (veto actors), shape policymaking responses to such pressures (e.g., Tsebelis 2002); that stresses how the institutional configuration of the economy, such as the coordination of wage-price bargaining, shapes the effect of certain policies, such as monetary policy (Franzese 2003b reviews). Examples could easily proliferate yet further.

In every case, and at each step of the analysis from interest structure to outcomes (and back), the role of institutions is to mediate, shape, structure, or condition[7] the impact of some other variable(s)[8] on the dependent variable of interest. That is, most (we believe: all) institutional arguments are inherently interactive. Yet, with relatively rare exceptions—regarding the above examples: see, e.g., Ordeshook and Shvetsova 1994; Franzese 2002b ch. 4; and

[7] Extending the list of synonyms might prove a useful means of identifying interactive arguments. When one says *x alters, changes, modifies, magnifies, augments, increases, intensifies, inflates, moderates, dampens, diminishes, reduces, deflates, etc.* some effect (of *z*) on *y*, one has offered an interactive argument.

[8] Institutions seem most often to condition the impact of structural variables: e.g., interest, demographic, economic, party-system structure, etc. We suspect that reflects some as-yet unstated general principle of institutional analysis.

Franzese 2001, respectively—empirical evaluations of institutional arguments have ignored this interactivity.

A more generic example further illustrates the common failure of empirical models to reflect the interactions that the theoretical models generating them imply. Scholars consider principal-agent (*i.e.*, delegation) situations interesting, problematic, and worthy of study because, if each had full control, agents would determine policy, y_1, by responding to some (set of) factor(s), $\boldsymbol{X}$, according to some function, $y_1=f(\boldsymbol{X})$, whereas principals would respond to some perhaps-different (set of) factor(s), $\boldsymbol{Z}$, perhaps-differently, following $y_2=g(\boldsymbol{Z})$. (For example, the principal might be a current government, which responds to various political-economic conditions in setting inflation policy, and the agent an unresponsive central bank, as in Franzese 1999.) Scholars then offer some arguments about how institutional and other environmental conditions determine the monitoring, enforcement, and other costs, $\boldsymbol{C}$, principals must incur to force agents to enact $g(\boldsymbol{Z})$ instead of $f(\boldsymbol{X})$. In such situations, realized policy, y, will follow some $y=k(\boldsymbol{C})\cdot f(\boldsymbol{X})+[1-k(\boldsymbol{C})]\cdot g(\boldsymbol{Z})$ with $0\leq k(\boldsymbol{C})\leq 1$ and $k(\boldsymbol{C})$ weakly increasing (see, *e.g.*, Lohmann 1992 on the banks, governments, and inflation example). Thus, the effect on y of each c‽$\boldsymbol{C}$ generally depends on $\boldsymbol{X}$ and $\boldsymbol{Z}$, and those of each x‽$\boldsymbol{X}$ and z‽$\boldsymbol{Z}$ generally depend on $\boldsymbol{C}$. That is, everything that contributes to monitoring and enforcement costs modifies the effect on y of all factors to which the principals and agents would respond differently, and, *vice versa*, the effect of everything that determines monitoring and enforcement costs depends on all such factors.[9] Most empirical models of principal-agent situations do not reflect this their inherent interactivity.

For those who study political behavior, opportunities to specify interactive hypotheses

[9] Franzese (1999, 2002) shows how to use nonlinear regression to mitigate the estimation demands of such highly interactive propositions.

also abound. Scholars who argue that the effects of some set of individual characteristics (e.g., partisanship, core values, or ideology) depend on another variable (e.g., race, ethnicity, or gender) are proposing hypotheses that can and should be explored with interactive terms. Research questions that ask how the impact of some experimental treatment or campaign communication depends on the level of some individual characteristic (e.g., political awareness) likewise imply interactive hypotheses. Questions that explore how context (e.g., minority neighborhood composition or news media coverage of an issue) conditions the effect of some other predictor (e.g., racism) also reflect interactive hypotheses.[10]

Interaction terms are widely used in statistical research in political science, and, in many more cases, theories suggest that interactions should be used although they are not. Despite their proliferation, some confusion persists regarding how to interpret these terms, which is all the more worrisome since that proliferation should continue and expand; accordingly, we turn now to addressing (and hopefully helping to redress) some of this confusion.

III. THEORY TO PRACTICE.

In this section, we provide some guidance for constructing statistical models to map onto substantive theory, implementing statistical analyses to test the theory, and presenting empirical results resulting from the analyses.

A. Specifying Empirical Models to Reflect Interactive Hypotheses

Optimally, theory should guide empirical specification and analysis. Thus, *e.g.*, empirical models of principal-agent and other shared-policy-control situations should reflect the convex-combinatorial form, with its multiple implied interactions, described above (Franzese 1999;

[10] These last also imply *spatial interdependence* (correlation), on the methodological issues involved in estimating which see, e.g., Franzese and Hays (2005), Beck & Gleditsch (2005), and contributions to *Political Analysis* 10(3).

2002b give examples and discussion).

To focus our discussion here and throughout this manuscript, we will use an empirical example from Cox's (1997) *Making Votes Count.* Cox's justifiably acclaimed book makes several institutional arguments in which some political outcome, *y*, say the effective number of parties elected to a legislature or the effective number of presidential candidates, is a function of some structural condition, *x*, say the number of societal groups created by the pattern of social cleavages (e.g., effective number of ethnic groups), and some institutional condition, *z*, say the proportionality or district magnitude of the electoral system or the presence or absence of a presidential runoff system. Theory in this case very clearly implies that the relationship between *y* and *x* should be conditional upon *z* and, conversely, that the relationship between *y* and *z* should be conditional upon *x*. As Cox (1997) theorizes, for example, "A polity will have many parties only if it <u>both</u> has many cleavages <u>and</u> has a permissive enough electoral system to allow political entrepreneurs to base separate parties on these cleavages. Or, to turn the formulation around, a policy can have few parties either because it has no need for many (few cleavages) or poor opportunities to create many (a constraining electoral system)" (206). (See, *e.g.*, Ordeshook and Shvetsova 1994; Neto and Cox 1997; and Cox 1997 for empirical implementation.)

One could perhaps conceive several ways in which some variable *z* (or *x*) may condition the relationship between *y* and *x* (or *z*) (see Wright 1976), but, most generally, the standard linear-interactive model would reflect a proposition that *x* and *z* affect *y* and that, in particular, the effects of *x* and of *z* on *y* each depend on the other variable. One simple way to write this (compound) proposition into a linear regression model would be to begin with a standard linear-additive model expressing a relation from *x* and *z* to *y*, along with an intercept, and then to allow the intercept and the coefficients on *x* and *z* each to depend on the level of *x* and *z*:[11]

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon \qquad [1]$$

$$\beta_0 = \gamma_0 + \gamma_1 x + \gamma_2 z$$

$$\beta_1 = \delta_1 + \delta_2 z$$

$$\beta_2 = \delta_3 + \delta_4 x$$

which implies that one may express the model of *y* for estimation by linear regression in the standard (linear) multiplicative-interactive manner:

$$y = \gamma_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon \qquad [2]$$

As originally expressed in [1], the coefficients in this linear-interactive model happen to be as follows: $\beta_x = \gamma_1 + \delta_1$, $\beta_z = \gamma_2 + \delta_3$, $\beta_{xz} = \delta_2 + \delta_4$. More importantly, in this model, the effects of *x* and *z* on *y* depend on *z* and *x*, respectively, as an interactive theory would suggest.

Theory, however, might dictate a different route to this same general model. For example, suppose one were to specify a system of relationships in which the effect of *x* and the intercept depend on *z*:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad [3]$$

$$\beta_0 = \gamma_0 + \gamma_1 z$$

$$\beta_1 = \delta_1 + \delta_2 z$$

which implies the following model for *y*:

$$y = \gamma_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon \qquad [4]$$

Where $\beta_x = \delta_1$, $\beta_z = \gamma_1$, $\beta_{xz} = \delta_2$.

[11] We begin with the simplest case, where the effects of *x* and of *z* are deterministically dependent on, respectively, *z* and *x*. Later, we relax this assumption to discuss probabilistic dependence (i.e., with error).

Note that the models actually estimated in [2] and [4] are identical, even though the theoretical stories told to derive the models from [1] and [3] seem to differ. Furthermore, if one were to propose a model in which y is a linear-additive function of z and the effect of z and the intercept depends on x, or, for that matter, simply to propose that the effect of x depends on z or the effect of z depends on x (and each effect may be nonzero when the other variable equals zero), this same empirical model would emerge again. Each of these seemingly different theoretical stories yields the same mathematical model: the linear-interactive model [2].[12]

This result demonstrates that, although the substance may determine which of these arguments is most intuitive to express, the distinction cannot be drawn mathematically. One must understand, however, that this mathematical ambiguity arises because the propositions being modeled are logically symmetric; *i.e.*, these statements all logically imply each other and, in that sense, they are identical; they mean precisely the same thing; they cannot be distinguished because they are not distinct. As Fisher (1988) writes, "prior theoretical specification is needed to interpret [in this sense] regression equations with product terms" (106). We concur but stress that the interpretive issues here are presentational and semantic because the alternatives are logical equivalents. These alternative theoretical stories may sound different in some substantive contexts, and some versions may seem more intuitive to grasp in certain contexts and others in other contexts. However, they are not actually alternatives; they are all the same tale.

Alternatively, one could propose the substantive argument that the effect of x on y depends on z, but that z matters for y only insofar as it alters the impact of x and, in particular, z has no effect when x is equal to zero (not present). This *is* a change in the theoretical account of

[12] Note: the linear-interactive model is not the only model form that would imply that the effects of x depend on z and *vice versa*, but, absent further theoretical elaboration that might suggest a more specific form of interaction, additive linear-interactive models like [2] are the logical, simple default in the literature.

the relationship between the variables, a logically distinct argument, and it produces a truly different equation to be estimated:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad [5]$$

$$\beta_1 = \delta_1 + \delta_2 z$$

$$y = \beta_0 + \beta_x x + \beta_{xz} xz + \varepsilon \qquad [6]$$

Where $\beta_x = \delta_1$, $\beta_{xz} = \delta_2$.

In this system of equations, we again see that *z* conditions the effect that *x* has on *y*. In fact, *z*'s sole effect is to determine the effect of *x* on *y*, and, in particular, movements in *z* have no effect on *y* when *x* is zero. Scholars will typically think of *z* in this scenario as the <u>intervening variable</u>: intervening in *x*'s relationship to *y*. However, notice that just as a value of *x*, namely *x*=0, exists where the effect of *z* is zero, a value of *z* exists, namely $z = -\beta_x / \beta_{xz}$, where the effect of *x* is zero. The substance of the context at hand may suggest whether to conceive *x=0* or $z = -\beta_x / \beta_{xz}$, or, for that matter, some other value of *x* or *z*, as the base from which to decide whether *x* or *z* is the one <u>intervening</u> in the other's relationship with *y*. Mathematically that determination is once again arbitrary because, <u>logically, all interactions are symmetric</u>.[13] Given this logical symmetry, *x* and *z* must necessarily <u>both</u> intervene in the other's relationship to *y*. In this sense, the language of one variable being the intervening, moderating, or mediating variable and the other being the one mediated may be best avoided; if an interaction exists, then all variables involved intervene, moderate, and mediate (etc.) in the others' relations to *y*.

[13] Mathematically, the proof of this logically necessary symmetry in all interactions is simply:

$$\frac{\partial\left(\frac{\partial f(x,z)}{\partial x}\right)}{\partial z} \equiv \frac{\partial^2 f(x,z)}{\partial x \partial z} \equiv \frac{\partial^2 f(x,z)}{\partial z \partial x} \equiv \frac{\partial\left(\frac{\partial f(x,z)}{\partial z}\right)}{\partial x} \; \forall \; f(x,z).$$

The equations above assume deterministic relations of z to the effect of x on y (and of x to the effect of z on y). That is, all of the models assume that the effect of x on y depends on z and the effect of z on y depends on x deterministically, i.e. without error. This should seem odd: in our theoretical model we expect that x and z predict y only with error (hence the inclusion of the term ε); but then we expect that the effect of x on y and of z on y each depend on the other variable without error. We can easily amend the linear-interactive model to allow a more sensible and logically consistent probabilistic or stochastic conditioning of the effects of variables by the others' levels as follows:

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon \qquad [7]$$

$$\beta_0 = \gamma_0 + \gamma_1 x + \gamma_2 z + \varepsilon_0$$

$$\beta_1 = \delta_1 + \delta_2 z + \varepsilon_1$$

$$\beta_2 = \delta_3 + \delta_4 x + \varepsilon_2$$

which implies that one may model Y for regression analysis in the now-familiar manner:

$$y = \gamma_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon^* \qquad [8]$$

where: $\varepsilon^* = \varepsilon + \varepsilon_0 + \varepsilon_1 x + \varepsilon_2 z$, $\beta_x = \gamma_1 + \delta_1$, $\beta_z = \gamma_2 + \delta_3$, $\beta_{xz} = \delta_2 + \delta_4$.

Notice, however, that, although the composite residual in [8] retains zero expected-value and non-covariance with the regressors x, z, and xz if its components, ε , ε_0, ε_1, and ε_2, do so. These key assumptions of the classical linear regression model (CLRM) ensure unbiasedness and consistency of OLS (Ordinary Least Squares); however, this compound residual does not retain constant variance, the CLRM assumption that assures OLS efficiency and accurate standard errors. In other words, if the conditionality of the x and z relationships with y themselves contain error, then the standard linear-interactive model has heteroskedastic error even if the individual stochastic terms comprising its compound residual are homoskedastic. Thus, OLS coefficient estimates are unbiased and consistent but not efficient. Furthermore, the non-constant variances are heteroskedastic as a function of the regressors *x* and *z*, which is more pernicious for OLS standard-error estimation than heteroskedasticity unrelated to the regressors would be, causing bias and inconsistency as well as inefficiency in these standard-error estimates. We return to this technical concern, which underlies calls both for random-coefficient and for linear-hierarchical models, below. To proceed for now, however, recall that heteroskedasticity alone, even if it is a function of the regressors, leaves OLS estimates of the coefficients unbiased and consistent, although inefficient. Heteroskedasticity as a function of the regressors renders OLS estimates of the variance-covariance of the estimated coefficients (i.e., standard-errors and estimated covariances) biased, inconsistent, and inefficient, but, as we show below, these problems are not necessarily as grave as they sound, are easy to redress by simple and familiar techniques, and, indeed, that the "robust" variance-covariance estimators political scientists typically employ already redresses these problems in many cases. Thus, we can proceed for now assuming the researcher estimates a model like [8] by OLS.

Let us return to our example of types of electoral systems, social cleavages, and number of parties or candidates to illustrate the above discussion. The analysis we will follow examines the effects of a presidential runoff system (*Runoff)* and the effective number of ethnic groups in a society (*Groups*) in determining the effective number of presidential candidates (*Candidates)* to emerge in the presidential democracy. The theory suggests that (1) the impact of social cleavages on the effective number of candidates depends on whether a runoff system is used, and (2) the impact of the runoff system on the effective number of candidates depends on the number of social cleavages in the society. (Recall that these are logically two sides of the same proposition.)

The confluence of a high number of social cleavages and the presence of a runoff system are hypothesized to produce a high effective number of presidential candidates, since the runoff system attenuates the incentives for pre-election coalition-building between groups. Given this theoretical structure, we can estimate the following:

$$Candidates = \beta_0 + \beta_G Groups + \beta_R Runoff + \beta_{GR} Groups * Runoff + \varepsilon \qquad [9]$$

The dataset we will use includes information from 16 presidential democracies in 1985.[14] The dependent variable, *Candidates,* refers to the number (specifically, the vote-share weighted, or so-called effective number) of presidential candidates in democracies; this variable ranges between 1.958 and 5.689, with a mean of 3.156 and a standard deviation of 1.202. The variable, *Groups,* refers to the effective (size-weighted) number of ethnic groups within a society.[15] This variable ranges between 1 and 2.756, with a mean of 1.578 and a standard deviation of 0.630. The variable, *Runoff,* refers to the presence or absence of a runoff system for the presidential election; this dummy variable takes the value of 0 if the system does not employ runoffs and 1 if the system does use runoffs. Within the sample of 16 presidential democracies, exactly half have a runoff system. The OLS regression results appear in Table 1.

[TABLE 1 ABOUT HERE]

Now that we have the results, how do we interpret them? What do these estimated coefficients mean? The next section provides guidance on these questions.

B. Interpreting Coefficients from Interactive Models.

[14] The data are freely available at http://dodgson.ucsd.edu/lij/pubs/. We have selected this dataset because it is freely available and researchers can easily replicate the results herein. The small N makes firm tests of statistical significance more than a little precarious—we discuss this issue as it arises below—but need not affect our pedagogical purposes.

[15] To avoid some tiresome repetition, we henceforth drop the adjectives effective, although they remain applicable.

In the simple linear-additive regression (indeed, only in that case), the coefficient on a variable and the effect on the dependent variable of a unit increase in that independent variable (*ceteris paribus* and controlling for other regressors) are identical. In interactive models, this equivalence between coefficient and effect no longer holds. To cope with this change, current practice in interpreting interactive effects often attempts to substitute some vague and potentially misleading terms, such as “main effect” and “interactive effect,” “direct effect” and “indirect effect”, and “independent effect” and “total effect” for the coefficients on *x* and *z* in the first case and on *xz* in the second.[16] Such terminology is usually unhelpful at best, misleading or wrong at worst, and, in any event can never sufficiently substitute for comprehension of the simple math of interactions. We encourage researchers instead to recall that each variable involved in the interaction terms of interactive models has multiple effects, neither any single, constant effect nor a “main” effect and an “interactive” effect, but multiple, different effects depending on the levels of the other variable(s) with which it interacts. This is, of course, exactly as implied substantively by interactive hypotheses in the first place.

An interactive hypothesis is implied when a researcher suggests that the effect of some variable *x* on *y* depends on *z*. This logically implies that the researcher believes that *x* has different effects on *y*, depending on the specific values of *z*. In the interactive case, the effects of *x* on *y* are therefore not a single constant, like the coefficient β_x on *x* is in the simple linear additive model. Instead, the effects of *x* on *y* vary depending on the coefficients on *x* and *xz*, as well as the value of *z*.

Neither are the terms “main” and “interactive” or “direct” and “indirect” effects for these

[16] Note that some of this terminology also refers to path-analytic models, which specify that some variable *x* affects the level (rather than, or in addition to, the effect) of some variable *z* which then determines *y*. This overlap in terminology provides even more confusion for the researcher.

coefficients on *x* and on *xz* , respectively, helpful; indeed, we warn researchers against use of these terms. As we elaborate below, first, simply terming one coefficient the "main*"* effect and another the "interactive" effect perilously confuses coefficients for effects, and, second, there may in fact be nothing whatsoever "main" or "direct" about the particular effect to which the coefficient on *x* actually does refer. In fact, those varying effects of *x* at particular values of *z* involve the coefficients on both *x* and *xz* and those specific values of *z*. This demonstrates that researchers cannot appropriately refer to the coefficient on *x* as "the main effect of *x*" or "the effect of *x*…independent of z" or "…considered independently of *z*" or, certainly not, "…controlling for *z*".

The coefficient on *x* is just one effect *x* may have, namely the effect of *x* at *z*=0. That is, the coefficient on *x* gives the estimated effect of a unit change in *x,* holding *z* fixed at zero. Far from necessarily being a "main effect": (1) This zero value of *z* may have nothing at all "main" about it and may even be out-of-sample or even logically impossible! (2) This effect at *z*=0 is obviously not "independent of *z"*! (3) This effect of *x* on *y* given that *z*=0 is a very different thing from the effect of *x* on *y* "controlling for *z,*" which latter is what the simple linear-additive multiple-regression model would estimate, not what the linear-interactive model estimates! This confusion of coefficients, on *x* and/or *z* and/or *xz* for the effects of *x* and *z* probably also explains most failures to report standard errors (or other uncertainty estimates) for those effects rather than only the standard errors on the individual coefficients involved in those effects, which are far less useful by themselves. In an interactive model, the effects of *x* depend on, *i.e.* they vary with, the value of *z*, and so too do the standard errors of those effects!

Our empirical example clarifies these points. The estimated coefficient on *Runoff* ($\hat{\beta}_R = -2.491$) refers to the estimate of the effect of runoff elections on the effective number of presidential candidates for the specific case where *Groups* takes a value of 0. But in our description of the dataset, we noted that the number of ethnic groups never takes the value of 0 in the sample; in fact, the number of ethnic groups in a society cannot logically ever equal 0. Thus, an interpretation of $\hat{\beta}_R$, the estimated coefficient on *Runoff*, as the "main" effect of a runoff system is nonsensical; far from a "main" effect, this is actually the effect at a value of ethnic heterogeneity that does not, and, indeed, could not, exist.

If, however, *Groups* were rescaled to include a value of zero, e.g., by subtracting some constant value, such as the mean, and calling the resulting variable *Groups**, then the estimated coefficient $\hat{\beta}_{R*}$ would be the estimated effect of *Runoff* when the rescaled variable *Groups** takes the value of zero. This would be logically possible and in-sample now, but the notion that the effect at this particular substantive value of ethnic heterogeneity is somehow "main" would remain strained and potentially misleading. That the effect of some variable when its moderating variable(s) happen to be at its (their) mean (centroid) should be called a "main effect" while all the other effects at all the other logically permissible or empirically existent values are something other than, or somehow less than, "main" seems an unnecessary strain on meaning, especially since the theoretical and substantive point of the interaction model in the first place is that the effect<u>s</u> of the interacting variables <u>vary</u> depending on each other's values. We return to this topic of mean-rescaling interactive variables below, in Section V.

Symmetrically, the estimated coefficient $\hat{\beta}_G$, the coefficient on *Groups*, refers to our estimate of the effect of the number of ethnic groups when *Runoff* equals zero. This value does logically and empirically exist, so the estimated value of $\hat{\beta}_G$ = -0.979 does tells us something substantively relevant. It reports an estimate that, in a system without runoffs, the number of ethnic groups has a negative impact on number of presidential candidates; specifically, an increase of 1 in the number of ethnic groups is empirically associated with a 0.979 reduction in the number of presidential candidates, in systems without runoff elections. (We find this result substantively puzzling, but that is the estimate.) Note, however, that the coefficient $\hat{\beta}_G$ only tells part of the story – it only reveals the estimated effect of Ethnic Groups in one condition: when Runoff equals zero. Our main point here is that the researcher who equates a coefficient in an interactive model to an effect is treading on hazardous ground. At best, the researcher will be telling a story about an effect that applies to only one of several possible conditions (e.g., when z = 0). At worst, the researcher will be telling a story about an effect that applies in no logically possible condition—an effect that is logically meaningless. In short and put simply: outside the simplest linear-additive case, coefficients and effects are different things.

Given these considerations, we suggest two possible methods of interpreting results from interactive models, such that the researcher can make meaningful statements about the effect of some variable x on y, and the effect of some variable z on y. The two techniques we suggest are differentiation (which requires working knowledge of entry-level calculus) and differences in predicted values (which does not require even basic calculus). We discuss each of these techniques, below.

1. Interpreting Effects through Differentiation.

Consider the following standard linear-interactive regression-model:

$$y = \gamma_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon \qquad [10]$$

Deriving conditional effects through first derivatives of the dependent variable with respect to the independent variables of interest (as, e.g., Friedrich 1982 or Stolzenberg 1979 suggest) is a direct and simple means of identifying the effects of x on y or the effects of z on y because first derivatives or differences, *e.g.*, *dy/dx* and *dy/dz*, are effects. One may, in fact, read *dy/dx* as "the change in *y*, *dy*, induced by a marginal (derivative) or unit (difference) change in *x*, *dx*, all else held constant." Differentiation is a reliable, methodical way of discussing interactive effects. For it to fulfill its promise of simplicity and to reduce the tendency to induce mistakes, political scientists who feel the need should refer to a table of differentiation rules, available in any introductory calculus text and in the appendix. Taking the standard linear regression model in [10], the first derivatives of *y* with respect to *x* and *z* are as follows:

$$dy / dx = \beta_x + \beta_{xz} z \qquad [11]$$

$$dy / dz = \beta_z + \beta_{xz} x \qquad [12]$$

As [11] and [12] exemplify, the first derivative of [10] with respect to *x* and *z* yields the conditional effect of those variables directly. Derivatives *are* effects, whether in basic linear-additive regression models, when they give just the coefficient on the variable of interest, in linear-interactive models like [10], when they give expressions like [11] and [12] involving two coefficients and the other interacting variable, or in any other model regardless of its functional form, when they can give expressions involving all manner of combinations of parameters (usually coefficients) and variables.

The effect of *x* on *y* in an interactive model like [10] is $\beta_x + \beta_{xz} z$, which reflects the conditional argument underlying that model. β_x is merely the effect of *x* on *y* when *z* happens to equal zero, which, as noted above, is certainly not necessarily "main" in any sense, may not even occur in sample, or may even be logically impossible. Nor does β_{xz} embody the "interactive" effect of *x* or of *z* exactly, as often suggested, although this statement is perhaps less problematic. The coefficient β_{xz} does indicate by how much the effect of *x* changes per unit increase *z*, and the logically and mathematically identical amount by which a unit increase in *x* changes the effect of *z*, but this is not precisely an effect. It is a statement of how an effect changes; an effect on an effect. In an interactive model, indeed in all models, the effect of a variable, *x*, on *y* is *dy/dx*: full stop. Here that effect is $\beta_x + \beta_{xz} z$: again, full stop. The sign and magnitude of β_{xz} tells us how that effect varies according to values of *z*, and so its significance tests whether *z* (linearly) conditions the effect of *x* on *y* (and *vice versa*), as noted next, but in no sense can one distinguish some part of this one conditional effect as main and another part as interactive.

Returning to our empirical example of the interaction between institutional structure and social cleavages in determining the number of presidential candidates, we are now prepared to interpret the results using differentiation. Recall that the theoretical model in equation [10] produces estimates that can be expressed as:

$$\hat{y} = \hat{\gamma}_0 + \hat{\beta}_x x + \hat{\beta}_z z + \hat{\beta}_{xz} xz \qquad [13]$$

Where $\hat{y}$ refers to the predicted values of *y*, and $\hat{\gamma}_0$, $\hat{\beta}_x$, $\hat{\beta}_z$, and $\hat{\beta}_{xz}$ are the OLS estimates of γ_0, β_x, β_z, and β_{xz} respectively.

Recall the results from our OLS regression:[17]

$$\textit{Predicted Candidates} = 4.303 - 0.979\textit{Groups} - 2.491\textit{Runoff} + 2.005\textit{Groups} * \textit{Runoff} \; [14]$$

Applying [11] and [12], we see that

$$d\hat{y}/dg = -0.979 + 2.005\textit{Runoff} \qquad [15]$$

$$d\hat{y}/dr = -2.491 + 2.005\textit{Groups} \qquad [16]$$

Given our results, we can thus see that the effect of Ethnic Groups on the number of presidential candidates varies with the presence or absence of a runoff, and the effect of a runoff on the number of presidential candidates varies with the number of ethnic groups in a society.

[17] Although, technically, one cannot strictly differentiate with respect to non-continuous variables, such as dummy variables, one can proceed ignoring this technicality without being misled, and we advise researchers to do so. (Do remember, however, that marginal increases cannot actually occur, only unit increases from 0 to 1 can.)

These conditional effects can be easily calculated by inserting substantively relevant values for the variables of interest into equations [15] and [16].

Recall that *Runoff* takes two values in the dataset: 0 in the absence and 1 in the presence of a runoff. Hence, we can calculate the effect of Ethnic Groups on the effective number of presidential candidates by calculating the conditional effect for the substantively (and logically) relevant values of Runoff. When Runoff = 0, $d\hat{y}/dg = -0.979 + 2.005 * 0 = -0.979$. When Runoff = 1, $d\hat{y}/dg = -0.979 + 2.005 * 1 = 1.026$. In this example, we can see that in the absence of a runoff, the conditional effect of ethnic groups is negative; in the presence of a runoff, the conditional effect of ethnic groups is positive. (The standard errors of these estimated effects and whether these effects are statistically significant are matters we discuss below.)

We can calculate the effect of runoff systems on the number of presidential candidates by again calculating the conditional effect for logically relevant values of *Groups*. Recall that *Groups* ranges between 1 and 2.756 in our dataset. Accordingly, we might want to present the estimated effect of *Runoff* from the sample minimum to the sample maximum values of *Groups*; or at evenly spaced intervals starting from the sample minimum to some logically relevant value; or at the minimum, mean, and maximum; or at the mean, the mean plus and the mean minus a standard deviation or two; or at any other substantively revealing set of values for *Groups*.

To take one of these options, we calculate conditional effects when *Groups* ranges from 1 to 3, at evenly spaced intervals of 0.5, which yields the following estimated conditional effects:[18]

When Groups = 1: $d\hat{y}/dr = -2.491 + 2.005 * 1 = -0.486$

When Groups = 1.5: $d\hat{y}/dr = -2.491 + 2.005 * 1.5 = 0.517$

[18] Although the sample maximum is 2.756, *Ethnic Groups* does extend beyond this value when we examine non-presidential systems (which Cox (1997) analyzes as well).

When Groups = 2: $d\hat{y}/dr = -2.491 + 2.005 * 2 = 1.520$

When Groups = 2.5: $d\hat{y}/dr = -2.491 + 2.005 * 2.5 = 2.522$

When Groups = 3: $d\hat{y}/dr = -2.491 + 2.005 * 3 = 3.525$

In this example, we can see that at the sample minimum (when there is only one ethnic group in the society), the runoff has a negative effect on the number of presidential candidates (which, again, seems substantively odd), but as the number of ethnic groups rises, the runoff begins to affect the number of presidential candidates positively (which is more sensible). Further, the size of the effect grows as ethnic groups become more numerous (also sensible). Again, the standard errors of these estimated effects and whether the effects are statistically significant are matters we will discuss below.

2. Interpreting Effects through Differences in Predicted Values.

A second strategy for examining the effects of *x* and *z* on *y* consists of examining differences in predicted values of *y* for logically relevant and substantively meaningful values of *x* and *z*. This strategy does not require the researcher to have working knowledge of even basic calculus; it is a bit more tedious but quite serviceable in its own right. Predicted values of *y* can be calculated by substituting the estimated values for the coefficients along with logically relevant and substantively revealing values of the covariates of interest. Taking our theoretical model (equation [10]) and substituting in estimated coefficient values:

$$\hat{y} = \hat{\gamma}_0 + \hat{\beta}_x x + \hat{\beta}_z z + \hat{\beta}_{xz} xz \qquad [17]$$

The researcher can calculate predicted values of *y* using varying values of *x* (between, say, x_a and x_b) while holding *z* constant at some meaningful value (e.g., the mean value, or some other interesting value, if, for example, *z* is a dummy and setting it at a mean value would be less meaningful). By doing so, the researcher can see how changes in *x* (from x_a to x_b) cause changes in $\hat{y}$ (from $\hat{y}_a$ to $\hat{y}_b$). Recall that as x changes from x_a to x_b, while z is held at some meaningful value, say z_0, this also implies that xz changes from $x_a z_0$ to $x_b z_0$. Similarly, the researcher can identify how $\hat{y}$ moves with changes in z (and xz) when x is held at some meaningful value.

Using our empirical example, we can examine how the predicted number of presidential candidates changes as we increase the number of ethnic groups, both in the presence and in the absence of runoff elections:

$$\textit{Predicted Candidates} = 4.303 - 0.979\textit{Groups} - 2.491\textit{Runoff} + 2.005\textit{Groups} * \textit{Runoff}$$

For example, when the number of ethnic groups = 1 and runoff = 0, we calculate the predicted effective number of candidates, $\hat{y}$ as:

$$(\hat{y} \mid \textit{Groups} = 1, \textit{Runoff} = 0) = 4.303 - 0.979 * 1 - 2.491 * 0 + 2.005 * 1 * 0 = 3.324$$

Table 2 presents the predicted effective number of presidential candidates, as *Groups* ranges from 1 to 3, when *Runoff* takes values of 0 and 1.

[TABLE 2 ABOUT HERE]

By examining changes in $\hat{y}$, the researcher can discern how the independent variables, *Groups* and *Runoff*, influence the predicted number of presidential candidates. By looking across single rows, we see the impacts of the presence of a runoff system at different numbers of *Groups*. When the value of *Groups* is at its minimum (a value of one), a runoff system has a small and negative effect, decreasing the number of parties by -0.486 (that same substantively odd result again). When the value of *Groups* is at a higher value, say 2.5, the impact of a runoff system is larger in magnitude and positive: in the presence of 2.5 *Groups*, the runoff system is estimated to increase the number of presidential candidates by 2.522 (the sensible one).

By looking down single columns, we see the impacts of changes in the number of ethnic groups, in the absence of a runoff system or in the presence of a runoff system. Here, we see that in the absence of a runoff system, a rise in the number of ethnic groups (from, say, 1 to 3) coincides with a decline in the number of presidential candidates (from 3.324 to 1.366: oddly). In the presence of a runoff system, however, a rise in the number of ethnic groups (from 1 to 3) coincides with an increase in the number of presidential candidates (from 2.838 to 4.891: sensibly). Once again, standard errors for these estimated changes and whether they are statistically distinguishable from zero will be addressed below.

These exercises in interpretation of coefficients should underscore the point that the variables in interactive specifications have varying effects. The size and sign of the effect of *x* can depend critically upon the value at which the other variable, *z*, is held; conversely, the size and sign of the effect of *z* can depend critically upon the value at which the other variable, *x*, is held. Calling one of the coefficients involved in these effects the "main effect" and another the "interactive effect" can be quite misleading and is no substitute for understanding the model's actual estimated effects. Differentiation and/or differences of predicted values offer the researcher simple alternative tools for examining the effect of variables *x* and *z* on *y* in general, and in interactive models in particular. But once we have calculated the estimated effects, how can we express the statistical certainty of those estimates? I.e., how can we calculate their standard errors? Are these estimated effects statistically significant? I.e., are these estimated effects distinguishable from zero? We now turn to these questions.

C. LINKING STATISTICAL TESTS WITH INTERACTIVE HYPOTHESES

In examining existing political-science articles published in the top journals of the field, we have found that statistical exploration of interactive propositions relied primarily on *t*-tests of significance of individual coefficients in the model. *I.e.*, researchers commonly compare each of the three key coefficient estimates in a typical interactive model, *e.g.* $\hat{\beta}_x$, $\hat{\beta}_z$, and $\hat{\beta}_{xz}$ in [17], to its respective standard error. Assuming the model exhibits the necessary statistical properties otherwise (*i.e.*, the Gauss-Markov conditions), the ratios in this comparison are *t*-distributed (or asymptotically normal), so these tests are statistically valid (asymptotically), yet scholars often mistake their meaning—i.e., they often mistake what these are tests of—reflecting the persistent confusion of coefficients for effects and use of the misleading main- and interactive-effect terminology. Just as the effects of variables involved in interactive terms vary according to two (or more) coefficients and the values of (an)other variable(s), so too do their standard errors and so the relevant *t*-statistics, confidence intervals, and hypothesis-test results (significance levels).

Thus, single *t*-tests on individual coefficients on variables involved in interactive terms require care to interpret because they refer to significance at only one empirical value of the other variables. For example, β_x and β_z in our standard model [17] indicate, respectively, *x*'s effect on *y* when *z* equals zero and *z*'s effect on *y* when *x* equals zero, so the standard *t*-tests on our estimates $\hat{\beta}_x$ and $\hat{\beta}_z$ indicate the significance of that variable's effect when the other variable equals zero, which substantive values as noted above, may be substantively, empirically, or even logically irrelevant.

In our empirical example predicting effective number of presidential candidates with social cleavages, electoral systems, and the product of the two variables, we noted that the number of ethnic groups in the sample never takes the value of 0 and logically could not do so. Given the discussion in the preceding paragraph, we immediately see that any inferences drawn about the statistical significance of our estimate of β_R (the coefficient on *Runoff*) are largely meaningless because they refer to a condition that does not and could not exist. Just as the interpretation of the coefficient is meaningless because it refers to a value that is logically impossible, any judgment of statistical significant based on said coefficient would also be logically irrelevant. Inferences drawn about the statistical significance from our estimate of coefficient β_G refer to the impact of *Ethnic Groups* in the specific case where *Runoff* equals zero. When there is no runoff system, the number of ethnic groups decreases the effective number of presidential candidates, with a p-value of 0.228. In this case, the interpretation of the statistical significance of the coefficient applies to a logically possible case that does in fact exist in our data and has important substantive meaning, but we remind the researcher that the judgment of statistical significance is still a limited one: it applies only to the effect of *Ethnic Groups* when *Runoff* equals zero and says nothing about that effect for other values of *Runoff* (here, the only other possible, meaningful value is *Runoff*=1).

To provide a framework for hypothesis testing of effects rather than coefficients in interactive models, consider the following types of theoretical questions that researchers often ask:

(1) Does *y* depend on *x*, or, equivalently, is *y* a function of *x*? Does *y* depend on *z*, or, equivalently, is *y* a function of *z*?

(2) Is *y*'s dependence on *x* contingent upon or moderated by *z,* or, equivalently, does the effect of *x* on *y* depend on *z*? Is *y*'s dependence on *z* contingent upon or moderated by *x,* or, equivalently, does the effect of *z* on *y* depend on *x*? **This is the classic interactive hypothesis; the symmetric pairs of questions are logically identical.**

(3) Does *y* depend on *x, z*, and/or their interaction, *xz*, at all, or, equivalently, is *y* a function of *x*, *x*, and/or *xz*?

Tables 3-5 link each of these sets of theoretical and substantive questions about interactive relationships to the mathematical expression of those relationships to the statistical tests in those empirical models corresponding to those questions, starting with the simpler propositions in

Table 3.

[TABLE 3 ABOUT HERE]

Notice that in Table 3, the statistical test that corresponds to each hypothesis holds the opposite of the mathematical expression for that hypothesis as the null; *i.e.*, as always: the null hypothesis is what one would like, theoretically, to reject statistically.

The first hypothesis examines whether *x* has any effect on *y*. The mathematical expression for the estimated effect of *x* on *y* includes $\hat{\beta}_x$ and $\hat{\beta}_{xz}z$ as stressed above. The standard *F*-test on the pair of coefficients, $\hat{\beta}_x$ and $\hat{\beta}_{xz}$, identifies whether *x* matters (*i.e.*, whether *y* depends on *x*). Only these coefficients both being zero would imply *y* does not depend on *x* in some fashion in this model.

The "simple" hypothesis that the effect of *x* on *y* is positive or negative, is actually ill-defined in linear-interactive models because, first, the effect*s* of *x* vary depending on values of *z*, and, second, they do so linearly, implying that the effects will be positive for some and negative for other *z* values (although not all values of *z* need necessarily be possible logically, occur empirically in sample, or be substantively meaningful, as noted above). Thus, no common practice exists for testing hypotheses that *x* or *z* generally increases or decreases *y* in linear-interactive models; we can now see this is because such hypotheses are mathematically (*i.e.*, logically) ambiguous in such models. Mathematically, the effect of *x* on *y* in a linear-interactive model can be negative for some values of *z* and positive for others, depending on the range of values taken on by the independent variables. Therefore, whereas the effect of *x* on *y* may statistically differ from zero at some values of *z*, it can be statistically indistinguishable from zero at some other values of *z* (*e.g.*, that value of *z* where the estimated effect of *x* on *y* is zero). To take one proposition often seen in many substantive contexts as a very illustrative example, suppose we hypothesized that x had increasingly positive effect on y as z increased, starting from no effect at z=0, and suppose even that z<0 is logically impossible. In this case, even if that proposition were completely true and the evidence were very strongly to support it, and even if we ignored the fact that the estimated effect of x would be negative at the (logically impossible) values z<0, the estimated effect of x on y would be zero at z=0 and therefore necessarily insignificant (i.e., insignificantly different from zero) at that point. Moreover, the estimated effect also must be insignificantly different from zero for some range near z=0, given that all estimates have some error. The range where this insignificance holds would obviously be larger the less precisely the effect of x given z is estimated, i.e., the less strongly the evidence supported the proposition, but even if the evidence extremely strongly supported the hypothesis that the effect of x was (weakly) positive and increasing in z, there would be some, likely appreciable, range of permissible values of z where we could not reject that the effect was zero. In a nutshell, hypotheses that the effects of x (or z) are generally positive or negative are poorly specified for a linear-interactive model.

For empirical testing of such propositions in a linear-interactive model, therefore, scholars must state their hypotheses more precisely to refer to some values or range of values of z.[19] Researchers can then calculate measures of uncertainty to determine whether the effect of x at some value of z is statistically distinguishable from zero. This approach is highlighted in the second and third hypotheses in Table 3. To make a general claim that the effect of x on y is

[19] Alternatively, the researcher could simply estimate a linear-additive model that omits the interaction in question and simply test whether the coefficient on x or z significantly differs from zero in the usual manner. If the interaction truly exists, the linear-additive model would tend to produce for coefficients on x and z their average effect across the sample values of the other variable, but with some attenuation bias and inefficiency due to the mis-specification of the truly linear-interactive model as linear-additive. If the interaction does truly exist, however, the researcher must note that this linear-additive model is mis-specified with the coefficient estimates on x and z very like subject to attenuation bias and inefficiency. Accordingly, these tests would tend to be biased toward failing to reject.

positive or negative, the researcher could test that the effect of *x* on *y* is positive over the entire logically possible, or substantively sensible, or sample range of *z*. To do so, the researcher should conduct several *t*-tests suggested in Table 3, calculated at several *z* values, over that entire range. Alternatively, but equivalently, one could plot *dy/dx* over the appropriate range of *z* along with confidence intervals, which latter would indicate rejection at that confidence-level if they contained zero at any point across that range of *z*. However, as noted above, in some cases, we would expect failure to reject (confidence intervals that span zero) at some levels of *z* even if the hypothesis were completely true and very strongly supported by the data.[20]

To execute this set of t-tests or generate these confidence intervals, the researcher will first need to differentiate *y* with respect to *x* or calculate predicted values of *y* at different values of *x*, holding *z* at some logically relevant value. Using the differentiation method, the estimated effect of *x* on *y* in [17] is $d\hat{y}/dx = \hat{\beta}_x + \hat{\beta}_{xz}z$. Given that OLS specifies that *z* is fixed in repeated sampling (or, roughly equivalently, that we interpret our estimates given *z*) and that our estimates of β carry some level of uncertainty, our estimate of the effect of *x* on *y* will have some level of uncertainty that depends on *z*. That is, just as the effects of *x* on *y* vary with the values of *z*, the standard errors of the effects of *x* on *y* also vary with values of *z*. We can express the uncertainty as a standard error or variance, or as a confidence interval around the marginal effect. Just as no single effect attributable to *x* exists when *x* interacts with another variable, no single variance exists either for the marginal effect of *x* when *x* interacts with another variable. Each unique

[20] Recognizing this fact, we suggest that researchers plot the estimated effects of *x* across the sample, logical, or substantively meaningful range of *z*, along with confidence intervals as described below, and then consider the share of these confidence intervals' covered area that lies above (or below) zero as indication of how strongly the evidence supports the proposition of a "generally" positive (or negative) effect. Since "generally" is imprecise and involves judgment, this test is imprecise and involves judgment too, but visualizing graphically the proportion of a confidence area that lies above or below zero should help in rendering this judgment.

value in the set of estimated marginal effects (one at each value of *z*) will have its own variance and corresponding standard error. Therefore, obviously, the statistical significance of these effects (*t*-statistics and confidence intervals) is also conditional on the substantive value of *z*, implying that the effect of *x* on *y* may be significant for *z* values in some range and insignificant in other ranges, as we have already explained. Our interpretation and presentation of results must clearly reflect this, and, again, no substitute exists for some basic math in doing so.

The variance of $d\hat{y}/dx$ is as follows:[21]

$$Var(d\hat{y}/dx) = Var(\hat{\beta}_x) + z^2 Var(\hat{\beta}_{xz}) + 2zCov(\hat{\beta}_x, \hat{\beta}_{xz}) \qquad [18]^{22}$$

Our level of uncertainty about the effect of *x* on *y* is some function of the variability in our parameter estimates of β_x, the variability in our parameter estimates of β_{xz}, the level of covariance between our estimates of β_x and β_{xz}, and the values of *z* at which the effects are evaluated. Var($\hat{\beta}_x$) is simply the square of the standard error that is typically reported in regression output. Var($\hat{\beta}_{xz}$) is, similarly, the square of the standard error that is reported for the

[21] If following the predicted-value method, one should distinguish between the error in $E(y|x,z{=}z_0)-E(y|x,z{=}z_1)$ and the prediction or forecast error in $E(y|x,z)$. The former refers predicted differences in *y* given *x* and *z* as the value of *z* changes. The former refers to a single prediction of *y* given *x* and *z*. Predictions, the latter case, will include two sources of error and variance, one due to estimation error in the $\hat{\beta}$ and one due to residual error, i.e. the stochastic residual/error term in the regression model. In predicted differences, the former case and the one more useful in illustrating estimated interactive effects as we have suggested, the differencing removes the second source of error because the same error is on both sides of the minus sign. The error in the estimated conditional effects of *x* in the linear-interactive model, as in estimated effects in the linear-additive model or, indeed, any model with additively separable stochastic component, contains only the first source of error, the estimation error. By the derivative method, this separability is more obvious because $\frac{dy}{dx} = \frac{d(XB+\varepsilon)}{dx} = \frac{d(XB)}{dx} + \frac{d\varepsilon}{dx} = \frac{d(XB)}{dx} + 0 = \frac{d\hat{y}}{dx}$.

[22] Given some constant *c* and some random variable *r*, $Var(cr) = c^2 Var(r)$. Given some constant *c* and two random variables r_1 and r_2, the variance of the expression $Var(r_1 + cr_2) = Var(r_1) + c^2 Var(r_2) + 2cCov(r_1, r_2)$. In our context, the *x* and *z* are fixed in repeated sampling, per the standard OLS assumptions, and the coefficients are the random variables. More generally, for random vector, $\hat{\beta}$, and constant vector, *X*, the variance of the linear-additive function $X\hat{\beta}$ is $Var(X\hat{\beta}) = X'Var(\hat{\beta})X$. The reader familiar with matrix algebra can verify that [18] is just one specific example of this more general formula.

coefficient estimate, $\hat{\beta}_{xz}$. The covariance between $\hat{\beta}_x$ and $\hat{\beta}_{xz}$, Cov($\hat{\beta}_x, \hat{\beta}_{xz}$), however, is *not* typically shown as part of the standard regression output and must be extracted from the estimated variance-covariance matrix of the coefficient estimates.

Recall that the measures of uncertainty that accompany OLS regression coefficient estimates are selected from the diagonal elements in the estimated variance-covariance matrix, which is computed using $s^2(X'X)^{-1}$, where s^2 is our estimate of σ^2, the variance of ε. The variance-covariance matrix $s^2(X'X)^{-1}$ is a symmetric matrix that contains the estimated variance of each estimated coefficient along the diagonal elements and the estimated covariance of each estimated coefficient with the other estimated coefficients in the off-diagonal elements:

$$\begin{bmatrix} Var(\hat{\beta}_1) & & & \\ Cov(\hat{\beta}_1, \hat{\beta}_2) & Var(\hat{\beta}_2) & & \\ \vdots & & \ddots & \\ Cov(\hat{\beta}_1, \hat{\beta}_k) & Cov(\hat{\beta}_2, \hat{\beta}_k) & \cdots & Var(\hat{\beta}_k) \end{bmatrix}$$

Returning to Equation [18], we see that our estimate of Cov($\hat{\beta}_x, \hat{\beta}_{xz}$) will appear in the off-diagonal representing $\hat{\beta}_x$ and $\hat{\beta}_{xz}$. When some software estimates an OLS regression, it stores the estimated variance-covariance matrix of the parameter estimates in memory, and the researcher can usually easily recall that estimated variance-covariance matrix through a single post-estimation additional command.

We reiterate that the standard errors of the effects of variables involved in interaction terms will vary depending on the values of other variables in the interaction, on the standard errors of the coefficients involved in that effect, and on the covariances of those coefficients.

To execute the tests or construct the confidence intervals suggested in the second and third rows of Table 3, the researcher would calculate the effect at some given value of z

($d\hat{y}/dx = \hat{\beta}_x + \hat{\beta}_{xz}z$) and calculate the estimated variance around that effect at the given value of z, $\hat{V}(d\hat{y}/dx)$. The researcher could then calculate the *t*-statistic for testing whether this estimated effect is statistically distinguishable from zero, by following the standard procedures: dividing the estimated effect $d\hat{y}/dx$ by the standard error of $d\hat{y}/dx$ and evaluating the resulting t-statistic against the t-distribution (with *n-k* degrees of freedom, where *n* refers to the number of observations and *k* refers to the number of regressors in the model, including the constant). The researcher would then repeat the process for another value of *z* to determine whether a general claim can be made about the direction of the effect (e.g., if *x* increases *y* over some set of *z*, or if *x* decreases *y* over some set of *z*).

To explore whether the effect of *x* on *y* is "generally," "typically," or "on-average" positive or negative, a common component of the typical complex of interactive hypotheses, requires more precise definition of the terms in quotations. If "on-average" refers to the effect at the sample-average value of *z*, then the single *t*-test of the effect of *x* at that value of *z* suffices. This value of *z* also gives the appropriate estimated effect and its statistical confidence for an "on-average" effect taken to mean the average in the sample of the effect of *x*. Incidentally, as we show in a later section, this hypothesis corresponds with the standard *t*-statistic reported for the coefficient on *x** in an interactive model where *x* and *z* have been mean-centered (had their sample means subtracted). If, however, one wishes to gauge the statistical certainty with which the hypothesis that *x*'s effect on *y* is "generally" or "typically" positive, we suggest plotting *dy/dx* over the sample range of *z* and *vice versa*, with confidence intervals along the entire range. The hypotheses that *dy/dx* is "generally" or "typically" positive or negative corresponds to most (unfortunately, no firm cut-off share exists) of this confidence interval lying appropriately above or below zero. One could obtain greater precision by quantifying the share of the area covered by the confidence interval that lies above or below zero to give more precision to this exploration, but, fundamentally, terms like “generally” and “typically” inherently entail exactly the imprecision that one cannot avoid here. An alternative strategy would be to estimate a different model, one without the interaction term(s), and simply evaluate the usual *t*-test on the appropriate coefficient, on *x* or on *z*. This alternative would reveal directly whether, “on average” or “generally,” *x* or *z* had a non-zero effect on *y*. However, if the true relationship really is interactive, then this alternative model is mis-specified, so these *t*-tests would be, at minimum, inefficient. (See also note 20.)

Aside from the most basic hypotheses that *x* affects *y* or that *x* increases/decreases *y* (over some range of *z*), researchers exploring interactive propositions are interested in whether and how the effects of *x* and of *z* on *y* depend on the other variable. Table 4 presents these sorts of hypotheses.

[TABLE 4 ABOUT HERE]

Notice two important facts from Table 4. First, because the coefficient on *xz* directly reflects the presence, sign, and substantive magnitude of a conditioning relationship, *i.e.*, the degree to which the effects of *x* and *z* on *y* depend on the other variable’s value, the standard *t*-test of the coefficient on the multiplicative term directly tests for the presence or sign of such a conditioning relationship. Second, the mathematical expression and the statistical test for the hypothesis that *x* conditions the effect of *z* on *y* are identical to those for the converse that *z* conditions the effect of *x* on *y*. This reflects the logical symmetry of interactive propositions. If *z* conditions the effect of *x* on *y*, then *x* logically must condition the effect of *z* on *y* and in the same amount. In fact, the second three rows of Table 4 simply state the logical converses of the first three rows, so the corresponding mathematical expressions and statistical tests are identical.[23]

As we established above, the effect of *x* on *y* is $dy/dx = \beta_x + \beta_{xz} z$, so a simple *t*-test of the null hypothesis that $\beta_{xz} = 0$ evaluates whether the effect of *x* changes as z changes. Given that β_x refers to the effect of *x* on *y* in the specific case when *z* is zero, β_{xz} displays the difference from that in the effect of *x* on *y* when *z* is nonzero. In essence, β_{xz} indicates whether the effect of *x* is the same when *z* is zero, as when *z* takes any other value. To restate: β_{xz} identifies whether the effect of *x* changes as *z* takes various values. A rejection of the null hypothesis that $\beta_{xz} = 0$ is consistent with the most central interactive hypothesis: the effect of *x* on *y* differs depending on the level of *z*.

Researchers are often, however, interested in hypotheses that contain a directional element: that the effect of *x* on *y* increases as *z* increases, or the effect of *x* on *y* decreases as *z* increases. Recalling that the effect of *x* on *y* is: $dy/dx = \beta_x + \beta_{xz} z$, a hypothesis that the effect of *x* on *y* increases as *z* increases could be tested with a one-tailed null hypothesis that $\beta_{xz} \leq 0$. Symmetrically, the hypothesis that the effect of *x* on *y* decreases as *z* increases could be tested with a one-tailed null hypothesis that $\beta_{xz} \geq 0$. These directional hypotheses are displayed in the second and third lines of Table 4.[24]

Finally, Table 5 reveals that the statistical test corresponding to the broadest sort of hypothesis one might have regarding an interactive model: that *y* depends in some manner,

[23] The order of differentiation in a cross-derivative never matters, so this symmetry does not rely on the linear-multiplicative form specifically. In any logical proposition/mathematical model, the effect of *x* depends on *z* implies that the effect of *z* depends, in identical fashion, on *x*: $d(dy/dx)/dz \equiv d(dy/dz)/dx$ for any function $y(x,z)$.

[24] Since assuming directionality in this way lowers the empirical hurdle for statistical rejection, many scholars opt more conservatively for always employing non-directional hypotheses and two-tailed tests.

linearly-additively and/or linear-interactively, on *x* and/or on *z*. In common language, some one or combination of *x* and *z* matters for *y*. This corresponds statistically, quite simply, to the *F*-test that all three coefficients involved in the interaction, $\beta_x, \beta_z, \beta_{xz}$, are zero. That all three of these are zero is the only condition that would render *x* and *z* wholly irrelevant to *y*.

Returning to our empirical example*,* let us walk it through the various tests outlined in Tables 3, 4, and 5.

First, does *x* affect *y*? Does the number of presidential candidates depend in some linear or linear-interactive way on the number of ethnic groups? To examine this general question, we conduct an *F*-test. The *F*-test will test the null hypothesis that $\beta_G = 0$ and $\beta_{GR} = 0$. The *F*-test produces the following results: F(2, 12) = 2.62; Prob > F = 0.1140. Whether to reject the null hypothesis depends on the researcher's desired level of certainty. At conventional levels ($p<0.10$, $p<0.05$, $p<0.01$), this null hypothesis would not (quite) be rejected.[25]

Does *z* affect *y*? Does the number of presidential candidates depend in some linear or linear-interactive way on the presence of a runoff system? The *F*-test on the null hypothesis that $\beta_R = 0$ and $\beta_{GR} = 0$ yields the following results: F(2, 12) = 2.96; Prob > F = 0.0903, which would (barely) satisfy a $p<0.10$ criterion, but fail the stricter $p<0.05$, $p<0.01$ criteria.

Next, we might be interested in whether *x* (generally) increases *y*. To answer this question, the researcher should conduct *t*-tests of or construct confidence intervals for the effect of *x*, across some range of values of *z* (corresponding to "generally"). To conduct these *t*-tests, one must first calculate the standard errors associated with the given marginal effect following

[25] A less-strictly classical approach to hypothesis testing would simply report the p-level and leave the reader to decide whether this is "significant enough" or, more generally, how much weight to assign a result with this level of statistical significance.

equation [18]. To do so, we first need to access the estimated variance-covariance matrix of the estimated coefficients. Recall that this symmetric matrix contains the estimated variance of each of the estimated coefficients along with the estimated covariance between each pair of estimated coefficients. Table 6 displays the estimated variance-covariance matrix from our example.[26]

[TABLE 6 ABOUT HERE]

The element in the first row and first column, 0.593, is the estimated variance of the coefficient estimate for β_G, which is the square of the standard error of $\hat{\beta}_G$, 0.770. The information we need from the variance-covariance matrix in this case that we do not see in the typical regression output is Cov($\hat{\beta}_G$, $\hat{\beta}_{GR}$), which is -0.593 in our empirical example. To calculate the variance of the estimated marginal effects, we simply substitute these values from the estimated variance-covariance matrix into equation [18].

$$Var(d\hat{y}/dg) = Var(\hat{\beta}_G) + Runoff^2 Var(\hat{\beta}_{GR}) + 2 * Runoff * Cov(\hat{\beta}_G, \hat{\beta}_{GR})$$

$$Var[(d\hat{y}/dg) \mid Runoff = 0] = 0.593 + 0^2 * 0.885 + 2 * 0 * -0.593 = 0.593$$

$$Var[(d\hat{y}/dg) \mid Runoff = 1] = 0.593 + 1^2 * 0.885 + 2 * 1 * -0.593 = 0.292$$

Testing the proposition that ethnic groups increase the number of presidential candidates requires positing the following null hypothesis: H$_0$: $\beta_G + \beta_{GR} Runoff \leq 0$. This null hypothesis can be evaluated at the two values of *z*: 0 (the absence of a runoff system) and 1 (the presence of a runoff system). These results appear in Table 7.

[TABLE 7 ABOUT HERE]

With a one-tailed p-value of 0.884, we cannot reject the null hypothesis that

[26] Appendix B provides step-by-step STATA commands.

$\beta_G + \beta_{GR} Runoff \leq 0$ when *Runoff* = 0. In the absence of a runoff system, one cannot reject the possibility that there is a negative or null relationship between *Groups* and *Candidates*. However, with a one-tailed p-value of 0.041, we can reject the null hypothesis that *Groups* decrease or have no effect on *Candidates* when *Runoff* = 1, in favor of the alternative that some positive relationship between *Groups* and *Candidates* exists when *Runoff* = 1. To test the reverse directional hypothesis: that the number of ethnic groups decreases the number of presidential candidates, we would pose the following null hypothesis: $\beta_G + \beta_{GR} Runoff \geq 0$ and re-evaluate the t-statistics against this null. In the absence of a runoff system, the one-tailed p-value is 0.116, which actually (substantively oddly, as we've noted) approaches significance. In the presence of a runoff system, the one-tailed p-value of 0.959 suggests that we are quite unable to reject the null hypothesis of a positive or null relationship between *Groups* and *Candidates*.

Next, the researcher might be interested in testing directional hypotheses with respect to the effect of a runoff system on the number of presidential candidates. To consider whether a runoff system increases the number of presidential candidates, the researcher could conduct a number of *t*-tests over a logically relevant range of *Groups*. These are displayed in Table 8.

[TABLE 8 ABOUT HERE]

Here, we see that evaluation of the null hypothesis of $\beta_R + \beta_{GR} Groups \leq 0$ changes for various values of *Groups*. As the number of ethnic groups increases, our ability to reject the null hypothesis also increases. One can see in this table that, only when *Groups* takes values over 3 does the hypothesis test begin to approach conventional significance levels. However, *Groups*=3 actually exceeds the realized sample maximum. The estimated effect, as such, is a prediction into what we might see if the number of ethnic groups reached that level (as it does in some non-presidential systems).

If we wanted to investigate the proposition that the effective number of ethnic groups decreases the number of presidential candidates, we would re-evaluate the t-statistics against the following null hypothesis: $\beta_R + \beta_{GR} Groups \geq 0$. The one-tailed p-values above show that we cannot remotely reject the null hypothesis in any case, thus lending no support at all to the proposition that the number ethnic groups decreases the number of presidential candidates.

To continue to the tests outlined in Table 4, the researcher might be most interested in the question of whether the effect of *Groups* on *Candidates* depends in some way on the presence or absence of a runoff system. The answer to this substantive question emerges directly from the coefficient on the interactive term, β_{GR}, and its standard error. A two-tailed test of the null hypothesis H_0: $\beta_{GR} = 0$ is in order. The p-value associated with this two-tailed test is 0.054. The determination of "statistical significance" in this case depends on the researcher's desired level of uncertainty: rejection at the $p<0.10$ threshold, near rejection at a $p<0.05$ threshold, and failure to reject at the tighter $p<0.01$ level. As discussed above, the mathematical symmetry in interaction terms also implies that if the researcher wanted to ascertain whether the effect of a presidential runoff system depends in some way on the number of ethnic groups, then the (same) answer also lies in the significance of that same interaction term, β_{GR}.

A directional hypothesis can also be posited. If the researcher were interested in whether the effect of *Groups* <u>increases</u> in the presence of a runoff system, then a one-tailed test is appropriate, using the following null hypothesis of H_0: $\beta_{GR} \leq 0$. The one-tailed test of the null hypothesis leads produces a p-value of 0.027, which suggests that the researcher can reject the null hypothesis of a negative or nonzero coefficient in favor of the alternative hypothesis of some positive coefficient. Thus, the results are consistent with an interpretation that the effect of *Groups* increases in the presence of a runoff system. The logical symmetry of the interaction term also implies that the results are consistent with an interpretation that the effect of *Runoff* increases as the number of ethnic groups increases.

Finally, let us examine the test outlined in Table 5: whether *x* and *z* have any effect on *y* in some linear or linear-interactive fashion. Here, the researcher is interested in whether *Groups*, *Runoff*, and/or their product have any effect on *Candidates*. An *F*-test on all three coefficients (β_G, β_R, and β_{GR}) will speak to the null hypothesis that $\beta_G = \beta_R = \beta_{GR} = 0$. The *F*-test yields the following results: F(3, 12) = 2.27, with an associated p-value of 0.132.

D. Presentation of Interactive Effects

The major point we convey in this section is that the mere presentation of regression coefficients and the accompanying standard errors is wholly inadequate. Hayduk and Wonnaccott (1980) noted that, “While the technicalities of these [interactive] procedures have received some attention…the proper methods for the interpretation and visual presentation of regressions containing interactions are not widely understood” (400). Unfortunately, their observation still holds too widely today, and in this section we aim to provide clear guidance on effective means of presenting the results from models that include interaction terms.

We have described two methods for interpreting effects: derivatives and differences in predicted values. Common presentations of effects that involve interactive terms now often do involve tables or graphs that depict the effect of *x* on *y*, when *z* equals particular values. What is too often missing still from these presentations is indication of the estimated uncertainty of the estimated effect and its magnitude, across a sufficiently wide and meaningful range of values of *z*. These conditional effect estimates can be effectively conveyed in tabular and graphical forms—we believe the latter are especially useful, but also especially underutilized, in the literature—but both should always also include some measure(s) of certainty.

Existing statistical packages can provide marginal effects and predicted values (or conditional expectations) as well as standard errors of the marginal effects and predicted values, typically as part of some post-estimation suite of commands. Further, other programs exist which will generate estimates of uncertainty around predicted values using simulation techniques (King, Tomz, and Wittenberg 2000). While we have no particular qualms against such pre-programmed commands and procedures,[27] the procedures we recommend below maximize the user's control over the values at which marginal effects and predicted values are calculated, and, we believe, will strengthen the user's understanding and intuition in interpreting models that contain interactive terms. We strongly recommend that the user be fully conversant with the elementary mathematical foundations underlying these procedures before taking pre-programmed commands "off the shelf."

1. Presentation of Marginal Effects.

We might be interested in conveying to the reader how the effect of *x* changes at various or along some range of *z* values. Recall that the estimated marginal effect of *x* consists of the first derivative based on the estimation results: $d\hat{y}/dx = \hat{\beta}_x + \hat{\beta}_{xz}z$. The marginal effect is, once again, a function of the estimated coefficient on *x*, the estimated coefficient on *xz*, and specific values of *z* at which we evaluate this effect. We will want to discuss the marginal effect of *x* over some substantively revealing range of values of *z*. One such revealing range and sequence of values, which may serve as a good default, might be to consider *z* taking evenly spaced values ranging from *a*, its minimum value in the sample, to *c*, its maximum value in the sample. More generally, the researcher could calculate the marginal effect of *x* on *y* for any set of values of interest for *z*.

[27] We would emphasize, however, that the researcher should verify that the uncertainty estimates produced by these procedures do not, as some unfortunately do, erroneously add stochastic error to estimation error in calculating the uncertainty of estimated effects in models with additively separable stochastic components (like linear regression).

Sample means, percentiles, means plus or minus one or two standard errors, etc. are all frequently useful default values for these considerations, but the substance and the researchers' presentational goals, not some relatively arbitrary default, should be determinate here. Using *z* values of particular observations—say, of some well-known, important, or illustrative case or cases in the sample—is also often a good idea. The researcher would then calculate the standard errors associated with each of these marginal effects. As a next step, the researcher could create a confidence interval around each estimated effect. Confidence intervals are usually more effective in graphical presentation and standard errors, t-statistics, or significance levels in tabular presentations, but visual clutter or researcher interest may reverse that (subjective) judgment. Confidence intervals can be generated with the following formula:

$$d\hat{y}/dx \pm t_{df,p}\sqrt{Var(d\hat{y}/dx)}$$

Where $t_{df,p}$ is the critical value in a *t*-distribution with *df* degrees of freedom (where *df* = *n-k*, where *n* refers to the number of observations and *k* refers to the number of regressors, including the intercept) that produces a p-value at which hypothesis tests are to be made and that produces ½ of the probability outside of the desired confidence interval. For example, to obtain lower and upper bounds of a 95% confidence interval, $t_{df,p}$ should correspond to critical values for a two-sided test of 0.05 level, i.e. one with 0.025 on each side; with large degrees of freedom, this will be approximately 1.96.

For our empirical example, we calculated two sets of marginal effects. The marginal effect of *Groups,* ($d\hat{y}/dg$), was calculated when *Runoff* equals 0 and when it equals 1. The marginal effect of *Runoff,* ($d\hat{y}/dr$), was calculated at various values of *Groups* (from 1 to 3, at evenly spaced intervals). To calculate a confidence interval, we need to calculate the variance of the marginal effect and to identify a desired level of confidence. For our empirical example, given the extremely small number of observations, we chose to accept lower certainty and so selected a 90% confidence interval. This interval implies a critical value of $t_{12,\alpha=.10}$ = 1.782. We would thus calculate the upper bound and lower bound for the confidence intervals as follows:

$$\text{Upper bound: } d\hat{y}/dx + 1.782 * \sqrt{Var(d\hat{y}/dx)}$$

$$\text{Lower bound: } d\hat{y}/dx - 1.782 * \sqrt{Var(d\hat{y}/dx)}$$

Tables 9 and 10 present the marginal effects, variance of the marginal effects, and 90% confidence intervals around these marginal effects.

[TABLE 9 ABOUT HERE]

[TABLE 10 ABOUT HERE]

If we are evaluating the marginal effect of *x* at several values of *z*, then a graphical display of marginal effects with confidence intervals would be effective. Figure 1 displays the marginal effect of *Runoff* for various values of *Groups* with confidence intervals around these estimated marginal effects. The straight line indicates the marginal effect of *Runoff* at given values of *Groups*, and the confidence intervals indicate the amount of uncertainty surrounding each marginal effect. The graph and accompanying confidence intervals thus convey to the reader both the marginal effects of *x* across values of *z* that are of interest and the degree of uncertainty that accompanies those estimated effects. This graph shows that over the range of sample-relevant values (varying *Ethnic Groups* from 1 to 3), the marginal effect of runoff increases, although the 90% confidence interval overlaps zero throughout the sample range.

[FIGURE 1 ABOUT HERE]

Recall that OLS estimates a *linear* relationship between the regressors and *y*, and these estimated relationships do not by themselves presume any bounding on the range of these regressors (here: *x* and *z*). As an exercise, to illustrate the mathematical properties of these effect lines and their associated standard errors, imagine if we were to extend the effect line out in both directions, by projecting into a much smaller number of ethnic groups and a much larger number of ethnic groups. Projecting into negative space with our independent variable is substantively nonsensical in this case, but it would not be for some variables, so let us imagine it were possible here, for illustrative purposes only. Calculating the marginal effect of *Runoff*, as the number of ethnic groups ranges from -4 to +6 produces Figure 2, illustrating several interesting properties.

The resulting graph shows a straight line (representing the marginal effect of *Runoff*) surrounded by two curves forming a confidence interval in the shape of an hourglass. First, note that the straight line (the marginal effect of *Runoff*) is a linear function of *Groups*; this is (part of) what it means to say these are linear-interactive models.[28] Second, as we noted above, the coefficient on *Runoff* indicates the impact of *Runoff* when *Groups* = 0. Thus, $\hat{\beta}_R$ =-2.49 is also our estimate of the intercept point for the marginal effect line (that is, the value on the y-axis when *Groups* takes on the value of zero), as the graph indicates. Third, notice that $\hat{\beta}_{GR}$ gives our estimate of the slope of the marginal effect line (2.01 according to our OLS estimates). Our coefficient estimate $\hat{\beta}_{GR}$ is positive and indicates that the effect of *Runoff* is estimated to increase at about 2 units *Candidates* for each one-unit change in *Groups*.

Next, note that the confidence intervals take an hourglass shape; this is characteristic of such conditional-effect lines. The smallest part of the hourglass is the point (i.e., the value of *z*) at which there is the highest degree of certainty concerning the size of the marginal effect (of *x* on *y*). This point, intuitively, will correspond to the mean in the sample of the counterpart term(s)

[28] If the effect of *Runoff* depended curvilinearly on groups, the linear-regression estimates of such linear-interactive models would give us a best (least-squares) linear approximation to the true curvilinear effect line. This is part of the reason that the common advice of "if including *xz*, then include *x* and *z*" is, indeed, a usually advisable practice, or at least something worth checking. See Section IV.A.

in the interactions (*z*); as always in regression analysis, we are most certain of our estimates for values around the mean (centroid for more-than-one variable) of our data. The wider parts of the hourglass are points at which there are lower degrees of certainty regarding the marginal effect, which, intuitively, are points farther from the mean (centroid) of the data. The characteristic hourglass shape of confidence region results from the appearance of z^2 in the expression for the variance of the effect and also from the covariance of the coefficient estimates in that expression, which is typically negative because the corresponding variables *x* and *xz* tend to correlate positively. Thus, the relative concavity of these hourglasses tends to sharpen with the magnitude of this correlation. In summary, the confidence intervals (regions) around conditional-effect lines will be (3D) hourglass in shape, with the "waist" located at the mean (centroid) of the conditioning variable(s) and generally becoming more accentuated as *x* and *xz* correlate more strongly, although this accentuation depends also on the relative (estimated) variances of $\hat{\beta}_R$ and $\hat{\beta}_{GR}$ and, in appearance anyway, graph and *z* scaling.

Note also from the graph above that the marginal effect of *Runoff* is statistically distinguishable from zero at negative values of *Groups* but is indistinguishable from zero at positive values of *Groups* until *Groups* exceeds about 5.5. These results illustrate clearly the following points made above. First, we see that the marginal effect of *Runoff* indeed varies with values of *Groups*. Second, we see that the effect lines, being linear, will extend above and below zero for some (not necessarily meaningful) values of the variables involved. Third, we see that our confidence regarding (i.e., standard errors and significance levels for) the marginal effect of *Runoff* also varies with values of *Groups*. Although we have plotted these effects and confidence intervals over ranges including substantively or even logically meaningless ranges, we wish to emphasize that, in actual research, the researcher bears responsibility to ensure that interpretation and presentation of the results corresponds with logically relevant and substantively meaningful values of the independent variables of interest. This implies that researchers must give (readers must demand, and editors must publish) such information about sample, substantive, and logical ranges necessary for the reader to recognize substantively and logically meaningful and sample-covering ranges. We have projected *Groups* into negative and positive values for pedagogical purposes only, to display properties of the marginal effects and confidence intervals, but we reiterate that these would *not* be logically relevant values in this case. Indeed, presenting a graph like Figure 2, which extends well beyond the sample and indeed the logically permissible range, would foster misleading conclusions regarding the substantive meaning of our estimates.

[FIGURE 2 ABOUT HERE]

2. Presentation of Predicted Values.

As discussed above, another effective method of interpreting coefficients and translating them into effects is by examining differences in predicted values across some values of a specific variable of interest. Tables and graphs of predicted values along with their associated confidence intervals can be generated according to the steps delineated below. Suppose we are interested in the prediction of y as x varies across a range of values, say from value a, its sample minimum, to c, its maximum, while holding z constant at some (meaningful and revealing) value. Changes in these predictions as x varies would reveal the effects of such changes in x on y at that level of z. We can easily create such tables or graphs of predicted values as x varies, holding z fixed. (*N.b.*, xz will also vary with x, even though its other component, z, is held constant.)

Changes in predictions, however, are only part of the story. Including measures of uncertainty around predictions is also imperative. As before, each predicted value has its own level of uncertainty attached to it. Thus, tables and graphs of predicted values should also include standard errors or confidence intervals (or variances, standard errors, or significance levels) around each of those predicted values.

The variance around each predicted value can be calculated as follows:

$$Var(\hat{y}) = Var(\hat{\beta}_0 + \hat{\beta}_x x + \hat{\beta}_z z + \hat{\beta}_{xz} xz) \qquad [19]$$

Expanding this expression:[29]

$$Var(\hat{y}) = Var(\hat{\beta}_0) + x^2 Var(\hat{\beta}_x) + z^2 Var(\hat{\beta}_z) + (xz)^2 Var(\hat{\beta}_{xz})$$

$$+ 2xCov(\hat{\beta}_0, \hat{\beta}_x) + 2zCov(\hat{\beta}_0, \hat{\beta}_z) + 2xzCov(\hat{\beta}_0, \hat{\beta}_{xz})$$

$$+ 2xzCov(\hat{\beta}_x, \hat{\beta}_z) + 2x(xz)Cov(\hat{\beta}_x, \hat{\beta}_{xz}) + 2z(xz)Cov(\hat{\beta}_z, \hat{\beta}_{xz}) \qquad [20]$$

In words, the variance of a sum equals the sum of the variances plus two times all the covariances; more specifically to this case, the variance of a sum of random variables (here, the coefficient estimates) multiplied by constants (here, independent variables) is equal to the sum of the variances times the associated constants squared plus two times all the covariances times the product of their corresponding constants.[30]

When an OLS regression is estimated, the estimated variance-covariance matrix of the parameter estimates of the β's is usually stored in memory, and the researcher can usually easily call the variance-covariance matrix through a post-estimation additional command.[31]

Let us take our empirical example and calculate the variance of the predicted values corresponding with various values of *Ethnic Groups* and *Runoff*. Recall that the variance-

[29] Note 21 gives the more general, linear-algebraic formula for variances of linear combinations of random variables and constants.

[30] These are variances and confidence intervals for $E(y|x,z=z_0)$ and not forecast or prediction errors, which would include also some uncertainty due to the variance of the regression's error term. See note 21.

[31] Appendix B provides step-by-step STATA commands.

covariance matrix of the coefficient estimates appears in Table 6. When *Ethnic Groups* = 1 and *Runoff* = 0, we calculate $\hat{y}$ as:

$$(\hat{y} \mid Groups = 1, Runoff = 0) = 4.303 - 0.979*1 - 2.491*0 + 2.005*1*0 = 3.324$$

Using Equation [20], substituting the values at which our independent variables are held (*Groups* =1 and *Runoff* =0) yields the following expression:

$$Var(\hat{y} \mid Groups = 1, Runoff = 0) = Var(\hat{\beta}_0) + 1^2 Var(\hat{\beta}_G) + 0^2 Var(\hat{\beta}_R) + (0)(1)^2 Var(\hat{\beta}_{GR})$$

$$+ 2*1*Cov(\hat{\beta}_0, \hat{\beta}_G) + 2*0*Cov(\hat{\beta}_0, \hat{\beta}_R) + 2*1*0*Cov(\hat{\beta}_0, \hat{\beta}_{GR})$$

$$+ 2*1*0*Cov(\hat{\beta}_G, \hat{\beta}_R) + 2*1*(1*0)Cov(\hat{\beta}_G, \hat{\beta}_{GR}) + 2*0*(1*0)Cov(\hat{\beta}_R, \hat{\beta}_{GR})$$

Substituting the estimated values of the variances and covariances of the coefficients:

$$Var(\hat{y} \mid Groups = 1, Runoff = 0) = 1.509 + 0.593 + 2*-0.900 = 0.302$$

Table 11 presents the variance of the predicted values, when *Runoff* takes the values of 0 and 1, and when *Ethnic Groups* ranges from 1 to 3.

[TABLE 11 ABOUT HERE]

One can easily see that these calculations could become quite cumbersome, quite quickly, in the presence of additional covariates. In fact, calculation of the variance of predicted values would require attention to the variance of each estimated coefficient and the covariances between each of the estimated coefficients (plus the levels of all the interacting variables). In our simple model, which includes just three variables plus an intercept, this implies attention to ten terms. Including just one additional regressor (which did not interact with any others) would require us to include five more terms into Equation [19]!

One way to simplify the expression is to use matrix algebra to depict $\hat{y}$ and to calculate Var($\hat{y}$) (see note 21). Note that a predicted value, $\hat{y}$, is generated by summing the products between set values of the right-hand-side variables and their corresponding coefficients. Let $\mathbf{M_h}$ consist of a *j* by *k* matrix of values at which *x, z*, and any other variables of interest in the equation are set, where *j* refers to the number of evenly spaced values at which the predicted value is calculated and *k* refers to the number of regressors, including the constant. Suppose we were to hold *z* (and any the other variables) at some logically relevant value(s), say z_0, and examine the changes in the predicted values of $\hat{y}$ at various values of *x* as *x* takes on *j* evenly-spaced values between x_a and x_c, and correspondingly, as *xz* takes on *j* evenly-spaced values between $x_a z_0$ and $x_c z_0$.

In our standard equation, we have estimated coefficients for *x*, *z*, and *xz*, in addition to an intercept. $\mathbf{M_h}$ would thus look like:

$$\mathrm{M_h} = \begin{bmatrix} x_a & z_0 & x_a z_0 & 1 \\ x_{a+1} & z_0 & x_{a+1} z_0 & 1 \\ & \vdots & & 1 \\ x_c & z_0 & x_c z_0 & 1 \end{bmatrix}$$

In $\mathbf{M_h}$ we can see that the value of *x* is varying between some value x_a and some other value x_c. The variable *z* is held to z_0. The interaction term *xz* varies, because *x* is varying. The column of one's represents the intercept that is estimated.

We can then express the vector of predicted values $\hat{y}$ that are produced by this particular set of values as:

$$\hat{y} = = \mathbf{M_h} \hat{\beta}$$

$$\text{Where } \hat{\beta} = \begin{bmatrix} \hat{\beta}_x \\ \hat{\beta}_z \\ \hat{\beta}_{xz} \\ \hat{\beta}_0 \end{bmatrix}$$

As a consequence, Var($\hat{y}$)=Var($\mathbf{M_h}\hat{\beta}$).

Since $\mathbf{M_h}$ represents the matrix of values at which our variables are set, and since OLS specifies that our independent variables are fixed in repeated sampling (and fixed for prediction purposes), the matrix $\mathbf{M_h}$ can be extracted from the expression and treated as a constant:

$$\text{Var}(\hat{y})=\mathbf{M_h}\text{Var}(\hat{\beta})\mathbf{M_h}'$$

Where Var($\hat{\beta}$) refers to the variance-covariance matrix of the estimated coefficients.

The diagonal elements in the resulting matrix will correspond with the variance of the predicted values of $\hat{y}$, along various values included in $\mathbf{M_h}$. In our empirical analysis, software programs will produce an estimate of Var($\hat{\beta}$), which we will denote as $\hat{V}(\hat{\beta})$.

Using our empirical example, we can thus calculate the variance of the predicted values of *y* as follows:

$$\mathbf{M_h}\text{ (varying values of } \textit{Groups} \text{ when } \textit{Runoff} = 0\text{)} = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 1.5 & 0 & 0 & 1 \\ 2 & 0 & 0 & 1 \\ 2.5 & 0 & 0 & 1 \\ 3 & 0 & 0 & 1 \end{bmatrix}$$

$$\text{Var}(\hat{y}\,|\text{Runoff=0})=\mathbf{M_h}\hat{V}(\hat{\beta})\,\mathbf{M_h}'$$

$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ 1.5 & 0 & 0 & 1 \\ 2 & 0 & 0 & 1 \\ 2.5 & 0 & 0 & 1 \\ 3 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.593 & 0.900 & -0.593 & -0.900 \\ 0.900 & 2.435 & -1.377 & -1.509 \\ -0.593 & -1.377 & 0.885 & 0.900 \\ -0.900 & -1.509 & 0.900 & 1.509 \end{bmatrix} \begin{bmatrix} 1 & 1.5 & 2 & 2.5 & 3 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Which produces the following symmetric matrix:

$$= \begin{bmatrix} 0.302 & & & & \\ 0.149 & 0.143 & & & \\ -0.005 & 0.138 & 0.281 & & \\ -0.159 & 0.133 & 0.424 & 0.715 & \\ -0.312 & 0.128 & 0.567 & 1.007 & 1.446 \end{bmatrix}$$

The diagonal elements represent Var($\hat{y}$) for respective values of *Groups* when *Runoff* =0.

Statistical computing software or even a basic spreadsheet program can make these matrix calculations very easy and quick to generate.

Predicted values are more effectively displayed when graphed with confidence intervals. As with the marginal effects discussed above, confidence intervals around predicted values $\hat{y}$ can be constructed as follows:

$$\hat{y} \pm t_{df,p}\sqrt{Var(\hat{y})}$$

Where, as before, $t_{df,p}$ is the critical value in a *t*-distribution with *df* degrees of freedom that produces a *p*-value corresponding to ½ of the probability outside of the desired confidence interval. For example, lower and upper bounds of a 95% confidence interval will again come from $t_{df,p}$ of approximately 1.96 in large samples.

In our empirical example, we calculated two sets of predicted values: $\hat{y}$ along evenly-spaced values of *Groups* as to ranges from from 1 to 3, when *Runoff* equals 0 and then when it equals 1. To calculate a confidence interval, we must first calculate the variance of the predicted value and identify a desired level of confidence. Again, given our tiny sample, we will accept appreciable uncertainty and so select a 90% confidence interval, implying a critical value of $t_{12,\alpha=.10}$ = 1.782. We would thus calculate the upper bound and lower bound for the confidence intervals as follows:

Upper bound: $\hat{y} + 1.782 * \sqrt{Var(\hat{y})}$

Lower bound: $\hat{y} - 1.782 * \sqrt{Var(\hat{y})}$

For example, $Var(\hat{y} \mid Groups = 1, Runoff = 0) = 1.509 + 0.593 + 2 * -0.900 = 0.302$.

The corresponding upper bound would be calculated as:

Upper bound: $3.324 + 1.782 * 0.302 = 4.304$

Lower bound: $3.324 - 1.782 * 0.302 = 2.345$

[TABLE 12 HERE]

Table 12 displays the confidence intervals calculated for the predicted values of number of presidential candidates, when *Runoff* takes on values of 0 and 1, and when *Groups* ranges from 1 to 3. These confidence intervals and predicted values could also be graphed. Examples appear in Figures 3 - 6.

[FIGURES 3 and 4 ABOUT HERE]

Figures 3 and 4 show the predicted values of number of presidential candidates, when *Groups* appears on the x-axis, the predicted values appear on the y-axis, and the value of *Runoff* is fixed (at 0 in Figure 3 and at 1 in Figure 4). To showcase the effect of *Runoff*, the researcher might overlay these graphs; the overlay of the marginal effects appears in Figure 5. Confidence intervals are added to this graph in Figure 6.

[FIGURES 5 and 6 ABOUT HERE]

As with the graphs of marginal effects, we extend the range of *Groups* purely for illustrative purposes in Figure 7. As with the graph of marginal effects, the graph of predicted values contains a straight line, indicating how the predicted value changes as *Groups* varies. The hourglass curves indicate the degree of certainty associated with each predicted value, $\hat{y}$. We can see that we have the greatest certainty around the mean of *Groups*, and less certainty at more-extreme and, especially, out-of-sample values of *Groups*, as explained above.

[FIGURE 7 ABOUT HERE]

To examine the effect of a *Runoff* using the difference method, we could present a table of the predicted value of $\hat{y}$, in the presence and absence of a *Runoff*, for meaningful values of *Ethnic Groups*. One could generate a graph, but there would only be two points on the X-Axis (Absence and Presence of a Runoff). Hence, a table may be just as informative (although one might still prefer a box-and-whiskers plot of the two estimates and their associated confidence intervals or even a pair of normal curves with means at the estimated effects and standard deviations given by the standard errors). Table 12 provides these values.

The table reveals substantial overlap in the predicted number of presidential candidates in the presence and absence of a runoff system when only one ethnic group exists. However, at higher numbers of *Groups*, much less overlap occurs in the 90% confidence intervals around the predicted number of presidential candidates. This pattern suggests that as *Groups* increases, the impact of *Runoff* on the predicted value of *y* becomes more discernible statistically.

3. Distinguishing the Two Methods.

What is the difference between graphs of marginal effects and graphs of predicted values of *y*? Both answer the question of what effect *x* has on *y*, but in slightly different ways. The graph of marginal effects shows how the <u>effect</u> of *x* changes as *z* changes, allowing direct comparison of the conditional effects in a single plot. The graph of predicted values shows how the <u>level</u> of $\hat{y}$, the prediction for *y*, changes as *x* changes, at particular levels of *z*. By plotting and comparing several of these prediction lines, one can grasp the <u>effects</u> of *x* or of *z* and how they change as the other variable changes. Selection of one type of plot over the other is therefore largely dependent on the researcher's presentational goals. Either method can effectively convey the substantive results from models involving interactive terms; we stress, though, that either sort of graph should incorporate measures of uncertainty into its presentation.

IV. THE MEANING, USE, AND ABUSE OF SOME COMMON GENERAL-PRACTICE RULES

Having discussed the formulation of interactive hypotheses, and the interpretation and presentation of effects, we turn now to some general-practice rules or advice given elsewhere and often followed in the discipline whose meaning and import seems sometimes less than fully understood.

A. Colinearity and Mean-Centering the Components of Interactions Terms

Having constructed the model and discussed its interpretation above, we turn here more directly to estimation of its parameters. One common concern regarding the estimation of interactive models is the (multi)colinearity, or high correlation, among independent variables usually induced by multiplying two (or more) of them together to obtain another (or others). Colinearity, as political scientists well-know, induces large standard-error estimates, reflecting our low degree of confidence in the partial coefficients estimated on these highly correlated factors. What is sometimes forgotten is that these large standard-errors are correctly large; the effect of some *x* controlling for other terms (*i.e.*, holding them constant) *is* very hard to determine with much certainty if *x* and those others correlate highly. These large standard errors accurately reflect our large uncertainty in these conditions. These perhaps unfortunate, but very real, facts regarding colinearity led Althauser (1971), *e.g.*, to argue against the use of interactive terms at all. However, to omit interactions simply because including them invites a greater degree of uncertainty in parameter estimates is to mis-specify intentionally our theoretical propositions, which assures at least inefficiency and probably bias (depending on standard omitted-variable-bias considerations: namely, if the omitted factor, *xz*, (partially) correlates with the included, say *x*, and (partially) correlates with the dependent variable, *y*, then a bias of sign and magnitude given by the product of these two partial coefficients is incurred).

Scholars, therefore, struggled valiantly for some technical artifice to reduce interaction-induced colinearity. However, the only "cures" for colinearity—which is just fancy jargon for too-little information—remains either to ask the data only questions easier to answer (*e.g.*, only first-order questions, like Table 3 or 5), which is often substantively unsatisfying, or to obtain more information, whether drawn from new data (preferably less-correlated, but more data will help regardless) or from heavier reliance upon the theoretical arguments/assumptions to specify models that ask more precise questions of the data than do generic linear-interactive models (*e.g.*, Franzese 1999, 2003a) .

Scholars devoted inordinate attention to illusory colinearity cures, the most commonly prescribed and taken of which in the interaction-term context, then and, unfortunately, now, was to "center" the variables *x* and *z* (i.e., subtract their sample means or "mean-deviate" them) that comprise the interactive term. Smith and Sasaki (1979) offered centering as a technique that would improve substantive interpretation of the individual coefficients, which it might in some substantive contexts. Tate argued that, although centering should not change the substantive effects (actually, it will not, see below), it "may improve conditioning through reduction of colinearity" (253), which 'conditioning' may have some relevance computationally given that computers use binary rather than base-ten yielding some rounding error (but see below). Morris, Sherman and Mansfield (1986), Dunlap and Kemery (1987), and others recommend centering less circumspectly. Cronbach's (1987) centering technique, apparently, has attained considerable acceptance in political science, perhaps due to its advocation in Jaccard, Turrisi and Wan's

(1990) *Interaction Effects in Multiple Regression*. Unfortunately, Cronbach's clarification on the extremely limited value of the technique seems less widely known.

Cronbach's (1987) centering procedure is still used in the discipline. It is harmless, but also is no help, if understood correctly; our concern is that it seems widely misunderstood and misinterpreted. Some existing scholarly research using the centering technique claim, wrongly, that centering helps evade colinearity in some manner that actually produces more-certain effect-estimates. Centering adds no new information of any sort to the empirical estimation, so it cannot possibly have produced more-precise estimates. It does not; it merely changes the substantive question to which the commonly reported coefficients and *t*-tests refer. Translating that answer back to the question answered by the commonly reported coefficients and *t*-tests in the non-centered equation gives the same answer. We explain these points now.

Recall our "standard" linear-interactive model:

$$y = \beta_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon \qquad [21]$$

Cronbach (1987) suggested subtracting the sample means from each of the independent variables involved in the interaction and multiplying the resulting demeaned variables together. The mean-centered model, then (using γ to represent coefficient values resulting from use of the centered data), is as follows:

$$y = \gamma_{0*} + \gamma_{x*} x* + \gamma_{z*} z* + \gamma_{x*z*} (x*)(z*) + \varepsilon * \qquad [22]$$

Where $x* = x - \bar{x}$ and $z* = z - \bar{z}$

Cronbach (1987) argued that rescaling variables could insure against computational errors—*i.e.*, literally computational: deriving from inescapable rounding errors in translating between computer binary and human base-10—that severe colinearity might induce.[32] Cronbach (1987) also noted that centered and non-centered models "are logically interchangeable, and under most circumstances it makes no difference which is used" (415). Given the many thousands of times computing precision has increased since Cronbach's writing, the computational concern has no current practical relevance in social science.

Notwithstanding the considerable care in discussing centering that most advocating it adopted, many political scientists seem to have adopted the practice credulously. Because centering does not affect the substance of any empirical estimation in any way, because it will not affect the computational algorithms of any modern statistical software, and because it is so widely misunderstood in the field, we join Friedrich (1982), Southwood (1978) and others in strongly advising the field to abandon the practice or, at least, to take great care in interpreting and presenting the results following it implementation. (In some contexts, as clarified below, mean-centering could actually enhance interpretation, as Smith and Sasaki (1979) suggested, provided the reader is fully reminded of its implications in that context, we would add.)

To clarify what centering does to the numeric and substantive estimates of an interactive analysis, which is something and nothing, respectively, consider again our basic linear-

[32] The computational issue here involves matrix inversion, namely the $(X'X)^{-1}$ in OLS formulae for coefficient and standard-error estimates, some of whose columns (*i.e.*, independent variables) may correlate nearly perfectly. If columns of X correlate perfectly, the determinant of $(X'X)$, which appears in the denominator of the formula for $(X'X)^{-1}$, is zero. Division by zero is, of course, impossible; therefore, obtaining distinct coefficient estimates (and so standard errors) when (some) columns of X correlate perfectly is impossible. All modern regression software warns of perfect colinearity when it obtains a zero determinant before allowing the computer to crash trying to divide by zero. Most warn of near-perfect colinearity well short of obtaining identically zero for that critical determinant, *i.e.*, well short of perfect colinearity, because, the translation from your base-10 data-matrix to the binary of computers involves rounding error. When something near zero appears in a denominator and contains slight rounding error, the final answer could exhibit massive error. This is the concern Cronbach raises. The multiplicative terms in interactive regressions, he feared, could be near enough to perfect collinearity to cause severe binary-to-base-ten rounding-error problems. However, since his writing, computers have become many thousands of times more exact in their binary calculations' approximation to base-10, meaning even this computational concern is no longer present in any practical social-science context.

interaction model and our centered model, which appear in equations [21] and [22], respectively. Starting from equation [22], and substituting terms, we see that:

$$y = \gamma_{0*} + \gamma_{x*}(x - \bar{x}) + \gamma_{z*}(z - \bar{z}) + \gamma_{x*z*}(x - \bar{x})(z - \bar{z}) + \varepsilon * \qquad [23]$$

$$y = \gamma_{0*} + \gamma_{x*}x - \gamma_{x*}\bar{x} + \gamma_{z*}z - \gamma_{z*}\bar{z} + \gamma_{x*z*}xz - \gamma_{x*z*}\bar{x}z - \gamma_{x*z*}x\bar{z} + \gamma_{x*z*}\overline{xz} + \varepsilon *$$

$$y = (\gamma_{0*} - \gamma_{x*}\bar{x} - \gamma_{z*}\bar{z} + \gamma_{x*z*}\overline{xz}) + (\gamma_{x*} - \gamma_{x*z*}\bar{z})x + (\gamma_{z*} - \gamma_{x*z*}\bar{x})z + \gamma_{x*z*}xz + \varepsilon * \qquad [24]$$

Comparing the centered equation in [24] with the original model in [21] highlights the exact correspondence of results between centered and uncentered regressions.

$$\beta_0 = \gamma_{0*} - \gamma_{x*}\bar{x} - \gamma_{z*}\bar{z} + \gamma_{x*z*}\overline{xz}$$

$$\beta_x = \gamma_{x*} - \gamma_{x*z*}\bar{z}$$

$$\beta_z = \gamma_{z*} - \gamma_{x*z*}\bar{x}$$

$$\beta_{xz} = \gamma_{x*z*}$$

Collecting terms, we see that the first set of terms in equation [24] serves as the intercept term. We see that the second set of terms in equation [24] serves as a coefficient for *x*. The third set of terms in equation [24] serves as a coefficient for *z*. The fourth term is the same in both versions of the model. Furthermore, trivially, since the right-hand-side models are mathematically interchangeable, the residuals and so the residual-variance from the centered and uncentered models are also identical.

As we discussed above, researchers' common troubles arise when they confuse coefficients with effects. We know, for example, that the marginal effect of *x* on *y* in equation [21] would be:

$$dy / dx = \beta_x + \beta_{xz} z$$

The marginal effect of $x*$ on *y* given equation [22] would be

$$dy / dx^* = \gamma_{x^*} + \gamma_{x^*z^*} z^*$$

With a little algebra, we can see that if $\beta_x = \gamma_{x^*} - \gamma_{x^*z^*}\bar{z}$, then $\gamma_{x^*} = \beta_x + \gamma_{x^*z^*}\bar{z}$. Therefore:

$$dy / dx^* = \beta_x + \gamma_{x^*z^*}\bar{z} + \gamma_{x^*z^*} z^*$$

Given that $z^* = z - \bar{z}$:

$$dy / dx^* = \beta_x + \gamma_{x^*z^*}\bar{z} + \gamma_{x^*z^*} z - \gamma_{x^*z^*}\bar{z} = \beta_x + \gamma_{x^*z^*} z$$

Since $\beta_{xz} = \gamma_{x^*z^*}$:

$$dy / dx^* = \beta_x + \beta_{xz} z$$

$$dy / dx^* = dy / dx$$

The marginal effect of a centered version of *x* is identical to the marginal effect of the uncentered version of *x*. The same identity, of course, applies to the effects of *z*. We reiterate: centering does <u>not</u> change the estimated effects of the variables of interest.

And the estimated variance-covariance (*i.e.*, standard errors, *etc.*) of those effects are also identical. Thus, there is no change in the estimated statistical certainty of the estimated <u>effects</u> either. For the uncentered data, we know that

Given the results above, $Var(d\hat{y}/dx) = Var(\hat{\beta}_x) + z^2 Var(\hat{\beta}_{xz}) + 2zCov(\hat{\beta}_x, \hat{\beta}_{xz})$

Using the mean-centered model:

$$Var(d\hat{y}/dx^*) = Var(\hat{\gamma}_{x^*}) + (z^*)^2 Var(\hat{\gamma}_{x^*z^*}) + 2(z^*)Cov(\hat{\gamma}_{x^*}, \hat{\gamma}_{x^*z^*})$$

Substituting $\hat{\gamma}_{x^*} = \hat{\beta}_x + \hat{\gamma}_{x^*z^*}\bar{z}$ and $\hat{\beta}_{xz} = \hat{\gamma}_{x^*z^*}$

$$Var(d\hat{y}/dx^*) = Var(\hat{\beta}_x + \hat{\beta}_{xz}\bar{z}) + (z^*)^2 Var(\hat{\beta}_{xz}) + 2(z^*)Cov(\hat{\beta}_x + \hat{\beta}_{xz}\bar{z}, \hat{\beta}_{xz})$$

$$Var(d\hat{y}/dx^*) = Var(\hat{\beta}_x) + \bar{z}^2 Var(\hat{\beta}_{xz}) + 2\bar{z}Cov(\hat{\beta}_x, \hat{\beta}_{xz})$$

$$+(z^*)^2 Var(\hat{\beta}_{xz}) + 2(z^*)Cov(\hat{\beta}_x + \hat{\beta}_{xz}\bar{z}, \hat{\beta}_{xz})$$

Re-arranging terms and substituting $z^* = z - \bar{z}$:

$$Var(d\hat{y}/dx^*) = Var(\hat{\beta}_x) + \bar{z}^2 Var(\hat{\beta}_{xz}) + (z-\bar{z})^2 Var(\hat{\beta}_{xz}) + 2(z-\bar{z})[\bar{z}Var(\hat{\beta}_{xz})]$$

$$+2(z-\bar{z})[Cov(\hat{\beta}_x \hat{\beta}_{xz})] + 2\bar{z}Cov(\hat{\beta}_x, \hat{\beta}_{xz})$$

$$Var(d\hat{y}/dx^*) = Var(\hat{\beta}_x) + z^2 Var(\hat{\beta}_{xz}) + 2zCov(\hat{\beta}_x, \hat{\beta}_{xz})$$

The variance of the estimated marginal effect of the centered *x* is identical to the variance of the estimated marginal effect of the uncentered *x*.

Although the exact numeric values of the elements in the variance-covariance matrices for the coefficients using uncentered and centered data will naturally differ from each other, exact correspondence in the estimated parameter variances can be derived through algebraic manipulation of the values in the variance-covariance matrices. As an example, recall that $\beta_x = \gamma_{x^*} - \gamma_{x^*z^*}\bar{z}$. This implies that

$$Var(\hat{\beta}_x) = Var(\hat{\gamma}_{x^*} - \hat{\gamma}_{x^*z^*}\bar{z}) = Var(\hat{\gamma}_{x^*}) + \bar{z}^2 Var(\hat{\gamma}_{x^*z^*}) - 2\bar{z}Cov(\hat{\gamma}_{x^*}, \hat{\gamma}_{x^*z^*})$$

Hence, while the estimated variance-covariance matrices will differ numerically from each other (*i.e., $Var(\hat{\beta}_x) \neq Var(\hat{\gamma}_{x^*})$*), the precision in the effects of the variables will be identical, regardless of whether the data are estimated with centered or uncentered data. Again, we warn the reader against confusing coefficients with effects.

We have just shown that estimates of the substantive effects and all estimates of the certainty of those substantive effects are identical whether the data is mean-deviated or left uncentered. Why, then, one might wonder, do some coefficient estimates, their standard errors, and corresponding *t*-statistics differ? The answer is simply that the coefficients and associated standard errors and *t*-statistics do not refer to the same substantive effect across centered and uncentered models. For example, in our standard model, $y = \beta_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon$, the coefficient β_x gives the effect of a unit increase in *x* when *z* equals zero; its standard error and the resulting *t*-ratio refer to the certainty of that *x* effect at that particular *z* value. In $y = \gamma_{0*} + \gamma_{x*} x^* + \gamma_{z*} z^* + \gamma_{x*z*}(x^*)(z^*) + \varepsilon^*$, the coefficient γ_{x*} gives the effect of a unit increase in *x** (or *x*, since a unit increase in *x* or *x** is the same thing) when *z** equals zero, which is not at all the same value of *z* as when *z*=0. Recall that $z^* = z - \bar{z}$, which means that *z** equals zero when *z* (the uncentered variable) equals its mean, not when *z* is equal to zero (assuming, of course that $\bar{z} \neq 0$). The standard error of this coefficient-estimate γ_{x*} and the resulting *t*-ratio refer to the certainty of the effect of a one-unit change in *x* at this <u>different</u> *z* value. Coefficients, standard errors, and t-statistics differ in centered from non-centered model because they refer to different substantive quantities, not because either model produces different, much less any better, estimates of effects than does the other.

Centering can, in this manner, actually be useful for substantive interpretation in some contexts. If interpreted carefully and understood fully, centering sometimes can facilitate a more substantively grounded discussion of the empirical analysis. If *z* cannot logically equal zero, *e.g.*, then substantive interpretation of β_x is vacuous, but examining the effect of *x* when *z* is equal to its sample mean might be substantively revealing. If so, researchers might advantageously center *z* around its mean to aid substantive interpretation and discussion. That is, centering *z* around its mean allows one to interpret the coefficient on *x*, as the effect of *x* when *z* equals its mean rather than when *z* equals zero and the former may be a more meaningful and revealing value in some contexts than the latter. Further, it allows the researcher to interpret the *t*-statistic on γ_{x*} as the statistical significance of that effect when *z* equals its mean, which may likewise simplify discussion in some contexts.

Accordingly, our concern is neither that centering, properly understood, does something damaging because it does nothing, really, nor that centering cannot sometimes facilitate discussion because it can. Our concern is that researchers too often misinterpret the results of centering – and have come to the mistaken conclusion that centering alters the estimates of *effects* or the estimated significance of *effects*. We recommend that centering transformations, if applied at all, be applied only with the aim to improve substantive presentation, not, *mistakenly*, to improve (apparent) statistical precision, and certainly not, *reprehensibly*, to yield more asterisks of statistical significance on reported *t*-tests simply by algebraically transforming variables to have the standard *coefficient* tests refer to the *effects* at most advantageous values of other variables. To understand this last comment, refer back to the characteristic hourglass shapes of confidence intervals around interactive effects. By subtracting the appropriate value from the variable on the x-axis before interacting, the researcher can shift the *coefficient* on the other variable to refer to the *effect* of this other variable at any value of the x-axis variable desired. This changes nothing at all substantively, but the t-statistic and significance of the *coefficient* on the other variable can in this manner be optimized over this hourglass relative to the x-axis. The *substantive* interpretation of the *effects* and the certainty of those effects is completely unaffected by this statistical sleight-of-hand. Therefore, as we have said, reviewers and other readers must demand effect-line plots or tabulations along with appropriate certainty estimates over meaningful ranges of the interacting variables and must not be confused into thinking of the coefficients on the component variables in the interactions as “main” effects and so giving their *t*-statistics inappropriate meaning.

B. INCLUDING *X* AND *Z* WHEN *XZ* APPEARS

To estimate models containing multiplicative interaction terms, most texts advise and political scientists have usually followed a hierarchical testing procedure: *i.e.*, if *xz* enters the model, then *x* and z must also. If *wxz* appears, then all (six) of the lower-order combinations (*x, w, z, xw, xz, wz*) must appear also, and so on analogously for higher-order interactions. Allison (1979), *e.g.*, writes, "[The] common rule… is that testing for interaction in multiple regression should only be done hierarchically... If a rationale for this rule is given at all, it is usually that additive relationships somehow have priority over multiplicative relationships" (149-150).

While this rule is probably a highly advisable rule, if one must have a rule, and certainly a much safer rule than an alternative proviso that one can include or not include components to interactions with little concern or consideration, we believe researchers must understand the logical foundations of the models they estimate and the meaning and purpose of any proffered rule, no matter how usually useful, instead of merely following such rules unthinkingly by rote. We argue instead for theoretically-driven statistical specifications with better appreciation of the assumptions underlying alternative models. In this context, that suggests that the rule to include *x* and *z* if including *xz*, while usually a quite reasonable application of *Occam's Razor*, frequently practically advisable, and almost always a good thing to check rather than assume, is neither logically nor statistically strictly necessary.

In short, including both *x* and *z* if *xz* appears is probably less harmful than the sin of omission, but inclusion of *x* or *z* is, first of all, neither logically nor statistically *necessary*. As proof that the rule cannot be logical necessity, notice that one can decompose *any* variable into the product of two or more others; therefore, strict adherence to this rule would actually entail infinite regress. As a substantive example, note that real GDP (*per capita*) equals nominal GDP times a price-index deflator (times the population inverse); conversely, nominal GDP (*per capita*) is real GDP times a price index (times the population inverse). Nothing statistically or logically requires researchers to include all these components in every model containing some subset of them. Researchers should, instead, estimate the models their theories suggest.

What then is the reason for this so-often useful rule? We see several, related, very good reasons. First, given the state of social-science theory, the models implied by theory will often be insufficiently specified as to whether to include *x* and/or *z* in an interactive model. Due scientific caution would then suggest including *x* and *z* to allow the simpler linear-additive theory a chance. Failing to do so would tend to yield falsely significant estimates of coefficients on *xz* if, in fact, *x* or *z* or both had just linear-additive effect on *y*. Second, inclusion of the *x* and *z* terms in models involving *xz* allows a non-zero intercept to the conditional effect lines, such as those plotted above. This is important because, even if the effect of *x* or of *z* is truly zero when the other variable is zero, if this conditional relationship is nonlinear, allowing a non-zero intercept to the linear-interactive estimate of the truly nonlinear interaction (by including *x* and *z*) will tend to enhance the accuracy of the linear approximation. Third, and perhaps most importantly, even when the theory clearly excludes *x* and/or *z*, from the model, *i.e.*, when it unequivocally establishes the effect of one (or both) variable(s) to be zero when the other is zero, the researcher can and should test that assumption and report the certainty with which the data supports the exclusion. If that test supports exclusion, then both theory and evidence recommend exclusion of these components, so their inclusion is misspecification, not, as the blind adherence to the rule would have it, their exclusion. For this sort of empirical exploration, only finding a coefficient expected to be zero in fact to be estimated as (very close to) zero and, highly preferably, with small standard error is clear evidence from the data that the assumption holds. That is, clear support for the assumption comes from failure to reject because the estimate is with some certainty near zero, rather than the other two reasons one might fail to reject: near zero but large standard error, which is more ambiguous, and far from zero and large standard error, which is probably not support. In sum, then, this rule, as an application of *Occam's Razor*, is a safer adage than its absence, but researchers should still, first, understand the basis for the rule and, second, should not shy from breaking it if their theory and the data strongly suggest doing so.

We now elaborate these points more fully and formally. If the theory expressly excludes *z* from having any effect on *y* when *x* is zero—*i.e.*, non-zero presence of *x* is a necessary condition for *z* to affect *y*, the correct model is:

$$y = \beta_0 + \beta_x x + \beta_{xz} xz + \varepsilon \qquad [25]$$

By this model, as theory demands, the effect of *z* on *y*, (*dy/dz*) equals $\beta_{xz} x$, which is zero when *x=0*. To estimate this model is to hold the necessity clause that *x* must be present for *z* to affect *y* true by assumption, not allowing the data to adjudicate the question. If, however, *z* does affect *y* even when *x* is zero, [25] would suffer omitted-variable bias; it should have included *z*. When OLS models omit such relevant factors, coefficient estimates will wrongly attribute the omitted effects to whatever does enter the model that correlates with the omissions. In this case, that will most likely imply a biased β_{xz} estimate (primarily). OLS will attribute some of the true-but-omitted effect-of-*z*-when-*x=0* to *z*'s interaction with *x*, so the estimate of β_{xz} will be too large (small) when this true-but-omitted effect is positive (negative). Thus, if the omitted effect is, *e.g.*, positive, not only will the estimated effect of *z* on *y* (*i.e.*, *dy/dz*=$\beta_{xz} x$) reflect a greater conditional effect than truly exists (*i.e.*, greater slope to this effect line), and not only does this imply underestimation of the effect of *z* on *y* at low *x* and over-estimation at high *x*, (*i.e.*, too-low effect-line intercept: zero in fact), but, conversely, the effect of *x* on *y* (*i.e.*, *dy/dx*=$\beta_x + \beta_{xz} z$) will be estimated as more conditional upon *z* than it truly is, implying too great a slope to this effect line also and, likely, therefore also too low an intercept (β_x) to that effect-line. Rather than assume such necessity clauses by omitting key interaction components, even when the underlying theory unequivocally demands their exclusion, we recommend researchers test them by first estimating the model including all multiplicative term lower-order components:

$$y = \beta_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon \qquad [26]$$

An insignificant coefficient of β_z here might then support the exclusion theory and provide some justification for proceeding with the necessity clause in place. But recall that a *t*-test on β_z only refers to the effect of *z* when *x* equals zero. However, here too, the basic caution of the hierarchical if-*xz*-then-*x*-and-*z* rule, seems well-taken. The theory concludes β_z should equal zero, so, even more so than usually, we would hardly want to accept that hypothesis merely because we fail to reject it at some high significance level (90% or $\alpha = 0.10$). Recall that failure to reject can occur with small coefficient-estimates and small standard-errors, small coefficient-estimates and large standard-errors, or large coefficient-estimates and larger standard-errors. Only the first of these cases should give the researcher great comfort that she may estimate the model that assumes the necessity clause by omitting (an) interaction component(s); the second gives less support for such a restriction; and that last gives very little or none at all.

In sum, although following the cautionary, *Occam's Razor*, hierarchical rule is certainly a more sensible starting point than ignoring it, and although empirical tests of exclusionary hypotheses over-riding that rule are not likely to support its abrogation frequently in social science, estimating models like [26] that include interaction components when true models, *e.g.* [25] perhaps, actually exclude them will cost researchers some inefficiency if not bias. That is, estimating [26] when the true model is [25] involves trying to estimate more coefficients than necessary, which implies inflated standard errors. Moreover, these included-but-unnecessary coefficients, β_x or β_z, are on variables, *x* or *z*, that are likely highly correlated with the necessary ones, *xz*, which implies greatly inflated standard errors. Thus, the inefficiency of over-cautious interaction-component inclusion could easily and often be severe enough to lead researchers to miss many interactions actually present in their subject.

Hence, we mildly disagree with rigid rules that component variables must necessarily appear in all models containing interaction terms. Here, as always, theory must guide model specification, and no effective substitute—such as rigid rule-following—exists for understanding the consequences of one's methodological decisions. Specifically on the hierarchical approach, we recommend that the researcher (a) acknowledge and discuss the assumptions/arguments underlying the decision to omit or include components of their interaction terms, (b) gauge statistically the certainty with which the data supports those assumptions, and then (c) apply *Occam's Razor* by following hierarchical procedures unless theory and data clearly indicate that doing so is unnecessary and over-cautious.

V. EXTENSIONS.

Having discussed how to model interactive hypotheses empirically, to evaluate them statistically, and to interpret them substantively, and now having discussed and hopefully clarified the role of some general practice rules, we turn next to some more-technical statistical concerns often raised regarding interaction-term usage in regression analysis. The first issue regards three-way interactions, which we can understand in the context of the framework already developed herein. The second issue, regarding separate *versus* pooled-sample estimation of interactive effects, should interest all readers. Readers less interested in more-technical matters, though, can skim the ensuing sections, being sure to take from them, however, the strong recommendation that one employ some version of White's robust (heteroskedasticity-consistent) variance-covariance (standard-error) estimator when estimating regression models that include multiplicative interaction-terms. The third issue concerns estimation and interpretation of interaction terms in nonlinear models. And the fourth issue concerns modeling and estimating stochastically (rather than determinatively) interactive relationships.

A. Three-Way (and Multiple) Interactions.

On occasion, a researcher might also seek to analyze a more complicated model that includes a three-way interaction. Nearly all of the discussion above applies to this case. Suppose a researcher wanted to estimate the following model:

$$y = \beta_0 + \beta_x x + \beta_z z + \beta_w w + \beta_{xz} xz + \beta_{zw} zw + \beta_{xw} xw + \beta_{xzw} xzw \qquad [27]$$

Interpretation of results follows from the above discussion. We reiterate the perils of confusing <u>coefficients</u> with <u>effects</u>. The coefficient β_x refers to the effect of *x* when *z* and *w* are both zero; the coefficient β_z refers to the effect of *z* when *x* and *w* are both zero; the coefficient β_w refers to the effect of *w* when *x* and *z* are both zero. Interpretation of each coefficient in this fashion can thus be quite cumbersome; we thus encourage either differentiation or calculation of predicted values to identify effects of the variables of interest.[33]

The conditional effects of *x*, *z*, and *w* are, respectively:

$$dy/dx = \beta_x + \beta_{xz} z + \beta_{xw} w + \beta_{xzw} zw \qquad [28]$$

$$dy/dz = \beta_z + \beta_{xz} x + \beta_{zw} w + \beta_{xzw} xw \qquad [29]$$

$$dy/dw = \beta_w + \beta_{xw} x + \beta_{zw} z + \beta_{xzw} xz \qquad [30]$$

[33] Indeed, other non-linearity in variables extensions of linear-additive models, such as squares or other powers of regressors, logs of regressors, etc., are even more difficult to interpret without differentiation. For example, the marginal effect of *x* on *y* in the model $y = \beta_0 + \beta_{x1} x + \beta_{x2} x^2 + \varepsilon$ is $\frac{dy}{dx} = \beta_{x1} + 2\beta_{x2} x$, which we believe would be hard to ascertain without differentiation.

The estimated variance of the estimated marginal effect of *x* on *y* is:

$$Var(d\hat{y}/dx) = Var(\hat{\beta}_x) + z^2 Var(\hat{\beta}_{xz}) + w^2 Var(\hat{\beta}_{xw}) + (zw)^2 Var(\hat{\beta}_{xzw})$$

$$+ 2zCov(\hat{\beta}_x, \hat{\beta}_{xz}) + 2wCov(\hat{\beta}_x, \hat{\beta}_{xw}) + 2zwCov(\hat{\beta}_x, \hat{\beta}_{xzw})$$

$$+ 2zwCov(\hat{\beta}_{xz,}\hat{\beta}_{xw}) + 2z(zw)Cov(\hat{\beta}_{xz,}\hat{\beta}_{xzw}) + 2w(zw)Cov(\hat{\beta}_{xw,}\hat{\beta}_{xzw})$$

Or, in matrix notation:

$$d\hat{y}/dx = \lfloor \hat{\beta}_x \quad \hat{\beta}_{xz} \quad \hat{\beta}_{xw} \quad \hat{\beta}_{xzw} \rfloor [1 \quad z \quad w \quad zw]' \qquad [31]$$

$$Var(d\hat{y}/dx) = [1 \quad z \quad w \quad zw] \, Var \lfloor \hat{\beta}_x \quad \hat{\beta}_{xz} \quad \hat{\beta}_{xw} \quad \hat{\beta}_{xzw} \rfloor [1 \quad z \quad w \quad zw]' \qquad [32]$$

In essence, the variance of the conditional effect is the sum of all the variances of the variables multiplied by their cofactor squared plus twice (or two times) all the covariances times the product of their cofactors.

Predicted values can be calculated following the procedures outlined above. The researcher might calculate a set of predicted values for *y* as *x* varies between x_a and x_b, holding *z* and *w* at logically relevant values. The variance of the predicted values would be facilitated by the matrix multiplication described above, where $\hat{y}=\mathbf{M_h}\hat{\beta}$, and Var($\hat{y}$)= $\mathbf{M_h}$Var($\hat{\beta}$)$\mathbf{M_h}'$.

B. SEPARATE- VERSUS POOLED-SAMPLE ESTIMATION OF INTERACTIVE EFFECTS

Researchers often explore the interactive effects of nominal (binary, categorical, *etc.*) variables by splitting their sample according to these nominal categories and estimating the same model separately in these subsamples.[34] In behavioral research, for example, scholars may explore interactive hypotheses arguing that race, gender, and party identification structure the

[34] 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 below, this will entail inefficiency as the gradations of ordinal or cardinal information are thrown away in the conversion to nominal categorization, although the practice may be justifiable in some cases (we suspect rarely) on other grounds.

impact of other variables by estimating the same model in subsamples separated by race, gender, etc. For example, a researcher might estimate the effect of socioeconomic status, *x*, on political participation, *y*, separately in samples of male and female respondents to explore whether socioeconomic status affects the propensity to vote differently depending on gender. This also occurs in comparative or international politics, such as in studies of exchange-rate regimes in political economy, or country or region in other aspects of comparative and international politics. A researcher 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. 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 recognize 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 (1) 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, at least presentationally, and, (2) more fundamentally, do not facilitate statistical comparison (i.e., standard errors) of the effects of "moderated" (or "moderating") variables—i.e., 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 samples 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 interactive” model. The two approaches (separate-sample versus fully interactive pooled sample) extract almost identical sets of information from the data, but pooled-sample estimation more easily extracts slightly more information, potentially more efficiently, and more easily allows for 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 2004), but we believe 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 relegated to exploratory and sensitivity-and-robustness-consideration stages of analysis and replaced by pooled-sample approaches for final analysis and presentation.

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 her sample into males and females and estimates:

Sample g=Male: $y_m = X\beta_m + u_m$ [33]

Sample g=Female: $y_f = X\beta_f + u_f$ [34]

Let *M* (*F*) be the number of observations in the male (female) sample. Let *j* index the columns of ***X*** (*e.g.*, x_{gj} represents the j^{th} independent variable, for the gender *g* sample; β_{gj} is the coefficient on the j^{th} independent variable, for the gender *g* sample), and let *J* be the number of independent variables. To obtain distinct coefficient estimates by gender, the researcher has several options. She could estimate models [33] and [34] separately, once *per* sub-sample, or she could pool the data into one sample and either reconfigure the ***X*** matrix by manually creating separate $\boldsymbol{X_m}$ and $\boldsymbol{X_f}$ variables for each column of ***X***, where $\boldsymbol{X_m}$ replaces each female respondent’s

$\boldsymbol{X}$-value with zero and $\boldsymbol{X_f}$ does analogously for male respondents. This allows distinct coefficients on $\boldsymbol{X_m}$ and $\boldsymbol{X_f}$ and, if the constant (intercept) is also separated into $\boldsymbol{X_m}$ and $\boldsymbol{X_f}$ 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=*male*, another indicator for g_f=*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 $g_m\boldsymbol{X}$ and $g_f\boldsymbol{X}$ here will equal the $\boldsymbol{X_m}$ and $\boldsymbol{X_f}$ from the manual procedure just described, so this too 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, 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_j (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_j plus the coefficient on the corresponding interaction term, g_fx_j, would refer to the effect on y of that x_j among females. And the coefficient on g_fx_j would refer to the difference from the effect of that x_j among females to the effect of that x_j 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 employing interaction terms exist and, as we show next, are at least equivalent on all grounds except, perhaps, convenience, where in some regards they might dominate and in other regards be dominated. Others may note more explicitly that, lacking *a priori* hypotheses about what differences in the effects of the various x_j 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, but we found no published article that made such an argument.

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

$$\begin{bmatrix} y_{1m} \\ \vdots \\ y_{Mm} \end{bmatrix} = \begin{bmatrix} 1 & X_{m11} & \cdots & X_{mk1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{m1M} & \cdots & X_{mkM} \end{bmatrix} \begin{bmatrix} \beta_{m0} \\ \beta_{m1} \\ \\ \beta_{mk} \end{bmatrix} \qquad [35]$$

and the exactly analogous equation for females. Typically, they 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 sample. That is the often-cited ease of interpretation. However, the second or third of the pooled-sample options described above (*i.e.*, creating distinct X_f and X_m variables manually or by dummy-variable interaction) exactly replicates these coefficients estimates, so if researchers prefer this sort of interpretability (another sort exists: see below), either pooled- or separate-sample estimation can 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, this direct interpretability of effects by subsample cannot adjudicate between pooled and separate-sample approaches since one can present the same results in the same fashion regardless of whether those results derived from pooled or separate-sample estimation. The other sort of interpretability—estimates that yield a direct statistical test of the significance of the difference between subsample estimates of the effects of ***X***—can adjudicate and decidedly favors pooled-sample estimation (see below).

The claimed convenience of separate-sample estimation, for its part, arises (we believe) from the user-interface particularities of many statistical-software packages. In many packages, restricting the sample on which a command (*e.g.*, OLS estimation) executes is easier than creating separate X_f and X_m variables directly or through appropriate interaction application. We do not dispute this convenience. Although such exploratory analysis has its own quite severe perils—namely, any such pre-vetting exploratory analysis strictly speaking invalidates *all* statistical tests of *any* estimates from the final-analysis models—almost all social scientists do conduct such preliminary explorations and have good reasons for doing so. Despite these known and severe perils, the state of most social-science theory probably mandates such exploration; rarely will our first approach to data even remotely approximate the true model we should have estimated. We do not condemn such exploration here. Therefore, given that such preliminary exploration remains perhaps an unavoidably necessary step in positive social-science, and that preliminary exploration should certainly be convenient, we suggest that separate-sample estimation remain common practice in preliminary analyses where convenient.

However, in final analyses, we will certainly want to report the most interpretable and accurate parameter-estimates possible, along with the most accurate and interpretable estimates of those parameters' uncertainty and that of any substantively important comparisons of those estimates. Pooled-sample estimation has, as noted above, at least identical interpretability and, as elaborated next, tends to dominate regarding the ease with which it can flexibly modify model assumptions that could improve the efficiency of coefficient estimates and the accuracy and interpretability of standard errors.

Underlying any separate-sample estimation in the first place is at least the hunch that the effect of some independent variable(s) differ(s) across the categories distinguished by the subsamples. Thus, certainly, anyone conducting such analysis will wish to compare coefficient estimates across such subsamples. However, separate-sample estimation encourages researchers merely to eyeball such comparisons. ***If*** the classical OLS assumptions apply in each subsample (so the OLS $\hat{\beta}$'s are BLUE), then the researcher could test the statistical significance of any estimated differences in parameters estimated separately across subsamples by difference in means tests of each β_f and corresponding β_m:

H$_0$: $\beta_f = \beta_m$, or $\beta_f - \beta_m = 0$

H$_a$: $\beta_f \neq \beta_m$, or $\beta_f - \beta_m \neq 0$

Conducting the standard *t*-test on this null hypothesis:

$$\frac{(\hat{\beta}_f - \hat{\beta}_m) - 0}{se(\hat{\beta}_f - \hat{\beta}_m)} = \frac{(\hat{\beta}_f - \hat{\beta}_m)}{\sqrt{\text{var}(\hat{\beta}_f) + \text{var}(\hat{\beta}_m) - 2\,\text{cov}(\hat{\beta}_f, \hat{\beta}_m)}} = \frac{(\hat{\beta}_f - \hat{\beta}_m)}{\sqrt{\text{var}(\hat{\beta}_f) + \text{var}(\hat{\beta}_m)}} \ \textbf{if}\ \hat{\beta}_f,\ \hat{\beta}_m \text{ independent}$$

We note, first, that few researchers in our review of the literature actually conducted this test; at best, they offered instead some vague reference to the *individual* standard errors of the two coefficient estimates in question. Although the subsample coefficient estimates will be independent by construction—the orthogonality of the indicator-variables assures this—which implies that the covariance in the second term is zero, 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 (since the researcher heretofore typically has not) 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*-test on that interaction-term coefficient directly reveals the statistical significance of that difference. (Likewise, the standard *F*-test on the set of interaction terms tests whether the set of effects of ***X*** jointly differ across subsamples.) Thus, regarding interpretability, pooled-sample estimation offers two ways of presenting the same substantive results, one that replicates the same interpretability of coefficients as effects in subsamples afforded by separate-sample estimation, and another that affords direct interpretation of coefficients as the estimated difference between effects across subsamples and of the standard *t*-tests or F-tests on those coefficients as revealing the statistical significance of that estimated difference.[35]

Formally, either manually or by dummy-variable interactions, one arranges their ***X*** matrix in block diagonal:

[35] Moreover, pooling not only produces identical effect estimates as those obtained from separate samples, but it also (under 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 more precise coefficient estimates (smaller standard errors), although only correctly so if these assumptions are correct (see below).

$$\underset{(M+F)x1}{\begin{bmatrix} y_{m1} \\ \vdots \\ y_{mM} \\ y_{f1} \\ \vdots \\ y_{mF} \end{bmatrix}} = \underset{(M+F)x(2(k+1))}{\begin{bmatrix} 1 & X_{m11} & \cdots & X_{mk1} & 0 & \ldots & & 0 \\ \vdots & & & & & & & \\ & & & \vdots & \vdots & & \ddots & \\ 1 & X_{m1M} & & X_{mkM} & 0 & \ldots & & 0 \\ 0 & \cdots & & 0 & 1 & X_{f11} & \cdots & X_{fk1} \\ & 0 & & \vdots & & & & \\ \vdots & & \ddots & & \vdots & & & \vdots \\ 0 & \ldots & & 0 & 1 & X_{f1F} & & X_{fkF} \end{bmatrix}} \underset{2(k+1)}{\begin{bmatrix} \beta_{m0} \\ \beta_{m1} \\ \vdots \\ \beta_{mk} \\ \beta_{f0} \\ \beta_{f1} \\ \vdots \\ \beta_{fk} \end{bmatrix}} \qquad [36]$$

Recall that $\hat{\Sigma} = s^2 (X'X)^{-1}$. Since the ***X***-matrix here is block-diagonal, the inverse will also be block diagonal, and the elements for males of $(\boldsymbol{X'X})^{-1}$ and ***X'y***, which comprise the coefficient estimate for males $\hat{\boldsymbol{\beta}}_m = (\boldsymbol{X_m'X_m})^{-1}\boldsymbol{X_m'y_m}$ 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 ($\boldsymbol{\beta}_f = \boldsymbol{\beta}_m$). Note, though, that the <u>single</u> s^2 estimated here naturally differs from the <u>two</u>, s_m^2 and s_f^2, estimated separately in the subsample estimates. I.e., 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*-*k* with the *N* reflecting the entire sample (*M*+*F*) and *k* reflecting the 2(k_g+1) coefficients in the pooled estimation. Separate-sample estimation allows s^2 to vary, producing one estimate for each subsample, and each sums only the squared residuals from its subsample and divides only by its subsample *N*. The latter is inefficient. In other words, we obtain better estimates of s^2 and, with it, of estimated coefficient-estimate variance-covariances in pooled- than in separate-sample estimation—<u>if</u>, indeed, the residual-variances <u>are</u> equal across subsamples—by imposing this restriction as in pooled OLS rather than estimating two separate s^2 terms. In this case, the inefficiency manifests as one of the

s_m^2 and s_f^2 being larger than it need be and the other smaller than it should be. More generally, some of the s_i^2 will be larger than need be and others smaller than should be. To explore whether such a common error-variance assumption is warranted, we can test whether heteroskedasticity instead prevails rather simply, and statistical techniques to address that if so are also simple, efficient, and well-known (see below).

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 coefficients and standard-error estimates' efficiency. For example, we may posit, or theory may establish, or even some accounting or mathematical identity may require, that some *x* affect males' and females' voting propensities equally (or equally and oppositely, or otherwise relatedly in some deterministic manner). Rather, then, than estimate both of these effects separately, as separate-sample estimation all-but requires[36], one could in pooled-sample estimation simply refrain from including that or those dummy-variable interactions (and reverse the sign of those variables in the male or female sample, or analogously impose the constraints directly for other restraints). As with a common-variance assumption, such cross-subsample restrictions can be tested, rather than assumed and imposed without testing, and again perhaps more conveniently in pooled than in separate-subsample estimation.

In multi-category cases of the approach of including ***X*** and its interactions with the category indicators, one can include ***X***, 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

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

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. Thus, whether obtaining separate coefficient estimates by category, which one can as easily do in pooled- as in separate-sample estimation, or obtaining one set of coefficients for a base category and another on a set of interactions with the other-category-indicator(s) better facilitates interpretation and presentation depends on what substantive questions the researcher wishes to highlight more directly with the reported coefficients. In either case, as we showed in previous sections, one can interpret these interactive effects by calculating first differences or derivatives (pretending the non-continuous indicators are differentiable).

In summary, the pooled-sample estimation facilitates at least equally and perhaps focuses attention more not just on comparison across subsamples but also on assessing the statistical certainty of such comparisons. Pooled-sample OLS estimation, moreover, imposes homoskedasticity across the subsamples, which, if true, adds some efficiency gains relative to separate-sample estimation.[37] This common-variance assumption can be easily tested and relaxed if unwarranted. Likewise, imposing cross-subsample coefficient restrictions or, for that matter, any other cross-subsample restrictions, is at a minimum far easier to do in pooled than in separate-subsample estimation, and these too will enhance coefficient and standard-error estimates' accuracy and efficiency if true. These gains could be large if subsamples are small and/or numerous, and they will be especially large if some of the coefficients may be constrained (correctly) to be equal across subsamples. If such homoskedasticity and/or equal-coefficient assumptions are false, however, the apparent efficiency gains will also be false, yielding

[37] 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 receive lower and the (some) other(s) higher.

inaccurately smaller standard errors and/or bias and inconsistency. Still, to test for variance or coefficient differences across subsamples is simple. If the data insist that coefficients differ, this is easily allowed, and, if they insist that heteroskedasticity (and/or correlation) among residuals patterned by gender exists, then one can model that variance (or variance-covariance) structure as a function of sex and employ weighted (or feasible-generalized) least-squares as usual in the pooled sample. 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 more than 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 and easily test whether they are (and relax them if they are not). Therefore, we generally recommend 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 return in section V.D below to some complicating considerations for this judgment raised by certain stochastic properties one might reasonably expect in some data structures, finding however that the core of our conclusion remains: in general, anything one can do in separate samples, one can do in pooled estimation with the corresponding interactions; some things will be easier or more convenient or presentationally advantageous in one approach than the other; and, in our view still, the techniques and emphases abetted by the pooled-sample estimation generally dominate.

C. NONLINEAR MODELS.

To this point, we have limited our discussion to interactive terms in linear models. It is fitting, though, to address interaction within nonlinear models also. For nonlinear models that explicitly include interactive terms, much of the discussion in preceding sections continues to apply. However, a complication generally arises when we think about the effect of a variable *x* on a dependent variable *y*, when these variables are nonlinearly related. (A further complication may arise in the contexts discussed in the next section and is discussed there.) When this occurs, as, for example, in logit or probit models, the effect of a variable *x* on *y* already depends on the values of the other variables *z* due to the imposed nonlinear structure of the model. Thus, nonlinear specifications implicitly consist of conditional relationships between the variables, although they may not be explicitly modeled as such. The issue, then, regards the proper interpretation of these variables upon which a conditional relationship has been imposed, or assumed by construction, by virtue of the particular model specification employed.

How much interaction substance can we extract from a model that imposes a conditional relationship upon all of the variables but does not explicitly model a particular interaction between two variables? The danger, as Frant (1991) and Nagler (1991) suggest, occurs when scholars overlook the assumptions embedded in nonlinear specifications such as logit and probit.

These nonlinear functional forms both 1) impose a conditional relationship on the variables by construction and 2) use an S-shaped functional form implying specific character to these interactions, namely that the effects of changes in one variable on *Y* are likely to be larger when the predicted probabilities are closer to the midpoint and smaller away from it. Nagler, for example, critiques Wolfinger and Rosenstone's (1980) claim that the effects of registration requirements in discouraging turnout are greater for low education groups. He argues that this larger effect derives from the functional form assumed *a priori* and do not necessarily result from an explicit or direct interaction between education and registration requirements, e.g., that the less educated find surmounting registration requirements more difficult. The logit form by itself implies only that education interacts with registration requirements and vice versa because and through the other variable's effect on the overall propensity to vote. Insofar as the 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 explores explicitly whether, additionally, education interacts with registration requirements (and vice versa) to help determine where on this S-shaped propensity-to-vote curve one actually lies, which would correspond more directly to Wolfinger and Rosentone's substantive interpretation, by including a specific interaction between education and registration requirement in the model connecting the independent variables to the logit or probit parameters and by disaggregating the sample to estimate these logit or probit coefficients separately for different education levels, with strict and lax registration requirements. He finds little support for Wolfinger and Rosenstone's conclusion.[38]

The notion that multiple, explicit interactions determine one's dependent variable

[38] Similarly, Frant (1991) reviews Berry and Berry's (1990) research 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 a 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. That is, as we suggest in the next paragraph 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, 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.

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. The functional form of the outcome, *i.e.*, that familiar S-shape, implies a particular and very specific set of interactions that produce such S-shapes (sigmoid functions); and, critically in our view, that such S-shapes should actually emerge is substantively and theoretically derived. Namely, it derives from the proposition that inducing probabilities to increase or to decrease becomes increasingly difficult, i.e. requires larger movements in independent variables, as those probabilities near one or zero, respectively (see also note 38). If the researcher wishes to infer beyond the specific forms of interactions that produce such S-shapes, we concur with Nagler that she must model those further interactions explicitly.

That said, we now discuss in more detail the interpretation of effects in two popular nonlinear models: probit and logit. For example, suppose the nonlinear function, often called a "link function", is used to relate a binary outcome *Y*, with **X***B,* where *Y* refers to a column vector of the dependent variable with *n* rows, **X** refers to a matrix of the regressors with dimension *n x k* (*n* observations by *k* regressors, including the intercept), and where *B* refers to a column vector of coefficients (the β's). In such a case, one could model the probability that a particular y_i for observation *i* takes on a value of 1, $p(y_i = 1) = F(X_i\mathrm{B})$: thus, more generally, $p(y = 1) \equiv p = F(X\mathrm{B})$.

In the probit case, $p = \Phi(X\mathrm{B})$, where Φ is the cumulative (standard) normal distribution. The cumulative normal distribution is indeed S-shaped so that ever larger increases or decreases in $X\mathrm{B}$ are required to increase or decrease the probability *y=1* as this probability draws closer to one or zero, respectively. In the logit case, $p = \Lambda(X\mathrm{B})$, where Λ refers to the logit function, which may be written as $\Lambda(X\mathrm{B}) = [1 + \exp(-X\mathrm{B})]^{-1}$. (Several other common formulations of the same function, such as $\Lambda(X\mathrm{B}) = \frac{e^{X\mathrm{B}}}{1+e^{X\mathrm{B}}}$, also exist.)

We begin with the simple equation modeled with the logit link function, omitting explicit interaction terms:

$$p = \Lambda(X\mathrm{B}) = [1+\exp(-(\beta_0 + \beta_x x + \beta_z z + ... + \beta_k w)]^{-1} \qquad [37]$$

As always, the effects of a variable *x* on *p* can be calculated by taking the first derivative of this function: using the chain rule.[39]

$$dp/dx = [dp/dF(XB)][dF(XB)/dx]$$

In the probit case, this is simply:

$$dp/dx = [d\Phi(X\mathrm{B})/dF(XB)][dF(XB)/dx] = \phi(X\mathrm{B})\beta_x$$

where $\phi(X\mathrm{B})$ is the (standard) normal probability density function evaluated at $X\mathrm{B}$.[40] Thus, as central to the theoretical proposition of an S-shaped relationship, the magnitude of effects of *x* on the probability that $y=1$ are largest at $p=0.5$ (at $X\mathrm{B}=0$) and become smaller as that probability approaches one or zero (as $X\mathrm{B}$ approaches infinity or negative infinity).

In the logit case, the marginal effect of a particular *x* on *p* would be:

$$dp/dx = [-(1+\exp(-X\mathrm{B}))^{-2}][\exp(-X\mathrm{B})\beta_x] = \frac{\exp(-X\mathrm{B})\beta_x}{(1+\exp(-X\mathrm{B}))^{-2}}$$

In the specific model of equation [37], this would be:

[39] Note the distinction here between conceptualizing marginal changes as a one-unit change in *x* literally computed (i.e., $\hat{p}_1$ - $\hat{p}_2$) versus an infinitesimal change in *x* (*dy/dx*). Generally, the former is recommended for discrete variables, and the latter for continuous variables. (See Greene 2003 for elaboration.)

[40] The derivative of any cumulative probability distribution, F, is the corresponding probability density function, f, so the derivative of Φ is ϕ.

$$dp/dx = \frac{[\exp\{-(\beta_0 + \beta_x x + \beta_z z + ... + \beta_k w)\}]\beta_x}{[1 + \exp\{-(\beta_0 + \beta_x x + \beta_z z + ... + \beta_k w)\}]^{-2}}$$

Obviously, the effect of *x* depends on the values of *x*, *z*,…, *w* as well as the estimated coefficients for $\beta_0,...,\beta_k$. Luckily, these somewhat complicated expressions simplify to something considerably more intuitive:

$$dp/dx = \frac{[\exp(X\mathrm{B})]}{[1+\exp(X\mathrm{B})]}\frac{1}{[1+\exp(X\mathrm{B})]}\beta_x = p(1-p)\beta_x$$

We can see directly in this formulation that the largest magnitude effects of *x* occur at $p = 0.5$ and these effects become progressively smaller in magnitude as *p* approaches one or zero, producing that familiar S-shape again (although a slightly different S-shape).

Now, when an explicit interaction term (for example, between *x* and *z*) is included in 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 above and now also movements along this S-shape induced by movements in *x* depend directly on the value of *z* as well:

$$p = \Lambda(X\mathrm{B}) = [1 + \exp(-(\beta_0 + \beta_x x + \beta_z z + \beta_{xz} xz + ... + \beta_k w)]^{-1} \qquad [38]$$

As always, the effects of *x* can be calculated using the derivative method or the method of differences in predicted values (in this case, predicted probabilities). A researcher following the first-derivative approach in interpreting a logit model with this explicit interaction in addition to its implicit ones would again apply the chain rule:

$$dp/dx = [-(1+\exp(-X\mathrm{B}))^{-2}][\exp(-X\mathrm{B})(\beta_x + \beta_{xz} z)] = \frac{\exp(-X\mathrm{B})(\beta_x + \beta_{xz} z)}{(1+\exp(-X\mathrm{B}))^{-2}}$$

which again can be expressed conveniently as:

$$dp/dx = [\Lambda(XB)][1-\Lambda(XB)][\beta_x + \beta_{xz} z]$$

This is the same expression as before except that, now, the effect of *x* depends not only on $[\Lambda(XB)][1-\Lambda(XB)]$ as, and in the manner, implied by the sigmoid function connecting *x* to *p*, but also and again on the value of *z* in the manner implied by the linear-interaction of *x* and *z* contained in *X*. Thus, *z* modifies the effect of *x* on *p* not only by its role in the calculation of $[\Lambda(X\mathrm{B})]$, where it enters in the $+\beta_z + \beta_{xz}xz$ terms, but also in the final term, $dX\mathrm{B}/dx$, where it enters in the expression $dX\mathrm{B}/dx = \beta_x + \beta_{xz}z$.

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 just has an additional step in projecting the linear index (i.e., the sum of the coefficients times their covariates) through the nonlinear model (in this case) onto probability space. For example, suppose we estimated the following relationship:

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

The predicted probabilities, generically, are $\hat{F} = F(X\hat{\mathrm{B}})$. The linear index $(X\hat{\mathrm{B}})$ is computed in identical fashion to the OLS case:

$$index = \hat{\beta}_0 + \hat{\beta}_x x + \hat{\beta}_z z + \hat{\beta}_{xz}xz$$

After calculation of the linear index, the researcher must use the link function (here, the logit $\Lambda(X\mathrm{B})$ or probit $\Phi(X\mathrm{B})$) 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_b), 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 $x_c\bar{z}$. The predicted probability in the logit case would thus be calculated first by computing the linear index:

$$index_a = \hat{\beta}_0 + \hat{\beta}_x x_a + \hat{\beta}_z \bar{z} + \hat{\beta}_{xz}x_a\bar{z}$$

And then projecting the linear index into probability space:

$$\hat{p}_a = 1 + \frac{1}{\exp^{-(index_a)}}$$

For probit, the process is the same, substituting $\Phi(index_a)$ for $1 + \frac{1}{\exp^{-(index_a)}}$, i.e. using the cumulative (standard) normal rather than the logit as the link function. By repeating this process with *x* set to a value of x_c, the researcher can thus see how a change in *x* (from x_a to x_c) induces a change in $\hat{p}$ (from $\hat{p}_a$ to $\hat{p}_b$) when *z* is held at this particular meaningful level.

We reiterate our strong recommendations for computation of measures of uncertainty around marginal effects and predicted probabilities. Standard errors for the marginal effects can be computed following the delta method, as described in any statistics text, *e.g.* Greene (2003). For any estimate of marginal effects, *dp/dx*, the asymptotic variance of that estimate is a function of the estimated variance-covariance matrix for $\hat{\mathrm{B}}$ and the derivative (or gradient) of the marginal effects with respect to B:

Asy. Var[*dp/dx*] = [$d(dp/dx)/d\mathrm{B}$]'**V** [$d(dp/dx)/d\mathrm{B}$], where **V** = Asy. Var[$\hat{\mathrm{B}}$].

In the logit case, this yields the (rather long) expression:

Asy. Var[*dp/dx*] = $[\Lambda(t)(1-\Lambda(t))]^2[\mathbf{I}+(1-2\Lambda(t)\beta x']\mathbf{V}[I+(1-2\Lambda(t))x\beta']$

$= [p(1-p)]^2[I-(1-2p)\beta x']\mathbf{V}[I-(1-2p)x\beta]$

Standard errors for the predicted probabilities can be calculated using the delta method also, so, analogously, the asymptotic variance for $\hat{F}$ is a function of the estimated variance-covariance matrix for $\hat{\mathrm{B}}$ and the derivative (or gradient) of $\hat{F}$ with respect to B:

Asy. Var [$\hat{F}$] = [$d\hat{F}/d\hat{\mathrm{B}}$]' **V** [$d\hat{F}/d\hat{\mathrm{B}}$],

where **V** = Asy. Var[$\hat{\beta}$] and $d\hat{F}/d\hat{\mathrm{B}}$ = [$d\hat{F}/dX\hat{\mathrm{B}}$][$dX\hat{\mathrm{B}}/d\hat{\mathrm{B}}$] = $\hat{f}X$

where $\hat{f}$ is the density function (corresponding to *F*) evaluated at the estimated $X\hat{\mathrm{B}}$, and *X* is the vector of values at which all of the independent variables are set for calculating the predicted probability. In the logit case, $\hat{f}$ conveniently equals $[\Lambda(X\hat{\mathrm{B}})][1-\Lambda(X\hat{\mathrm{B}})]$, yielding:

$$\text{Asy. Var}\,[\,\hat{F}\,] = \hat{f}^{2} X'\mathbf{V}\,X$$

For practical purposes, this means squaring the product $\hat{p}(1-\hat{p})$ and multiplying the result by $X'\mathbf{V}X$. This latter term consists of pre-and post-multiplying the variance-covariance matrix of the coefficients by the *x* vector specified at the values of interest.

The standard error for the difference between the two estimates of probabilities being used to illustrate the effect of *x* in this example, is a bit more complicated but follows the same general delta method:

$$\text{Asy. Var}\,[\,\hat{F}_a - \hat{F}_c\,] = [\,d\big(\hat{F}_a - \hat{F}_c\big)/\,d\hat{\mathrm{B}}\,]'\;\mathbf{V}\;[\,d\big(\hat{F}_a - \hat{F}_c\big)/\,d\hat{\mathrm{B}}\,],$$

$$\text{where } d\big(\hat{F}_a - \hat{F}_c\big)/\,d\hat{\mathrm{B}} \;=\; d\hat{F}_a\,/\,d\hat{\mathrm{B}} - d\hat{F}_c\,/\,d\hat{\mathrm{B}} = \hat{f}_a X_a - \hat{f}_c X_c$$

which implies:

$$\text{Asy. Var}\,[\,\hat{F}_a - \hat{F}_c\,] = \big[\hat{f}_a X_a - \hat{f}_c X_c\big]'\,\mathbf{V}\big[\hat{f}_a X_a - \hat{f}_c X_c\big]$$

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 provide readers with a sense of the mathematics underlying the production of these standard errors and to reemphasize the importance of examining effects rather than simply coefficients (or predicted levels), be they estimated in a linear or nonlinear specification.

D. 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 versus stochastic nature of that relationship. 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 this sort probabilistic relationship. (The related multi-level-model or hierarchical-model literature addresses a similar issue, though with different assumptions about the properties of the stochastic aspects of the relationships: see below.)

Let us start thus:

$$y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon \qquad [39]$$

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

$$\beta_0 = \gamma_0\,,\ \beta_1 = \delta_1 + \delta_2 z \ \text{ and } \ \beta_2 = \delta_3 + \delta_4 x \qquad [40]$$

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

$$y = \beta_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon \qquad [41]$$

$$\text{where } \beta_x = \delta_1\ , \beta_z = \delta_3\,, \beta_{xz} = \delta_2 + \delta_4$$

The linear-interactive model with stochastic effects would instead be:

$$\beta_0 = \gamma_0 + \varepsilon_0\,,\ \beta_1 = \delta_1 + \delta_2 z + \varepsilon_1\,, \text{ and } \ \beta_2 = \delta_3 + \delta_4 x + \varepsilon_2 \qquad [42]$$

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

$$y = \gamma_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon^* \qquad [43]$$

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

Thus, as easily seen now, the distinction between the deterministically interactive and the stochastically interactive models occurs only in the “error” term; i.e., the two models are identical except in their error terms, ε versus ε^*. In the first case, where the conditioning effects are assumed to be deterministic in nature, OLS (provided the specified model is correct, of course) would be BLUE. In the second case, where the conditioning effects are assumed to be of a stochastic, or probabilistic, nature, one suspects OLS estimates might not be BLUE. Notice, however, that, assuming these stochastic terms have mean zero, E($\varepsilon_0,...,\varepsilon_k$) = 0, and do not covary with the regressors, Cov[($\varepsilon_0,...,\varepsilon_k$),X]=0, as commonly done in most regression contexts including hierarchical modeling, OLS estimation would still yield unbiased and consistent coefficient-estimates. On the other hand, the composite residual's variance, V(ε^*), is not constant, which violates homoskedasticity (even if V(ε_0), … ,V(ε_k) are all constant), rendering coefficient estimates and standard errors inefficient. Moreover, this non-constant variance, i.e. heteroskedasticity, in fact depends on 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 had constant variance, 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 above, similar modeling issue arises in the literature on hierarchical, or multi-level models (see, e.g., Bryk and Raudenbush 2001). Often these models propose that some individual-level y_{ij} is dependent on a contextual-level variable varying only across and not within the *j* contexts, z_j, and an individual-level variable, x_{ij}, and furthermore that the effect of the individual-level variable x_{ij} depends (deterministically or stochastically) on z_j:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + \varepsilon_{ij} \qquad [44]$$

$$\beta_0 = \gamma_0 + (\varepsilon_{0ij})$$

$$\beta_1 = \delta_1 + \delta_2 z_j (+\varepsilon_{1j})$$

$$\beta_2 = \delta_3 + \delta_4 x_{ij} (+\varepsilon_{2ij})$$

which implies that one may model *y* for regression analysis as:

$$y = \gamma_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon^* \qquad [45]$$

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

Assuming deterministic conditional relationships so that $\varepsilon^* = \varepsilon_{ij}$ and the parenthetical terms are all zero and assuming that this residual is well-behaved (mean zero, constant variance, and no correlation with regressors, as usual), OLS is BLUE of course. 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 non-constant variances and correlations were themselves uncorrelated with the regressors (their cross-products or squares). In the case that these patterns correlated in some fashion with the regressors (their cross-products or 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, as usual in this context. These standard-error inconsistency problems could be redressed as usual 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 a common correlation of stochastic components within units *i*.

With stochastic dependence such that $\varepsilon^* = \varepsilon_{ij} + \varepsilon_{0ij} + \varepsilon_{1j} x_{ij} + \varepsilon_{2ij} z_j$, OLS coefficient-estimates are still unbiased and consistent, but we see that the error term presents us with two issues even in the case of well-behaved ε_{ij}: heteroskedasticity (with the composite residual term, ε^*, depending on some linear combination of *x* and *z*) as well as autocorrelation (since ε_{1j} will be shared amongst all individuals *i* in context *j*).[41]

Thus, the random-effects case and the multi-level case produce identical problems with OLS. Accordingly, the same solutions will apply. To begin, notice that some form of the familiar White's or Huber-White's style consistent variance-covariance estimators, i.e. so-called "robust" standard errors, will redress the inconsistency in the OLS estimates of the estimated coefficients' variance-covariance. Recall that, given nonspherical disturbances,

$$
\begin{aligned}
\text{V}(\hat{B}) &= \text{E}[(\text{B} - \hat{B})(\text{B} - \hat{B})'] \qquad [46]\\
&= \text{E}[(\text{X'X})^{-1}\text{X}'(\varepsilon)[(\text{X'X})^{-1}\text{X}'(\varepsilon)]']\\
&= \text{E}[(\text{X'X})^{-1}\text{X}'\varepsilon\varepsilon'[(\text{X'X})^{-1}\text{X}]]\\
&= (\text{X'X})^{-1}\text{X}'\ \text{E}[\varepsilon\varepsilon'](\text{X'X})^{-1}\text{X}
\end{aligned}
$$

Under classical linear regression model assumptions, ε~N$(0,\sigma^2)$, E(εX)=0.

In the random coefficient case, $\varepsilon^* = \varepsilon + \varepsilon_0 + \varepsilon_1 x + \varepsilon_2 z$. In the hierarchical model case, $\varepsilon^* = \varepsilon_{ij} + \varepsilon_{0ij} + \varepsilon_{1j} x_{ij} + \varepsilon_{2ij} z_j$. Both of these cases violate the assumptions of the classical linear regression model in essentially the same way:

[41] Some of the current literature suggests that OLS coefficients are *biased* in the presence of multi-level random effects. This is false. Provided that the context-specific or other components of the composite error term do not correlate with the regressors, OLS will remain unbiased and consistent. The fact that Z_j and ε_j are both common to all individuals in context *j* implies that the pattern of the non-sphericity in the composite V(ε^*) relates to a regressor, *Z*, producing biased, inconsistent, and inefficient OLS standard-error estimates, but does not imply that the Cov(Z_j,ε^*) is non-zero. As long as this latter term is zero, OLS coefficient-estimates are unbiased and consistent, although inefficient. The inefficiency may be sufficiently great as to render unbiasedness and consistency of little practical comfort, but the problem is not bias. The "problem" with OLS for hierarchical models therefore resides solely in the efficiency of OLS coefficient-estimates and in the OLS estimates of the variance-covariance matrix of b. 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. Again, the inefficiency in coefficient estimates can be dramatic if the within-context correlation of individual errors is great, but, even so, bias and inconsistency do not correctly describe the problem.

In our random coefficient case:

E($\varepsilon\varepsilon'$) = $E(\varepsilon+\varepsilon_0+\varepsilon_1 x+\varepsilon_2 z)(\varepsilon+\varepsilon_0+\varepsilon_1 x+\varepsilon_2 z)'$ [47]

$=E(\varepsilon\varepsilon'+\varepsilon_0\varepsilon'+\varepsilon_1 x\varepsilon'+\varepsilon_2 z\varepsilon'+\varepsilon\varepsilon_0'+\varepsilon_0\varepsilon_0'+\varepsilon_1 x\varepsilon_0'+\varepsilon_2 z\varepsilon_0'+\varepsilon x'\varepsilon_1'+\varepsilon_0 x'\varepsilon_1'+\varepsilon_1 xx'\varepsilon_1'+\varepsilon_2 zx'\varepsilon_1'$

$+\varepsilon z'\varepsilon_2'+\varepsilon_0 z'\varepsilon_2'+\varepsilon_1 xz'\varepsilon_2'+\varepsilon_2 zz'\varepsilon_2')$

Even assuming $\varepsilon,\varepsilon_0,...,\varepsilon_2$ iid ~N(0, σ^2), the variance-covariance matrix for $\hat{B}$ in the random coefficient model is:

V($\hat{B}$)$_{RC}$ = $2\sigma^2$ + xx' σ^2 + zz' σ^2 = σ^2(2**I** + xx'+ zz') [48]

In the hierarchical model case, the basic structure is the same but the assumption that $\varepsilon,\varepsilon_0,...,\varepsilon_2$ iid ~N(0, σ^2), is less credible because it is unlikely that the context-level variance (ε_{1j}) would be equal to individual-level variance ($\varepsilon_{ij},\varepsilon_{0ij},\varepsilon_{2ij}$). It might be more plausible to assume that between-level variation differs but within-level variation is constant. In that case, the variance-covariance matrix for $\hat{B}$ in the hierarchical model is:

V($\hat{B}$)$_{HC}$ = $2\sigma_{ind}{}^2$ + $\sigma_{context}{}^2$ [xx'] + $\sigma_{ind}{}^2$ [zz'] [49]

Notice from these last two equations that the expression for V(b)$_{HM}$ in the hierarchical case has almost identical mathematical form to the expression for V(b)$_{RC}$ in the random-coefficient case. The only difference is the separation we allow for the variances of components of ε^* in the hierarchical case, V($\hat{B}$)$_{HC}$, because they seem substantively sensible, that we do not allow in the random-coefficient case, V($\hat{B}$)$_{RC}$, because they may not. In either case, the familiar class of robust estimators (and/or reasonably familiar versions of Feasible Generalized Least Squares) will redress the problems sufficiently in a relatively straightforward manner.

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

Est. Asy. Var[b] = n (**X'X**)$^{-1}$ S$_0$ (**X'X**)$^{-1}$

where $S_0 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 x_i x_i'$

As Greene (2003) writes, the 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 (non-constant variance) whose pattern is somehow related to a pattern in *xx'*, i.e. to some pattern in the regressors, the regressors squared, or the cross-products of the regressors. In the random coefficient case, White's estimator provides consistency ("robustness") to precisely the heteroskedasticity issue raised because the pattern of non-constant 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 component, 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.

One issue with that would remain with either sort of "robust" standard-error estimator is that coefficient and standard error estimates would remain inefficient. To redress this issue, a Feasible Weighted Least Squares (FWLS) approach may be appropriate for the pure heteroskedasticity induced by simple random effects, and a Feasible Generalized Least Squares (FGLS) approach 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 as usual 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 root of these predicted values as weights for the FWLS re-estimation.[42] 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 root in step (6). The procedure for implementing FGLS if appropriate and desired is similar, except that both variance and covariance parameters are to be estimated in steps (3)-(4) for insertion into the estimated V($\hat{\varepsilon}$) matrix whose "square-root inverse" is to provide the weighing matrix in step (6). Also, the "square-root inverse" of a matrix with non-zero off-diagonal elements is not a simple inversion of the square root of each of the elements, as it is in the FWLS case. However, most existing statistical software packages will find the "square-root inverse" of a matrix, so we need not detain the reader with these computations.

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.[43] In all cases, the data were actually generated using hierarchical model structures (with several alternative relative variances and covariances of the

[42] One could iterate this procedure

[43] The variance-covariance matrix for coefficients estimated with the particular robust cluster we implemented (using Stata) is: V(b) = $\mathbf{(X'X)^{-1}S_J(X'X)^{-1}}$ where $S_j = \sum_{j=1}^{J} u_j ' u_j$ and where $u_j = \sum_{i=1}^{nj} e_{ij} x_{ij}$. We estimated the random effects model using HLM software.

error components and the right-hand-side variables) and in samples with 50 *i* units and 100 observations per unit (to correspond to a rather small survey conducted in each of the 50 US 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 nearly as good with the robust-cluster-estimation strategy as with the full random effects model (the standard errors were within 5% of each other). Appreciable efficiency gains in coefficient estimates from the hierarchical models relative to the OLS ones were also notably absent. Accordingly, the main conclusion of our exercise was that one seemed generally to have little to gain from complicated random-coefficients and hierarchical modeling strategies as any of the more-familiar and far-easier-to-implement OLS with robust Variance-Covariance estimators appended to OLS (e.g., in STATA, one simply appends ", robust" or ", robust cluster" to the end of the estimation command) seemed generally to suffice.[44]

VI. SUMMARY

In this brief manuscript, we have emphasized the importance of understanding the links between substantive theory and empirical tests of those theories. Political scientists often formulate hypotheses that demand some complexity beyond the simple linear additive model. Multiplicative interaction terms provide one simple means often sufficient to enhance the match of these complex theories to appropriate empirical statistical analyses.

We conclude with this summary of our recommendations on the use and interpretation of

[44] Of course, we would demand much further simulation, across wider and more-systematically varying ranges of parameters and sample dimensions to support this conclusion more whole-heartedly as a general one, but it certainly seems, across a reasonably broad range of parameters in this sample-dimension context at least, that the simpler strategies work almost indistinguishably from the more complex ones, so we're happy to argue for simplicity.

interactive terms in linear regression models. In order:

- *Theory*: What is the scientific phenomenon to be studied? Does your theory suggest that the effects of some variable(s) X depend on some other variable(s) Z (which implies the converse that the effect(s) of Z depend(s) on X)? Does it imply anything more specific about the manner in which the effects of X and of Z depend on each other?
- *Model*: What is the appropriate mathematical model to express your theory? Write the formal mathematical expression that encapsulates your theory. In the case where the theory implies that the effect(s) of X depend(s) on Z and *vice versa*, (a) simple multiplicative interaction term(s) will often suffice to express that (those) proposition(s). If the theory implies something more specific, ideally one would specify that more-specific (perhaps nonlinear) form of the interactions.
- *Estimation*: Estimate the model with an appropriate estimation strategy; OLS (or nonlinear regression model) with some "robust" or "robust cluster" standard-error estimator will usually suffice for most interactive propositions.
- *Interpretation*: What are the substantive effects of interest? Calculate marginal effects using first derivatives to describe the effects of the variable(s) of interest, X and/or Z at various, meaningful levels of the other variables. Calculate changes in the predicted values of y induced as some variable(s) of interest, X and/or Z, change(s) at various, meaningful levels of the other variables. Either way, calculate also the standard errors of these estimated effects.
- *Presentation*: Present tables or graphs including **both** marginal effects and accompanying measures of uncertainty or including **both** predicted values and

accompanying measures of uncertainty, plotting or tabulating both at various, meaningful levels of the other variables.

VII. APPENDIX A: DIFFERENTIATION RULES.

Table of differentiation rules (for a more complete list of differentiation rules, we refer the reader to Kleppner and Ramsey 1985).

Let b = a constant
y = a function of some variable(s)
X, Z, W = variables
f(), g() = functions

Expression	**dy/dX**	**Verbal Description**	**Example**
y = a	$da / dx = 0$	The derivative of a constant is zero.	*d*7/*d*X=0
y = bZ	$d(bZ) / dx = 0$	The derivative of a term that does not depend on *X* is zero.	*d*3Z/*d*X=0
y = bX	$d(bX) / dX = b$	The derivative of a term involving a linear coefficient and *X* is that coefficient.	*d*3X/*d*X=3
y = bX$^{\mathbf{n}}$	$d(bX^n) / dX = nbX^{n-1}$	The derivative of a term involving a linear coefficient and *X* raised to the nth power is the product of n, b, and X raised to the (n-1) power.	*d*3X^4/*d*X=12X^3
y = bXZ	$d(bXZ) / dy = bZ$	The derivative of a term involving a linear coefficient, *X*, and another variable *Z* is the product of the coefficient and the variable (we can treat the variable essentially as a constant here, with respect to *X*).	*d*3XZ/*d*X=3Z

y = bXZW	$d(bXZW)/dx = bZW$	The result extends to higher-order interactions, where again the variables that are not a function of the variable with respect to which one is differentiating are treated as fixed.	d3XZW/dX=3ZW
y = a + b$_x$X + b$_z$Z + b$_{xz}$XZ + b$_w$W	$\frac{\partial a}{\partial X}+\frac{\partial b_x X}{\partial X}+\frac{\partial b_z Z}{\partial X}+\frac{\partial b_{xz} XZ}{\partial X}+\frac{\partial b_w W}{\partial X}$ $= b_x + b_{xz}Z$	The derivative of some linear additive function equals the sum of the derivative of each of the terms.	d(3X+4Z+5XZ+6W)/dX =3+0+5Z+0=3+5Z
y=f(g(X))	$(df/dg)(dg/dx)$	This is the chain rule for nested functions.	d(3X+4Z+5XZ+6W)^3/dX =3(3X+4Z+5XZ+6W)^2·(3+5Z)
F ≡ a cumulative probability function for the probability function f.	$\frac{\partial F(x)}{\partial x} = f(x)$	The derivative of any cumulative probability function is the corresponding probability density function.	$\frac{\partial \Phi(x)}{\partial x} = \phi(x)$

VIII. APPENDIX B: STATA SYNTAX.

Many statistical software packages are available to researchers. Because STATA is prominent in the discipline, we have provided STATA-based syntax for readers to use in following our advice.

A. Marginal Effects, Standard Errors, and Confidence Intervals.

To estimate the following "standard model," given variables *y, x,* and *z* in the dataset.

$$y = \beta_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon$$

The first step is to generate the multiplicative term, *xz*.

```
gen xz = x*z
```

Then the next step, naturally, is to run the linear regression model:

```
regress y x z xz
```

This command regresses y on *x*, *z*, and *xz.*

Recall that the marginal effects of *x* and *z* consists of the first derivative based on the estimation results:

$$dy / dx = \beta_x + \beta_{xz} z$$

$$dy / dz = \beta_z + \beta_{xz} x$$

The *estimated* marginal effects of *x* and *z* are calculated using the estimates of β_x, β_z, and β_{xz}:

$$dy / dx = \hat{\beta}_x + \hat{\beta}_{xz} z$$

$$dy / dz = \hat{\beta}_z + \hat{\beta}_{xz} x$$

Suppose one wanted to calculate the estimated marginal effects of *x* as *z* takes on evenly values from its minimum to its maximum. One could, technically, enter each evenly spaced value into a dataset (e.g., 1; 2; 3; etc.), but if the dataset contains more than a handful of observations, then this could quickly become tedious. A more efficient way of setting values of *z* is easily found. First, we create variables that capture the sample minimum and maximum values of *z*:

```
egen zmin = min(z)
egen zmax = max(z)
```

We can take advantage of STATA's stored assignment of "_n" which contains the number of the current observation. We create a variable z0 that takes on evenly spaced values between *z's* minimum to *z's* maximum*:*

$$\texttt{gen z0 =} \ \frac{(\_n-1)}{(\_N-1)}(z\max - z\min) + z\min$$

Theoretically, if the dataset contains 1001 observations, and, for example, we wanted z_0 to take on values between 2 and 5, for example, the expression above returns the proper values; observation 1 in the dataset is assigned a value of $\frac{(1-1)}{(1001-1)}(5-2)+2=2$, observation 501 in the dataset is assigned a value of $\frac{(501-1)}{(1001-1)}(5-2)+2=3.5$, and observation 1001 in the dataset is assigned a value of $\frac{(1001-1)}{(1001-1)}(5-2)+2=5$. This would be an excellent approach, but it also requires handling a very large matrix, and with even larger datasets, it would quickly become unwieldy.

Instead, we advise plotting some reasonable number (*v)* (for example, 10 or 100) of evenly spaced values between **`zmin`** and **`zmax`**:

$$\texttt{gen z0 =} \ \frac{(\_n-1)}{(v-1)}(z\max - z\min) + z\min \ \ \texttt{in 1/v}$$

Marginal effects are calculated by adding the estimated $\hat{\beta}_x$ to the product of each value in z0 with the estimated coefficient $\hat{\beta}_{xz}$. The user could simply take the estimated coefficients from the regression output and thus generate a new variable:

gen dyhatdx=$\hat{\beta}_x$+**z0***$\hat{\beta}_{xz}$

Where the estimated value $\hat{\beta}_x$ from the regression output (*e.g.,* "-2") is entered instead of "$\hat{\beta}_x$", and the estimated value of $\hat{\beta}_{xz}$ from the regression output (*e.g,* "10") is entered in place of "$\hat{\beta}_{xz}$". This command line would thus create *v* values corresponding with the marginal effects of *x*, as *z* varies from its minimum to its maximum. One disadvantage to the procedure above is that it is entirely possible that one might mistype one of the estimated values, and this would of course impose error in the calculated values. A less error-prone way of calculating the marginal effects of *x*, then, would be to take advantage of the estimates that STATA stores in memory.

For our purposes, STATA stores the estimated coefficient $\hat{\beta}_x$ in memory as **_b[x]** and the coefficient $\hat{\beta}_{xz}$ in memory as **_b[xz]**, so a variable that consists of the marginal effects of *x* as the variation in *z* is captured by **z₀** is generated as follows:

```
gen dyhatdx=_b[x]+_b[xz]*z0
```

The variable **dyhatdx** thus contains the estimated marginal effects of *x* along values of **z₀**. A table of selected marginal effects for evenly spaced values of interest could thus be easily created.

Marginal effects, however, are only part of the story. We reiterate that discussions of marginal effects should also include an indication of our level of certainty or uncertainty regarding these marginal effects. Recall that the variance of each marginal effect, following the specific example above, is formally expressed as:

$$\mathrm{var}(d\hat{y}/dx) = \mathrm{var}(\hat{\beta}_x) + z^2\ \mathrm{var}(\hat{\beta}_{xz}) + 2zCov(\hat{\beta}_x, \hat{\beta}_{xz})$$

Thus, calculating $\mathrm{var}(d\hat{y}/dx)$ is a fairly straight-forward computational expression. The variance-covariance matrix of the estimated coefficients is available for display by typing "**`vce`**" after viewing the regression output.

The user could then simply generate a new variable by taking the specific values of Var($\hat{\beta}_x$), Var($\hat{\beta}_{xz}$), and Cov($\hat{\beta}_x$, $\hat{\beta}_{xz}$), acquired from viewing the values in the variance-covariance matrix.

`gen vardyhatdx` = $\mathrm{var}(\hat{\beta}_x) + z0 * z0 * \mathrm{var}(\hat{\beta}_{xz}) + 2 * z0 * Cov(\hat{\beta}_x, \hat{\beta}_{xz})$

Where Var($\hat{\beta}_x$), Var($\hat{\beta}_{xz}$), and Cov($\hat{\beta}_x$, $\hat{\beta}_{xz}$), would be replaced by their estimated values (*e.g.,* "2").

Again, although this "enter by hand" method is transparent, human error in data entry could lead to erroneous values. A less error-prone way of calculating the variances of the marginal effects, then, would be to take advantage of the estimates that STATA stores in memory. The variance can be calculated by utilizing the estimates that STATA stores in memory: Var($\hat{\beta}_x$) is stored as its square root in the form of "**`_se[x]`**"; Var($\hat{\beta}_{xz}$)is stored as its square root in the form of "**`_se[xz]`**", and Cov($\hat{\beta}_x$, $\hat{\beta}_{xz}$) is stored as the element in the respective cell in the estimated variance-covariance matrix of the coefficient estimates, **`VCE`**. In this particular case, given the order in which the variables are entered in the estimated equation, Cov($\hat{\beta}_x$, $\hat{\beta}_{xz}$) is located in the third row, first column of the estimated variance-covariance matrix (and, because the variance-covariance matrix is symmetric, it is also located in the first row, third column as well).

First, we create a new matrix **`V`** to represent the variance-covariance matrix of the coefficient estimates, **VCE**.

```
matrix V = get(VCE)
```

We can pull out the stored element Cov($\hat{\beta}_x$, $\hat{\beta}_{xz}$) as follows:

```
matrix C= V[3,1]
```

This command line pulls out the Cov($\hat{\beta}_x$, $\hat{\beta}_{xz}$) value, which is stored as element [3,1] in the variance-covariance matrix (note that the specific row and column numbers will depend on the order in which the variables of interest are entered into the model; in our model, *x* was entered first and *xz* third, which is why element [3,1] contains the covariance between the respective coefficients).

The command line extracts this value as a 1x1 matrix (a scalar).

We then generate a variable containing only values of 1, to allow us to apply this covariance to all observations:

```
gen column1 = 1 in 1/v
```

We then convert this column of 1's into a vector, to allow us to apply this covariance to all observations:

```
mkmat column1, matrix(col1)
```

Finally, we multiply our 1x1 matrix **C** by this vector of 1's:

```
matrix cov_x_xz = C*col1
```

The vector **cov_x_xz** contains the value of Cov($\hat{\beta}_x$, $\hat{\beta}_{xz}$) for all *v* observations. Finally, we convert the vector into a variable **cov_x_xz1**, which contains the covariance of interest.

```
svmat cov_x_xz, name(cov_x_xz)
```

The variance of each marginal effect can thus be calculated as:

```
gen vardyhatdx=(_se[x])^2+(z0*z0)*(_se[xz]^2)+2*z0*cov_x_xz1
gen sedyhatdx=sqrt(vardyhatdx)
```

Tables of marginal effects can thus include either the standard errors that correspond with those marginal effects. A table could be generated as follows:

```
tabdisp z0, cellvar(dyhatdx sedyhatdx)
```

This command line would present a table featuring all *v* values of **z₀**, with the appropriate marginal effect and standard error of the marginal effect. This table is likely to be useful for the researcher for interpretation, but for presentational purposes, only a set of selected values of **z₀** might be incorporated into an abbreviated table.

Alternatively, marginal effects can be graphed. Recall that confidence intervals can be generated with the following formula:

$$d\hat{y}/dx \pm t_{df,p}\sqrt{Var(d\hat{y}/dx)}$$

STATA stores the degrees of freedom from the previous estimation as "e(df_m)," and the researcher can utilize the inverse t-distribution function to create the multiplier $t_{df,p}$. For a 95% confidence interval, the lower and upper bounds are calculated as follows:

```
gen LBdyhatdx=dyhatdx-invttail(e(df_m),.05)*sedyhatdx
gen UBdyhatdx=dyhatdx+invttail(e(df_m),.05)*sedyhatdx
```

The graph of marginal effects and accompanying confidence intervals can be created by the following command line:

```
twoway connected dyhatdx LBdyhatdx UBdyhatdx z0
```

These procedures are summarized in Table B1.

[TABLE B1 ABOUT HERE]

B. Predicted Values, Standard Errors, and Confidence Intervals.

To estimate the following "standard model," given variables *y, x,* and *z* in the dataset.

$$y = \beta_0 + \beta_x x + \beta_z z + \beta_{xz} xz + \varepsilon$$

The first step is to generate the multiplicative term, *xz*.

```
gen xz = x*z
```

Then the next step, naturally, is to run the linear regression model:

```
regress y x z xz
```

This command regresses y on *x*, *z*, and *xz.*

The predicted values, $\hat{y}$, are generated by summing the products between set values of the right-hand-side variables and their corresponding coefficients: $\hat{y} = \mathbf{M_h} \hat{\beta}$, where $\mathbf{M_h}$ is a matrix of values at which *x*, *z*, and any other variables in the model are set.

The next step is create the matrix $\mathbf{M_h}$ which sets the values of the variables *x, z,* and *xz* for which $\hat{y}$ will be calculated. Suppose we are interested in the effects of *x* as it varies across a range of values, from value *a*, its minimum, to *c*, its maximum, for example, and we want to hold *z* at a specific value z_h (for example, its sample mean).

As above, we recommend calculating these values for the first *v* observations in the dataset.

```
egen xmin = min(x)

egen xmax = max(x)
```

$$\texttt{gen xh =} \ \frac{(\_n-1)}{(v-1)}(x\max - x\min) + x\min \quad \texttt{in 1/v}$$

Next, we create a variable that sets z at some specified value; here, we set it specifically to equal its mean.

```
egen zh=mean(z)
```

Next, we create the interactive term that reflects the values to which *x* and *z* are held:

```
gen xhzh=xh*zh
```

This step creates a variable that consists of a column of 1's (to be multiplied by the intercept term).

```
gen col1h=1
```

Now that we have created these separate variables, we need to assemble them into a matrix:

```
mkmat xh zh xhzh col1h in 1/v, matrix(Mh)
```

This command creates matrix **`Mh`** which contains the specified values at which our variables are set: **`zh`** is fixed at its mean, **`xh`** varies at regular intervals between *a* and *c*, and **`xhzh`** correspondingly varies, since it is the product of **`zh`** and **`xh`**.

Recall that $\hat{y} = \mathbf{M_h}\hat{\beta}$. $\hat{\beta}$ is a column vector of coefficients, but STATA stores the estimated coefficients as a *1 x k* row vector, **`e(b)`**. So we want to create **`betas`**, a column vector with *k x 1* dimensions, that takes the stored coefficients and transposes them into ($\hat{\beta}$):

```
matrix betas=e(b)'
```

This command creates the column vector of estimated coefficients.

Calculating the predicted values is simply a matter of multiplying **`Mh`** by **`betas`**. This command creates a column vector that contains the product of the specified values and the coefficient vector.

```
matrix yhat=Mh*betas
```

Then convert the column vector into a variable, **`yhat1`**.

```
svmat yhat, name(yhat)
```

Recall that Var($\hat{y}$)=$\mathbf{M_h}$Var($\hat{\beta}$)$\mathbf{M_h}$'**.** As mentioned above, STATA stores the estimated variance-covariance matrix of the estimated coefficients, Var($\hat{\beta}$), as VCE in its memory. We want to create a matrix V to represent the estimated Var($\hat{\beta}$):

```
matrix V=VCE
```

We can now calculate the variance of the predicted values as follows:

```
matrix VYH=Mh*V*Mh'
```

This command creates a matrix, VYH, that contains variances and covariances of predicted values. The diagonal elements in the variance-covariance matrix of predicted values are those of interest to us, as they correspond with the estimated variance of the predicted values. We want to extract these diagonal elements into a vector, and we do so as follows:

```
matrix DVYH= vecdiag(VYH)
```

This command creates a row vector from the diagonal elements of the variance-covariance matrix of the predicted values. But we want a column vector rather than a row vector, so that we can convert the column vector into a variable that can later be graphed. This command transposes the row vector into a column vector.

```
matrix VARYHAT=DVYH'
```

This command creates a new variable, **`varyhat1`**, which contains a unique variance to correspond with each predicted value **`yhat1`**.

```
svmat VARYHAT, name(varyhat)
```

Taking the square root produces the standard error of each predicted value **`yhat1`**.

```
gen seyhat1 = sqrt(varyhat1)
```

The researcher can next present a table of predicted values with corresponding standard errors of the predicted values:

```
tabdisp xh, cellvar(yhat1 seyhat1)
```

Graphically, predicted values are more effectively displayed when graphed with confidence intervals. The confidence intervals around predicted values $\hat{y}$ can be constructed as follows:

$$\hat{y} \pm t_{df,p}\sqrt{Var(\hat{y})}$$

Where $\hat{y}$ corresponds with the values in **yhat1**, $\sqrt{Var(\hat{y})}$ corresponds with the values in **seyhat1**, and $t_{df,p}$ refers to the relevant critical value from the *t*-distribution. STATA stores the degrees of freedom from the previous estimation as "e(df_m)," and the researcher can utilize the inverse t-distribution function to create the multiplier $t_{df,p}$.

For a 95% confidence interval, the lower and upper bounds are calculated as follows:

```
gen LByhat1=yhat1-invttail(e(df_m),.05)*sqrt(varyhat1)
gen UByhat1=yhat1+invttail(e(df_m),.05)*sqrt(varyhat1)
```

These values can easily be graphed along values of X_h as follows:

```
twoway connected yhat1 LByhat1 UByhat1 xh
```

These procedures are summarized in Table B2.

[TABLE B2 ABOUT HERE]

IX. REFERENCES

Achen, Christopher H. 1982. *Interpreting and using regression.* Thousand Oaks, CA: Sage Publications.

Allison, Paul D. 1979. “Testing for interaction in multiple regression.” *American journal of sociology* 83 (1): 144-53.

Althauser, Robert P. 1971. “Multicollinearity and non-additive regression models.” In *Causal models in the social sciences,* ed. H. M. Blalock, Jr. Chicago: Aldine Atherton, 453-72.

Baron, Reuben M. and David A. Kenny. 1986. “The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations.” *Journal of Personality and Social Psychology* 51 (6): 1173-82.

Berry, Frances Stokes and William D. Berry. 1990. “State lottery adoptions as policy innovations: An event history analysis.” *American political science review* 84 (2): 395-415.

-----.1991. “Specifying a model of state policy innovation.” *American Political Science Review* 85 (2): 573-9.

Blalock, H. M., Jr. 1971. *Causal models in the social sciences*. Chicago: Aldine Atherton.

Blalock, Hubert M., Jr. 1969. *Theory construction: From verbal to mathematical formulations*. Englewood Cliffs, NJ: Prentice-Hall.

Bryk, Anthony S. and Stephen W. Raudenbush. 2001. *Hierarchical linear models, second edition: Applications and data analysis methods*. Newbury Park, CA: Sage.

Cox, Gary W. 1997. *Making votes count: Strategic coordination in the world’s electoral systems*. Cambridge, UK: Cambridge University Press.

Cronbach, Lee J. 1987. “Statistical tests for moderator variables: Flaws in analyses recently proposed.” *Psychological bulletin* 102 (3): 414-7.

Dunlap, William P. and Edward R. Kemery. 1987. “Failure to detect moderating effects: Is multicollinearity the problem?” *Psychological bulletin* 102 (3): 418-20.

Fisher, Gene A. 1988. “Problems in the use and interpretation of product variables.” In *Common problems/Proper solutions: Avoiding error in quantitative research,* ed. J. Scott Long. Newbury Park: Sage, 84-107.

Frant, Howard. 1991. “Specifying a model of state policy innovation.” *American Political Science Review* 85 (2): 571-3.

Franzese, Robert J., Jr. 1999. "Partially independent central banks, politically responsive governments, and inflation." *American Journal of Political Science* 43 (3): 681-706.

-----2001. "Institutional and sectoral interactions in monetary policy and wage-price bargaining." In *Varieties of capitalism: The institutional foundations of comparative advantage,* ed. Peter A. Hall and D. Soskice. Cambridge: Cambridge University Press, 104-44.

-----.2002a. "Electoral and partisan cycles in economic policies and outcomes." *Annual reviews of political science* 5: 369-421.

-----.2002b. *Macroeconomic policies of developed democracies*. Cambridge, UK: Cambridge University Press.

-----.2003a. "Multiple hands on the wheel: Empirically modeling partial delegation and shared control of monetary policy in the open and institutionalized economy." *Political Analysis* 11 (4): 445-74.

-----2003b. "Strategic interactions of the ECB, wage/price bargainers, and governments: A review of theory, evidence, and recent experience." In *Institutional conflicts and complementarities: Monetary policy and wage bargaining institutions in EMU,* ed. Robert J. Franzese, Jr., Peter Mooslechner, and Martin Schürz. New York: Kluwer, 5-42.

Friedrich, Robert J. 1982. "In defense of multiplicative terms in multiple regression equations." *American journal of political science* 26 (4): 797-833.

Greene, William H. 2003. *Econometric analysis, 5th ed.* Upper Saddle River, NJ: Prentice Hall.

Hall, Peter A. 1986. *Governing the economy: The politics of state intervention in Britain and France*. Cambridge: Polity.

Hayduk, Leslie A. and Thomas H. Wonnacott. 1980. "'Effect equations' or 'effect coefficients': A note on the visual and verbal presentation of multiple regression interactions." *Canadian journal of sociology* 5 (4): 399-404.

Ikenberry, G. John. 1988. "Conclusion: An institutional approach to American foreign economic policy." *International organization* 42 (1): 219-43.

Jaccard, James, Robert Turrisi, and Choi K. Wan. 1990. *Interaction effects in multiple regression.* Newbury Park, CA: Sage Publications.

Jusko, Karen Long, and W. Phillips Shively. 2004. "A two-step strategy for the analysis of cross-national public opinion data." Paper prepared for presentation at the Conference on Models of Analysis for Multilevel Data, Princeton, NJ, October 22-24.

Kam, Cindy D., Robert J. Franzese, Jr., and Amaney A. Jamal. 1999. "Modeling interactive hypotheses and interpreting statistical evidence regarding them." Presented at the 1999 Annual Meeting of the American Political Science Association, Atlanta.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the most of statistical analyses: Improving interpretation and presentation." *American Journal of Political Science* 44 (2): 347-62.

Kleppner, Daniel and Norman Ramsey. 1985. *Quick calculus, 2nd edition*. New York: John Wiley & Sons.

Lijphart, Arend. 1994. *Electoral systems and party systems: A study of twenty-seven democracies, 1945-1990*. Oxford: Oxford University Press.

Lohmann, Susanne. 1992. "Optimal commitment in monetary policy: Credibility versus flexibility." *American economic review* 82(1): 273-286.

Long, J. Scott. 1988. *Common problems/Proper solutions: Avoiding error in quantitative research*. Newbury Park: Sage.

Marsden, Peter V. 1981. "Conditional effects in regression models." In *Linear models in social research,* ed. Peter V. Marsden. Beverly Hills: Sage, 97-116.

-----1981. *Linear models in social research*. Beverly Hills: Sage.

Morris, J. H., J. D. Sherman, and E. R. Mansfield. 1986. "Failures to detect moderating effects with ordinary least squares-moderated multiple regression: Some reasons and a remedy." *Psychological bulletin* 99: 282-8.

Nagler, Jonathan. 1991. "The effect of registration laws and education on U.S. voter turnout." *American political science review* 85 (4): 1393-405.

Neto, Octavio Amorim and Gary W. Cox. 1997. "Electoral institutions, cleavage structures, and the number of parties." *American Journal of Political Science* 41 (1): 149-74.

Ordeshook, Peter and Olga Shvetsova. 1994. "Ethnic heterogeneity, district magnitude, and the number of parties." *American Journal of Political Science* 38 (1): 100-23.

Schuessler, Karl F. 1979. *Sociological methodology 1980*. San Francisco: Jossey-Bass.

Shepsle, Kenneth. 1989. "Studying institutions: Some lessons from the rational choice approach." *Journal of theoretical politics* 1: 131-47.

Smith, Kent W. and S. W. Sasaki. 1979. "Decreasing multicollinearity: A method of models with multiplicative functions." *Sociological methods and research* 8: 35-56.

Southwood, Kenneth E. 1978. "Substantive theory and statistical interaction: Five models." *American journal of sociology* 83 (5): 1154-203.

Steinmo, Sven, Kathleen Thelen, and Frank Longstreth. 1992. *Historical institutionalism in comparative politics* . Cambridge: Cambridge University Press.

Stolzenberg, Ross M. 1979. "The measurement and decomposition of causal effects in nonlinear and nonadditive models." In *Sociological methodology 1980,* ed. Karl F. Schuessler. San Francisco: Jossey-Bass, 459-88.

Tate, Richard L. 1984. "Limitations of centering for interactive models." *Sociological methods and research* 13 (2): 251-71.

Tsebelis, George. 2002. *Veto players: How political institutions work*. New York: Russell Sage Foundation.

Western, Bruce. 1998. "Causal heterogeneity in comparative research: A bayesian hierarchical modelling approach." *American journal of political science* 42 (4): 1233-59.

Wolfinger, Raymond E., and Steven J. Rosenstone. 1980. *Who votes?* New Haven: Yale University Press.

Wright, Gerald C., Jr. 1976. "Linear models for evaluating conditional relationships." *American journal of political science* 20 (2): 349-73.

Zedeck, Sheldon. 1971. "Problems with the use of 'moderator' variables." *Psychological bulletin* 76 (4): 295-310.

X. TABLES.

Table 1. OLS Regression Results.

	Coefficient (standard error)	2-Sided P-Level (probability |T|>t) referring to Null Hypothesis that β=0
Ethnic Groups	-0.979 (0.770)	0.136
Runoff	-2.491 (1.561)	0.228
Ethnic Groups * Runoff	2.005 (0.941)	0.054
Intercept	4.303 (1.229)	0.004
N	16	
Adjusted R^2	0.203	
P>F	0.132	

Table 2. Predicted Effective Number of Presidential Candidates

	When Runoff = 0	When Runoff = 1
Ethnic Groups = 1	3.324	2.838
Ethnic Groups = 1.5	2.835	3.351
Ethnic Groups = 2	2.345	3.865
Ethnic Groups = 2.5	1.855	4.378
Ethnic Groups = 3	1.366	4.891

Table 3. Does *Y* Depend on *X* or *Z*?

Hypothesis	Mathematical Expression[45]	Statistical test
X affects Y, or *Y is a function of (depends on) X*	$Y=f(X)$ $dy/dx=b_x+b_{xz}Z \neq 0$	*F- test:* H_0: $b_x = b_{xz}=0$
X increases Y	$dy/dx= b_x+b_{xz}Z > 0$	*Multiple t-tests:* H_0: $b_x+b_{xz}Z \leq 0$
X decreases Y	$dy/dx = b_x+b_{xz}Z < 0$	*Multiple t- tests:* H_0: $b_x+b_{xz}Z \geq 0$
Z affects Y, or *Y is a function of (depends on) Z*	$Y=g(Z)$ $dy/dz= b_z+b_{xz}X \neq 0$	*F- test:* H_0: $b_z=b_{xz}=0$
Z increases Y	$dy/dx= b_z+b_{xz}X > 0$	*Multiple t-tests:* H_0: $b_z+b_{xz}X \leq 0$
Z decreases Y	$dy/dx= b_z+b_{xz}X < 0$	*Multiple t- tests:* H_0: $b_z+b_{xz}X \geq 0$

[45] The mathematical expression refers to the standard linear-interactive model, such as that in Equation [17].

Table 4: Is *Y*'s Dependence on *X* Conditional on *Z* and *vice versa*? How?

Hypothesis	Mathematical Expression[46]	Statistical test
The effect of X on Y depends on Z	$Y=f(XZ,\bullet)$ $dy/dx=b_x+b_{xz}Z=g(Z)$ $d(dy/dx)/dZ$ $=d^2Y/dXdZ=b_{xz}=0$	t-test: H_0: $b_{xz}=0$
The effect of X on Y increases in Z	$d(dy/dx)/dZ$ $=d^2Y/dXdZ=b_{xz}>0$	t-test: H_0: $b_{xz}\leq 0$
The effect of X on Y decreases in Z	$d(dy/dx)/dZ$ $=d^2Y/dXdZ=b_{xz}<0$	t-test: H_0: $b_{xz}\geq 0$
The effect of Z on Y depends on X	$Y=f(XZ,\bullet)$ $dy/dz=b_z+b_{xz}X=h(X)$ $d(dy/dz)/dX$ $=d^2Y/dZdX=b_{xz}=0$	t-test: H_0: $b_{xz}=0$
The effect of Z on Y increases in X	$d(dy/dz)/dX$ $=d^2Y/dZdX=b_{xz}>0$	t-test: H_0: $b_{xz}\leq 0$
The effect of Z on Y decreases in X	$d(dy/dz)/dX$ $=d^2Y/dZdX=b_{xz}<0$	t-test: H_0: $b_{xz}\geq 0$

[46] The mathematical expression refers to the standard linear-interactive model, such as that in Equation [17].

Table 5: Does Y Depend on X, Z, or Some Combination Thereof?

Hypothesis	Mathematical Expression[47]	Statistical Test
Y is a function of (depends on) X, Z, and/or their interaction	$Y=f(X,Z,XZ)$	*F-test:* H_0: $b_x=b_{xz}=b_z=0$

[47] The mathematical expression refers to the standard linear-interactive model, such as that in Equation [17].

Table 6. Variance-Covariance Matrix of Coefficient Estimates

	Ethnic Groups	*Runoff*	*Ethnic Groups * Runoff*	*Intercept*
Ethnic Groups	0.593			
Runoff	0.900	2.435		
Ethnic Groups Runoff*	-0.593	-1.377	0.885	
Intercept	-0.900	-1.509	0.900	1.509

Table 7. Hypothesis Tests of whether *Ethnic Groups* affects *Number of Presidential Candidates*.

	*d*Y/*d*EG	Var (*d*Y/*d*EG)	s.e. (*d*Y/*d*EG)	t-statistic	One-tailed p-value H_0: $\beta_{EG}+\beta_{EG*R}$Runoff ≤ 0	One-tailed p-value H_0: $\beta_{EG}+\beta_{EG*R}$Runoff ≥ 0
Runoff = 0	-0.979	0.593	0.770	-1.260	0.884	0.116
Runoff = 1	1.026	0.292	0.540	1.900	0.041	0.959

Table 8. Hypothesis Tests of whether *Runoff* affects *Number of Presidential Candidates*.

	*d*Y/*d*EG	Var (*d*Y/*d*EG)	s.e. (*d*Y/*d*EG)	t-statistic	One-tailed p-value H_0: $\beta_{R}+\beta_{EG*R}$Runoff ≤ 0	One-tailed p-value H_0: $\beta_{R}+\beta_{EG*R}$Runoff ≥ 0
Ethnic Groups=1	-0.486	2.134	1.461	-0.333	0.627	0.373
Ethnic Groups=1.5	0.517	2.647	1.627	0.318	0.378	0.622
Ethnic Groups=2	1.520	3.603	1.898	0.801	0.219	0.781
Ethnic Groups=2.5	2.522	5.001	2.236	1.128	0.141	0.859
Ethnic Groups=3	3.525	6.842	2.616	1.348	0.101	0.899

Table 9. Confidence Interval around Marginal Effect of *Ethnic Groups.*

	*d*Y/*d*EG	Var(*d*Y/*d*EG)	90% Confidence Interval
Runoff = 0	-0.979	0.593	[-2.351, 0.393]
Runoff = 1	1.026	0.292	[0.063, 1.989]

Table 10. Confidence Interval around Marginal Effect of *Runoff.*

	*d*Y/*d*Runoff	Var*(d*Y/*d*Runoff)	90% Confidence interval
Ethnic Groups = 1	-0.486	2.134	[-3.089, 2.118]
Ethnic Groups = 1.5	0.517	2.647	[-2.383, 3.417]
Ethnic Groups = 2	1.520	3.603	[-1.863, 4.903]
Ethnic Groups = 2.5	2.522	5.001	[-1.463, 6.508]
Ethnic Groups = 3	3.525	6.842	[-1.137, 8.187]

Table 11. Variance of Predicted Values of Effective Number of Presidential Candidates

	$\hat{y}$ |Runoff = 0	Var($\hat{y}$ |Runoff=0)	$\hat{y}$ |Runoff = 1	Var($\hat{y}$ |Runoff=1)
Ethnic Groups = 1	3.324	0.302	2.838	0.264
Ethnic Groups = 1.5	2.835	0.143	3.351	0.152
Ethnic Groups = 2	2.345	0.281	3.865	0.186
Ethnic Groups = 2.5	1.855	0.715	4.378	0.366
Ethnic Groups = 3	1.366	1.446	4.891	0.692

Table 12. Confidence Intervals for Predicted Values of Effective Number of Presidential Candidates

	$\hat{y}$ | Runoff = 0	90% Confidence Interval ($\hat{y}$ |Runoff=0)	$\hat{y}$ | Runoff = 1	90% Confidence Interval ($\hat{y}$ |Runoff=1)
Ethnic Groups = 1	3.324	[2.345, 4.304]	2.838	[1.923,3.754]
Ethnic Groups = 1.5	2.835	[2.160, 3.509]	3.351	[2.657, 4.046]
Ethnic Groups = 2	2.345	[1.400, 3.290]	3.865	[3.096, 4.633]
Ethnic Groups = 2.5	1.855	[0.348, 3.363]	4.378	[3.299, 5.456]
Ethnic Groups = 3	1.366	[-0.777, 3.509]	4.891	[3.408, 6.373]

Table B1. STATA commands for marginal effects of x on y, standard errors, and confidence intervals

Procedures	`Command syntax`
Generate the multiplicative term.	`gen xz = x*z`
Run the linear regression model.	`regress y x z xz`

Create range of *v* evenly spaced values for *z* from its minimum to its maximum.	`egen zmin = min(z)`
	`egen zmax = max(z)`
	`gen z0 = ((_n-1)/(v-1))*(zmax-zmin) in 1/v`

Calculate the estimated marginal effect.	`gen dyhatdx=_b[x]+_b[xz]*z0`
Create a matrix from the estimated covariance matrix of the coefficient estimates.	`matrix V = get(VCE)`
Pull out the stored element estimating Cov($\hat{\beta}_x$, $\hat{\beta}_{xz}$).	`matrix C= V[3,1]`
Generate a variable containing only values of 1.	`gen column1 = 1 in 1/v`
Convert this column of 1's into a vector.	`mkmat column1, matrix(col1)`
Create a vector with values of the estimated Cov($\hat{\beta}_x$, $\hat{\beta}_{xz}$).	`matrix cov_x_xz = C*col1`
Convert the vector into a variable.	`svmat cov_x_xz, name(cov_x_xz)`

Calculate the estimated variance of each estimated marginal effect.	`gen`
	`vardyhatdx=(_se[x])^2+(z0*z0)*(_se[xz]^2)+2*z0*cov_x_xz`

Calculate the estimated standard error of each estimated marginal effect.	`gen sedyhatdx=sqrt(vardyhatdx)`
Generate a table displaying estimated marginal effects and standard errors for all *v* values of $\mathbf{z}_0$	`tabdisp z0, cellvar(dyhatdx sedyhatdx)`

Generate lower and upper confidence interval bounds.	`gen LBdyhatdx=dyhatdx-invttail(e(df_m),.05)*sedyhatdx`
	`gen UBdyhatdx=dyhatdx+invttail(e(df_m),.05)*sedyhatdx`

Graph the estimated marginal effects and the upper and lower confidence intervals.	`twoway connected dyhatdx LBdyhatdx UBdyhatdx z0`

Table B2. STATA commands for calculating predicted values of y, standard errors, and confidence intervals

Procedures	Command syntax
Generate the multiplicative term.	gen xz = x*z
Run the linear regression model.	regress y x z xz
Create the variables which set the values of the variables *x, z,* and *xz* for which $\hat{y}$ will be calculated.	egen xmin = min(x) egen xmax = max(x) gen xh = ((_n-1)/(*v*-1))*(xmax-xmin) in 1/*v* egen zh=mean(z) gen xhzh=xh*zh gen col1h=1
Assemble the variables into a matrix, **Mh**	mkmat xh zh xhzh col1h in 1/*v*, matrix(Mh)
Create **betas**, a column vector with *k x 1* dimensions.	matrix betas=e(b)’
Calculate the predicted values.	matrix yhat=Mh*betas
Convert the vector into a variable.	svmat yhat, name(yhat)
Create a matrix from the estimated covariance matrix of the coefficient estimates.	matrix V = get(VCE)
Calculate the variance of the predicted values.	matrix VYH=Mh*V*Mh’
Extract the diagonal elements into a row vector.	matrix DVYH= vecdiag(VYH)
Tranpose elements into a column vector.	matrix VARYHAT=DVYH’
Convert the vector into a variable.	svmat VARYHAT, name(varyhat)
Calculate the estimated standard error of each predicted probability.	gen seyhat1 = sqrt(varyhat1)
Present a table of predicted values with corresponding standard errors of the predicted values.	tabdisp xh, cellvar(yhat1 seyhat1)
Generate lower and upper confidence interval bounds. Graph the predicted probabilities and the upper and lower confidence intervals.	gen LByhat1=yhat1-invttail(e(df_m),.05)*seyhat1 gen UByhat1=yhat1+invttail(e(df_m),.05)*seyhat1 twoway connected yhat1 LByhat1 UByhat1 xh

XI. FIGURES.

Figure 1. Marginal Effect of *Runoff* with 90% Confidence Interval.

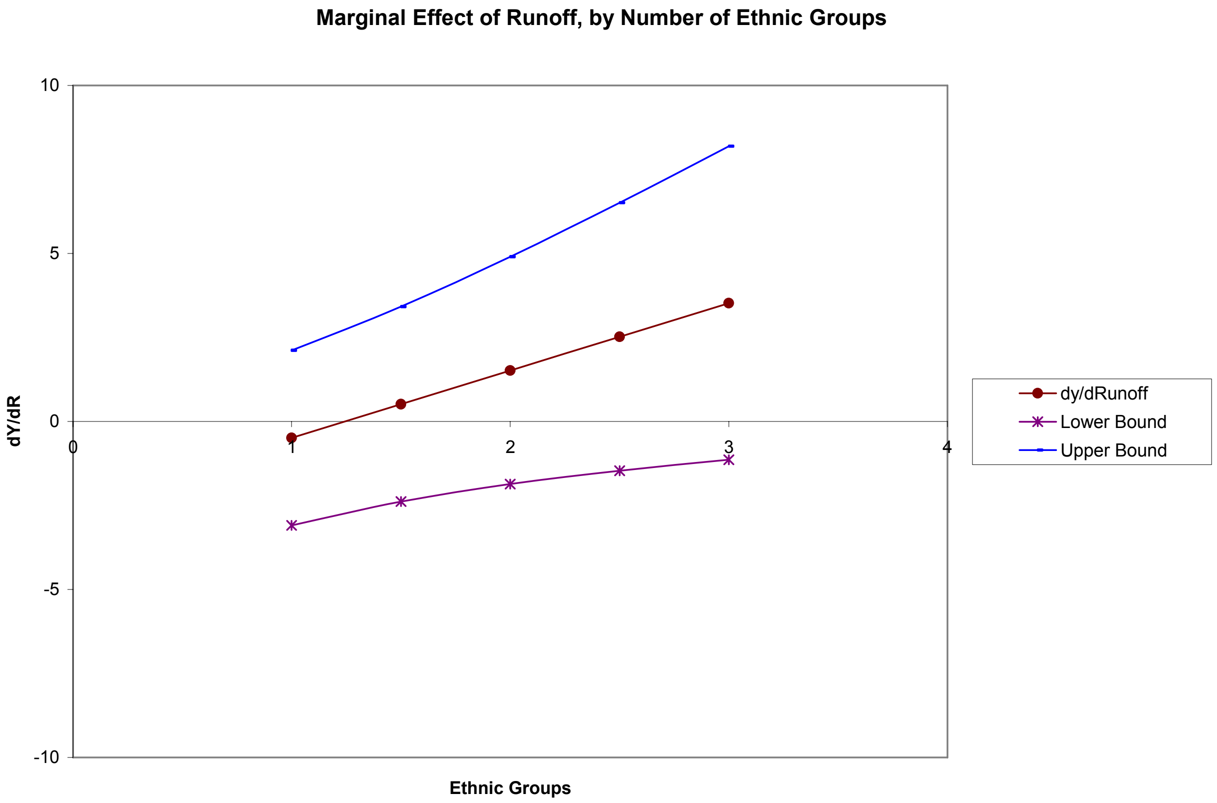


Figure 2. Marginal Effect of *Runoff*, Extending the Range of *Ethnic Groups*.

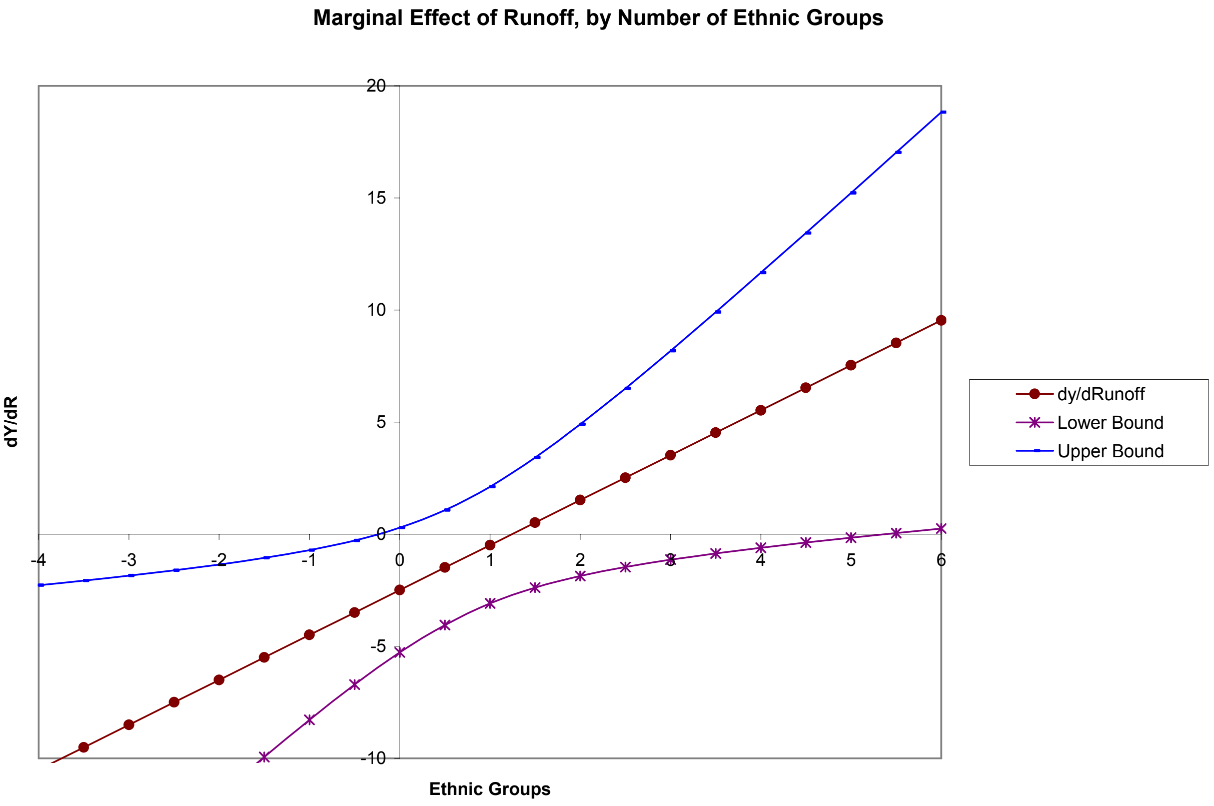


Figure 3. Predicted Number of Candidates, when *Runoff*=0.

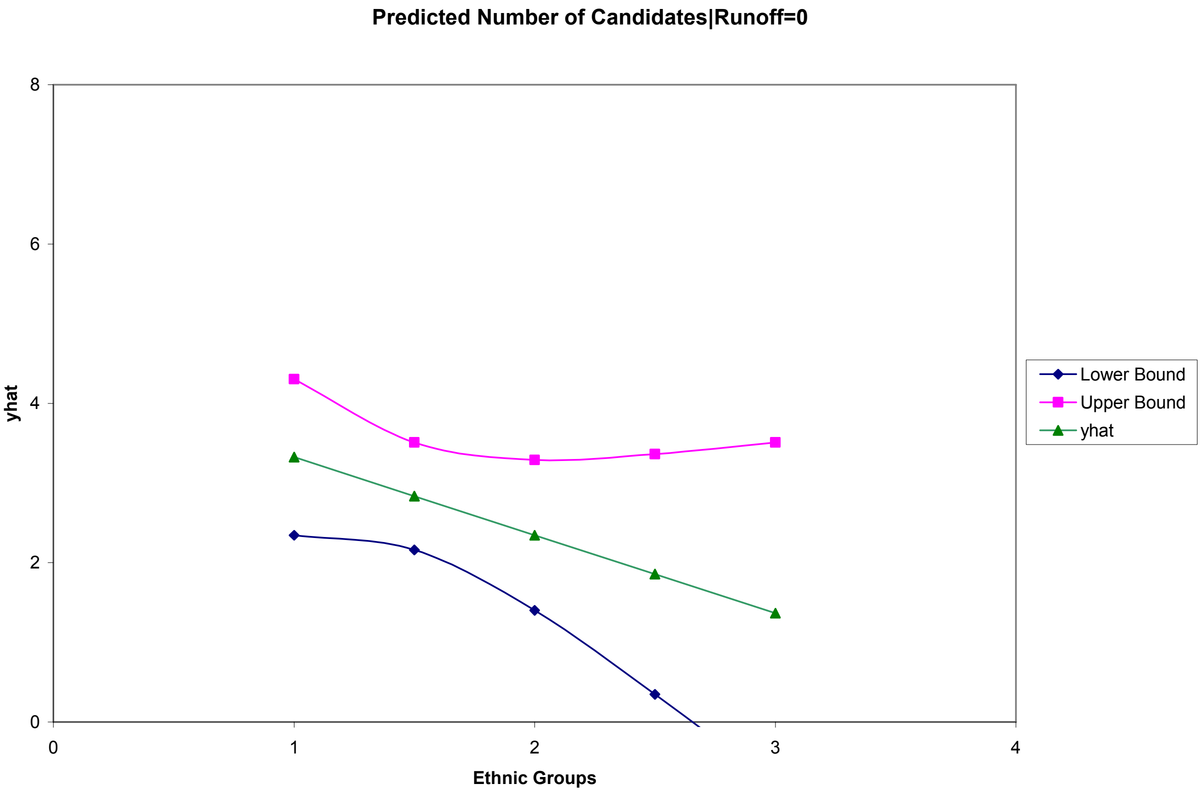


Figure 4. Predicted Number of Candidates, when *Runoff*=1.

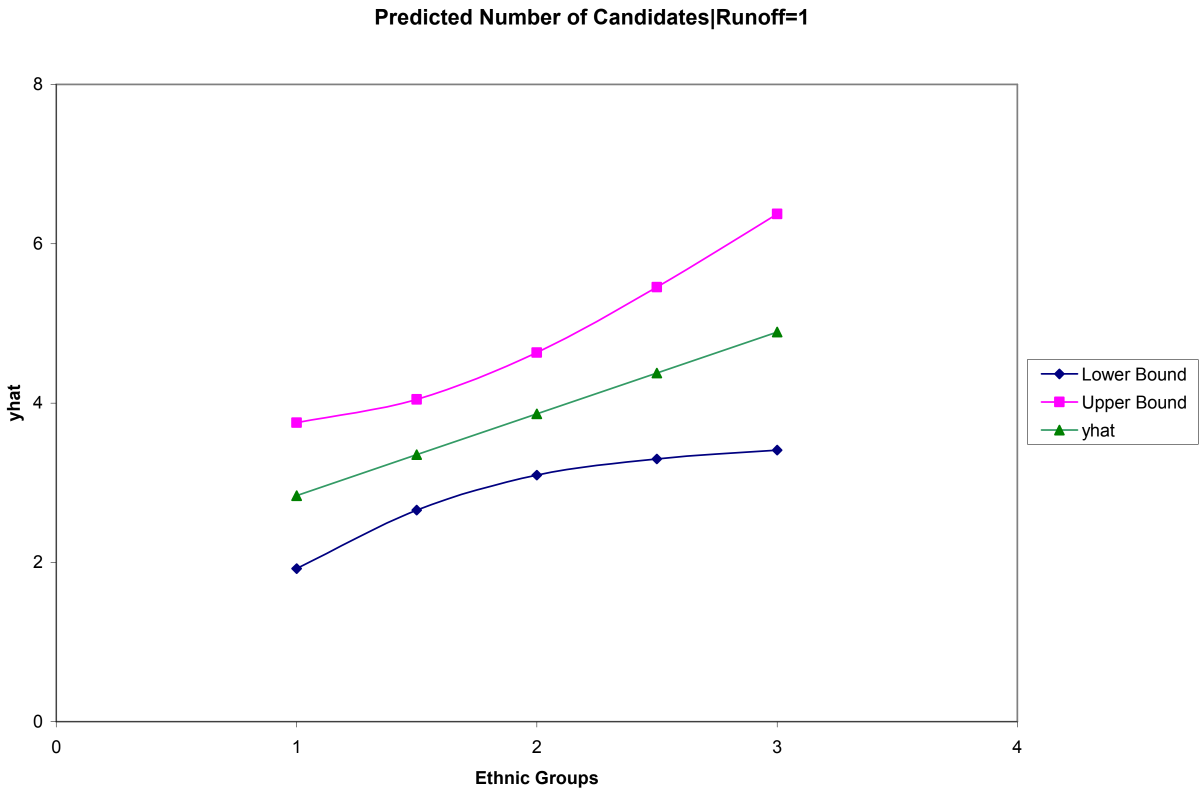


Figure 5. Predicted Number of Candidates.

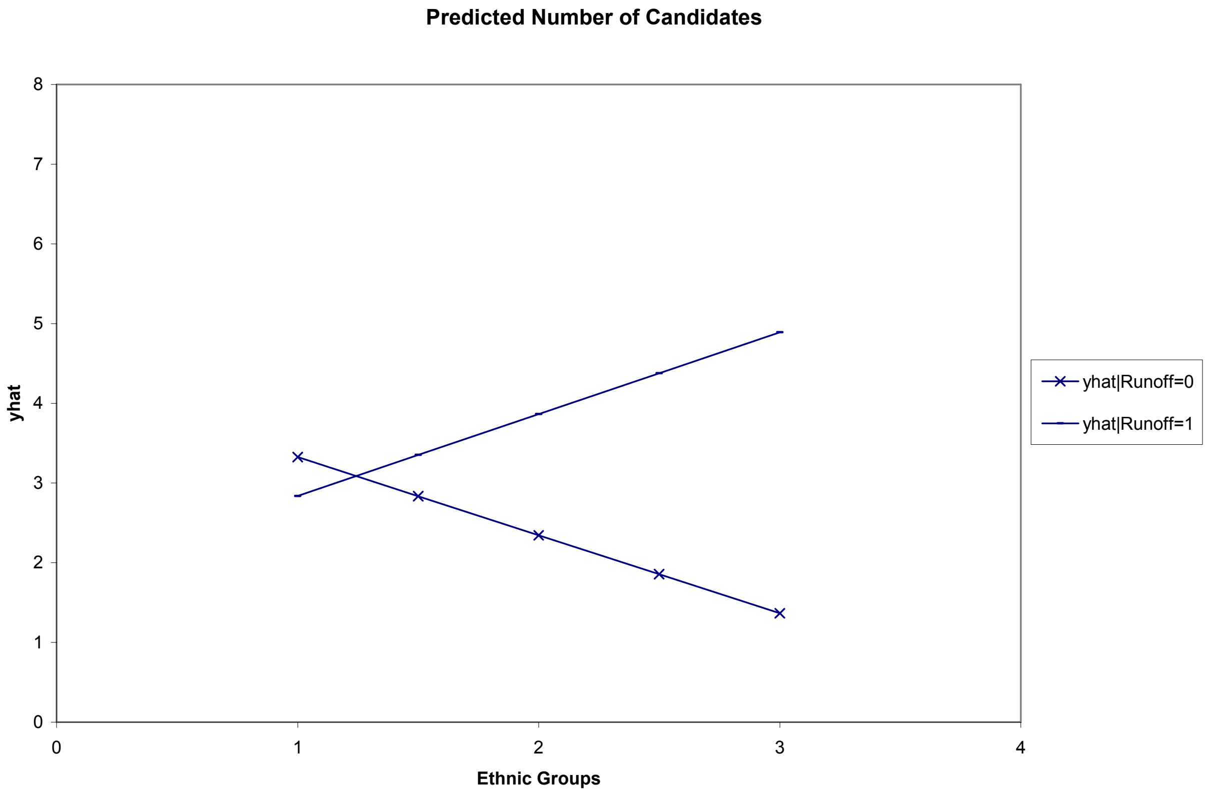


Figure 6. Predicted Number of Candidates, with 90% Confidence Intervals.

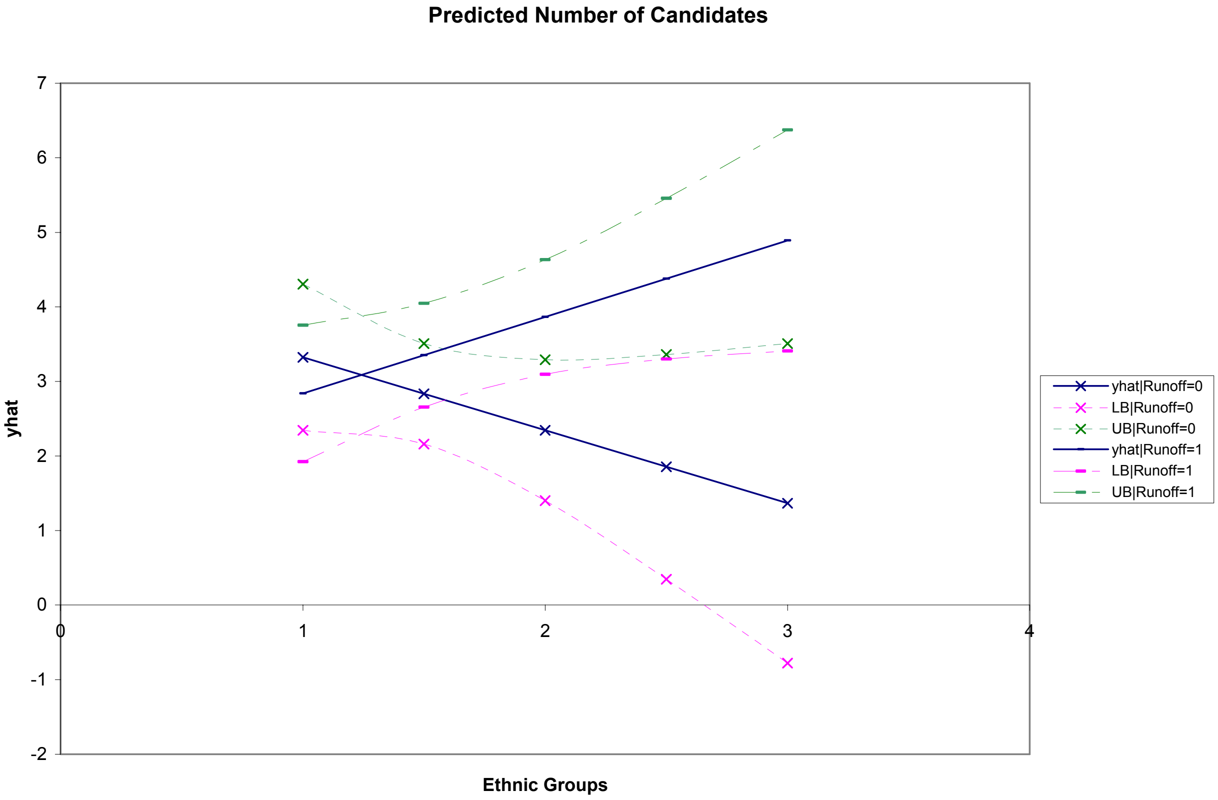
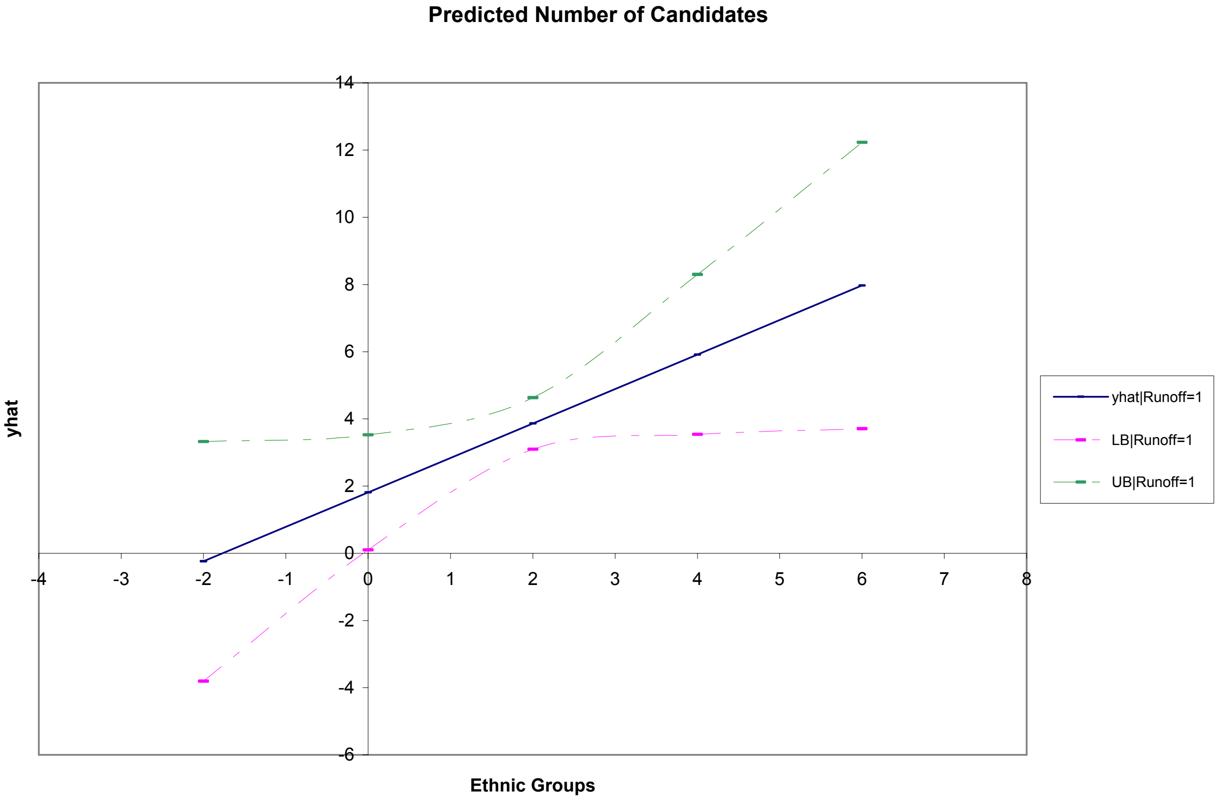


Figure 7. Predicted Number of Candidates, Extending the Range of *Ethnic Groups*.



Predicted Number of Candidates
yhat
Ethnic Groups
14
12
10
8
6
4
2
0
-2
-4
-6
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
yhat|Runoff=1
LB|Runoff=1
UB|Runoff=1