TECHNICAL WORKING PAPER SERIES

INTERACTION EFFECTS AND DIFFERENCE-IN-DIFFERENCE ESTIMATION IN LOGLINEAR MODELS

John Mullahy

Technical Working Paper 245
http://www.nber.org/papers/T0245

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
November 1999

Thanks are owed to Josh Angrist, Jon Gruber, Don Kenkel, Gene Laska, Will Manning, Mike Morrisey, Edward Norton, Jon Skinner, Doug Staiger, and Aaron Stinnett for insightful comments and suggestions on earlier drafts. The views expressed herein are those of the authors and not necessarily those of the National Bureau of Economic Research.

Interaction Effects and Difference-In-Difference Estimation
in Loglinear Models
John Mullahy
NBER Technical Working Paper No. 245
November 1999
JEL No. C2, I1

ABSTRACT

In applied econometric work, analysts are concerned often with estimation of and inferences about interaction effects, e.g. "Does the magnitude of the effect of z_1 on y depend on z_2?" This paper develops tests for and proper interpretation of various forms of interaction effects in one prominent class of regression models – loglinear models – for which the nature of estimated interaction effects has not always been given due attention. The results obtained here have a direct bearing on the interpretation of so-called difference-in-difference estimates when these are obtained using loglinear models. An empirical example of the impacts of health insurance and chronic illness on prescription drug utilization underscores the importance of these issues in practical settings.

John Mullahy
Departments of Preventive Medicine and Economics
University of Wisconsin - Madison
787 WARF, 610 Walnut Street
Madison, WI 53705
and NBER
jmullahy@facstaff.wisc.edu

I. Introduction

In applied work, analysts are concerned often with estimation of and inferences about interaction effects, e.g. "Does the magnitude of the effect of z_1 on y depend on z_2?".[1] Such concerns are operationalized most commonly in terms of some particular second-order properties of the conditional mean of outcomes (y) given exogenous covariates ($\mathbf{z}$), $E[y|\mathbf{z}]$.[2] Popular estimation strategies like difference-in-difference estimation or linear index function model specifications with second-order terms in the elements of $\mathbf{z}$ typically involve considerations of such interaction effects. This paper develops tests for and proper interpretation of various forms of interaction effects in one prominent class of regression models -- loglinear or log-link models -- for which consideration of the nature of estimated interaction effects has not always been given due attention.

It is of some importance to note at the outset that intellectual or policy-related questions pertaining to "interaction effects" -- whether structured in terms of

---

[1] For instance, Wedig, 1988, considers the extent to which the response of health service utilization to price depends on individuals' health status.

[2] Of course, the focus on properties of $E[y|\mathbf{z}]$ is by no means necessary. For instance, Josh Angrist has suggested to me in an email that "Most labor economists seem to think directly in terms of" $E[\ln(y)|\mathbf{z}]$ for "wages or hours worked."

quantities like $\frac{\partial^2 E[y|\mathbf{z}]}{\partial z_1 \partial z_2}$ or $\frac{\Delta^2 E[y|\mathbf{z}]}{\Delta z_1 \Delta z_2}$ as considered here, or in terms of other second-order properties of other features of the conditional distribution $\phi(y|\mathbf{z})$ that may be relevant to particular circumstances -- are *logically* distinct from issues pertaining to the specification of functional forms of quantities like $E[y|\mathbf{z}]$. How to *infer* the presence and/or magnitude of such interaction effects given particular functional form assumptions is part of the task at hand. As such it is helpful to bear in mind that the "policy question" issue and the "functional form" issue are in fact *logically* separate yet ultimately *analytically* intertwined.

The plan for the paper is as follows. Section II lays out the baseline modeling assumptions. Section III describes some alternative characterizations of "interaction effects" and presents a set of test statistics for undertaking inference about such effects. Section IV discusses the implications of *a priori* restrictions on "interaction parameters." Section V sketches differences between interpretations of interaction *parameters* and intereaction *effects*. Sections VI and VII illustrate these issues in the contexts of synergy/antagonism and of quantile regression analysis, respectively. Section VIII presents an empirical illustration. Section IX concludes.

II. Loglinear Models

Given scalar outcomes $y>0$ or $y\geq 0$, analysts work frequently with a variety of loglinear econometric specifications, i.e. specifications where

$$\ln(E[y|\mathbf{z}]) = \mathbf{z}\beta$$

i.e.

$$E[y|\mathbf{z}] = \exp(\mathbf{z}\beta).$$

This specification encompasses log-link models in the GLM framework, exponential conditional mean (ECM) models, etc. Estimation is undertaken via suitable loglinear or nonlinear regression methods, accounting properly for the fact that $E[\ln(y)|\mathbf{z}]$ is not necessarily proportional with respect to $\mathbf{z}$ to $\ln(E[y|\mathbf{z}])$[3] (see Manning, 1998, or Mullahy, 1998).

[3] That is, issues properly involving retransformation must be accommodated in practice, but will be ignored in what follows. Their consideration here cannot simplify the results presented below, but may complicate them. In addition, several readers of earlier drafts have emphasized that the basic arguments advanced here apply *mutatis mutandis* to other nonlinear link function ("transformation") models as well; the results are exposited here in the context of the log-link specification because of the dominance of its use in practice.

III. Characterizing Interaction Effects

Let $\mathbf{z}$ be partitioned as $\mathbf{z}=[z_1, z_2, z_1 \times z_2, \mathbf{z}_a]$, where z_1 and z_2 are scalars whose possible "interaction" is of primary interest in the analysis, and let β be partitioned conformably as $\beta=[\beta_1, \beta_2, \beta_{12}, \beta_a]$. $E[y|\mathbf{z}]$ thus has the exponential conditional mean (ECM) or log-link specification,

$$E[y|\mathbf{z}] = \exp(\beta_1 z_1 + \beta_2 z_2 + \beta_{12} z_1 \times z_2 + \mathbf{z}_a \beta_a),$$

i.e.

$$\ln(E[y|\mathbf{z}]) = \beta_1 z_1 + \beta_2 z_2 + \beta_{12} z_1 \times z_2 + \mathbf{z}_a \beta_a.$$

It might be tempting to infer something about the presence or absence of interaction effects via point estimates $\hat{\beta}_{12}$ and their estimated standard errors, possibly -- though neither necessarily nor necessarily appropriately -- obtained via a linear regression of $\ln(y)$ on $\mathbf{z}$. It may turn out, however, that such an inference is not proper to draw, a point underscored by means of the following particular cases encountered commonly in applications.

Case 1: Continuous z_1 and z_2

In the case where z_1 and z_2 are continuous, a common conceptualization of an interaction effect is a nonzero value of the cross-partial derivative

$$\frac{\partial^2 E[y|\mathbf{z}]}{\partial z_1 \partial z_2} = E[y|\mathbf{z}] \times \left\{ \frac{\partial^2 \ln E[y|\mathbf{z}]}{\partial z_1 \partial z_2} + \frac{\partial \ln E[y|\mathbf{z}]}{\partial z_1} \frac{\partial \ln E[y|\mathbf{z}]}{\partial z_2} \right\}$$

$$= E[y|\mathbf{z}] \times \{\beta_{12}[1 + (\beta_1 z_1 + \beta_2 z_2 + \beta_{12} z_1 \times z_2)] + \beta_1 \beta_2\}.$$

As such, a proper test for and interpretation of this particular form of interaction effect involves particular values of z_1 and z_2. At such prespecified values, a test of the nonlinear-in-parameters null hypothesis

$$H_{01}: \beta_{12}[1 + (\beta_1 z_1 + \beta_2 z_2 + \beta_{12} z_1 \times z_2)] + \beta_1 \beta_2 = 0$$

is a proper test for the interaction effect of interest.

Importantly, note that even if $\beta_{12}=0$, a nonzero value of the product $\beta_1 \beta_2$ would be indicative of a nonzero interaction effect -- thus characterized -- at any values of z_1 and z_2.

Case 2: Dummy Variable z_1, z_2 ("Difference-in-Difference" Estimation)

When z_1 and z_2 are zero-one dummies, the obvious analog to the cross-partial derivative is the difference-in-difference ("DID"),

$$\frac{\Delta^2 E[y|\,\mathbf{z}]}{\Delta z_1 \Delta z_2} = \exp(\mathbf{z}_a\beta_a) \times [\exp(\beta_1+\beta_2+\beta_{12}) - \exp(\beta_1) - \exp(\beta_2) + 1]$$

$$= \exp(\mathbf{z}_a\beta_a) \times T(\beta_1,\beta_2,\beta_{12}),$$

where T(.) is shorthand for the term in square brackets. In this case, a proper test for the absence an interaction effect (i.e. the null is a zero difference-in-difference) would be a test of the nonlinear-in-parameters null hypothesis

$$H_{02}: T(\beta_1,\beta_2,\beta_{12}) = 0.$$

Again, even if $\beta_{12}=0$ an interaction effect characterized as a difference-in-difference is generally nonzero.

One prominent application of the DID method is in situations where z_1 represents a discrete time measure (0="before"; 1="after") and z_2 represents a discrete treatment-status measure (0="control"; 1="treated"). In some such applications, the treatment effect net of any secular trends captured by z_1 may be estimated by some quantity like the difference in the difference of the (z_1, z_2)-subsample y-means,

$$\left(\overline{\ln(y_{11})} - \overline{\ln(y_{10})}\right) - \left(\overline{\ln(y_{01})} - \overline{\ln(y_{00})}\right),$$

where the ij subscripts denote the respective values of z_2 and z_1. If ln(y) is specified to have a linear regression structure, then this expression would tend in probability to

$$[(\beta_1 + \beta_2 + \beta_{12} + E\mathbf{z}_{a1}\beta_a) - (\beta_2 + E\mathbf{z}_{a1}\beta_a)] - [(\beta_1 + E\mathbf{z}_{a0}\beta_a) - E\mathbf{z}_{a0}\beta_a] = \beta_{12},$$

where $\mathbf{z}_{a1}$ and $\mathbf{z}_{a0}$ represent the treatment and control $\mathbf{z}_a$ variables, respectively. Yet as shown below in section V this DID will not generally capture the quantity that is probably of interest, viz $\frac{\Delta^2 E[y|\,\mathbf{z}]}{\Delta z_1 \Delta z_2}$.

Case 3: Dummy Variable z_1 and z_2

In this dummy variable case, the effect of primary interest may not be the difference-in-difference *per se*, but rather may involve whether for a particular value of one of the dummy variables (say z_1=1, without loss of generality), there is a difference in the conditional mean of y between the z_2=0 and z_2=1 subpopulations. Interaction is interpreted here as a conditional *first* difference, or as a *treatment effect among the treated* in the terminology of the treatment effect literature (Heckman, 1997). For example, in the empirical analysis undertaken below in section VIII, a main concern is whether for a subpopulation of chronically ill individuals (CHRONIC≡z_1=1) there is a difference in the conditionally expected utilization of prescription medicines between uninsured (INSURED≡z_2=0) and insured (z_2=1) subpopulations.

In this instance, the difference of interest is

$$\frac{\Delta E[y|\, z_1 = 1, \mathbf{z}_a]}{\Delta z_2} = E[y|\, z_1=1, z_2=1, \mathbf{z}_a] - E[y|\, z_1=1, z_2=0, \mathbf{z}_a]$$

$$= \exp(\mathbf{z}_a\beta_a) \times [\exp(\beta_1 + \beta_2 + \beta_{12}) - \exp(\beta_1)],$$

and a proper test for the absence of such an interaction effect would be a test of the linear-in-parameters null hypothesis

$$H_{03}: \beta_2 + \beta_{12} = 0.$$

Once again, β_{12} is only partially informative about the presence or absence of interaction effects thus defined, as opposed to the linear model situation where testing $H_0:\beta_{12}=0$ would be the proper test of the null hypothesis.

In all three cases, the main point is that $\beta_{12}=0$ or, perhaps more practically, failure to reject $H_0:\beta_{12}=0$ does *not* provide sufficient grounds for concluding that second-order interaction effects when characterized as differences-in-differences are zero or insignificant, power considerations notwithstanding. Even when $\beta_{12}=0$, second-order interaction effects will generally be nonzero in the loglinear model so long as either or both of the respective first-order or main parameters are themselves important. Other cases -- z_1 continuous and z_2 dummy; quadratics in continuous z_1 and/or z_2 -- can be imagined, but the basic idea will still apply.

IV. *A Priori* Restrictions on β_{12}

No less important to note is that loglinear model specifications that restrict *a priori* $\beta_{12}=0$ are implicitly imposing a sign structure on such interaction effects.

Specifically, when $\beta_{12}=0$, the hypothesized interaction effect will be positive if $\text{sgn}(\beta_1)=\text{sgn}(\beta_2)$ and will be negative if $\text{sgn}(\beta_1)=-\text{sgn}(\beta_2)$ for cases 1 and 2, and will be equal to $\text{sgn}(\beta_2)$ for case 3.

V. Interpreting β_{12} and Interpreting "Interaction Effects"

What "$\beta_{12}=0$" *does* imply in the dummy-dummy case (case 2) is that the "ratio of ratios" ("ROR"),

$$\frac{\left\{\frac{E[y|z_1=1,z_2=1,\mathbf{z}_a]}{E[y|z_1=0,z_2=1,\mathbf{z}_a]}\right\}}{\left\{\frac{E[y|z_1=1,z_2=0,\mathbf{z}_a]}{E[y|z_1=0,z_2=0,\mathbf{z}_a]}\right\}} = \exp(\beta_{12}),$$

equals one. When all E[.] are positive, one could thus take an ROR of one to indicate "no interaction," an ROR$\in(0,1)$ to

indicate a "negative interaction," and an ROR>1 to indicate a "positive interaction."

While conceptually kindred to the difference-in-difference characterization of the interaction effect, this ratio-of-ratios measure is clearly conveying a different sense of "interaction" than does the difference-in-difference. Specifically, it is evident that one can find the presence of interaction effects in one characterization while at the same time finding their absence in the other.

Consider for concreteness the following three scenarios. In each table the southeast 2×2 matrix contains the expectations at the indicated values of the binary covariates indicated on the top and left margins:

Scenario 1: $\beta_a=0$, $\beta_1=\ln 2$, $\beta_2=\ln 3$, $\beta_{12}=0$	$z_1=0$	$z_1=1$
$z_2=0$	1	2
$z_2=1$	3	6

Scenario 2: $\beta_a=0$, $\beta_1=\ln 2$, $\beta_2=\ln 3$, $\beta_{12}=\ln(2/3)$	$z_1=0$	$z_1=1$
$z_2=0$	1	2
$z_2=1$	3	4

Scenario 3: $\beta_a=0$, $\beta_1=\ln 5$, $\beta_2=\ln 3$, $\beta_{12}=\ln(2/3)$	$z_1=0$	$z_1=1$
$z_2=0$	1	5
$z_2=1$	3	10

The key insight here is that the DID and ROR characterizations of "interaction" need not lead to the same conclusion about either the presence or the sign of an interaction effect. In Scenario 1, ROR=1 ("no interaction") but DID=2 ("positive interaction"); in Scenario 2, ROR=2/3 ("negative interaction") but DID=0 ("no interaction"); in Scenario 3, ROR=2/3 ("negative interaction") while DID=3 ("positive interaction").

As a general matter obtaining sensible empirical answers requires specifying well-structured questions; the case of what is meant by an "interaction effect" is clearly no different in this regard.

VI. Synergy and Antagonism

Synergy and antagonism are in an important sense the more general statistical counterparts of the economic concepts of substitutes and complements, respectively (Laska et al., 1997). In the context of the previous discussion with z_1 and z_2 measured

as dummy variables, and assuming that $\operatorname{sgn}\left\{\frac{\Delta E[y|\mathbf{z}]}{\Delta z_1}\right\} = \operatorname{sgn}\left\{\frac{\Delta E[y|\mathbf{z}]}{\Delta z_2}\right\}$, synergy is the case where

$$\operatorname{sgn}\{E[y|\mathbf{z}_a, z_1=1, z_2=1] - E[y|\mathbf{z}_a, z_i=1, z_j=0]\} = \operatorname{sgn}\left\{\frac{\Delta E[y|\mathbf{z}]}{\Delta z_i}\right\},$$

for $i,j \in \{1,2\}$, whereas antagonism replaces $\operatorname{sgn}\left\{\frac{\Delta E[y|\mathbf{z}]}{\Delta z_i}\right\}$ with $-\operatorname{sgn}\left\{\frac{\Delta E[y|\mathbf{z}]}{\Delta z_i}\right\}$ on the rhs. That is, synergy (resp. antagonism) is the case where z_1 and z_2 acting in concert produce a result (measured in $E[y|\mathbf{z}]$ units) of greater (resp. smaller) magnitude than that obtained by either z_1 or z_2 acting alone.

In this context, it is straightforward to see that the proper test of the null hypothesis of {no synergy, no antagonism} is the test of H_{02} as described above.

VII. Loglinear Quantile Regressions

Suppose the α-th conditional quantile of $\phi(y|\mathbf{z})$ is given by

$$Q_\alpha(y|\mathbf{z}) = \exp(\mathbf{z}\zeta_\alpha).$$

Given the monotonicity property of (conditional) quantiles (Powell, 1991), this implies that the α-th conditional quantile of $\phi(\ln(y)|\mathbf{z})$ is given by

$$Q_\alpha(\ln(y)|\mathbf{z}) = \mathbf{z}\zeta_\alpha,$$

i.e. a linear conditional quantile relationship. It thus follows that a proper test of the null hypothesis of no interaction effects between z_1 and z_2 at the α-th conditional quantile of $\phi(y|\mathbf{z})$ when z_1 and z_2 are dummy variables, i.e. H_0: $\dfrac{\Delta^2 Q_\alpha(y|\mathbf{z})}{\Delta z_1 \Delta z_2} = 0$, is from a test of the null hypothesis

$$H_{04}: \exp(\zeta_{\alpha 1}+\zeta_{\alpha 2}+\zeta_{\alpha 12}) - \exp(\zeta_{\alpha 1}) - \exp(\zeta_{\alpha 2}) + 1 = 0,$$

i.e. the quantile-model analog of H_{02}. Parallel considerations would result in a quantile-model analog of the other null hypotheses described above.

VIII. An Application

Mullahy, 1999, uses a sample of 4,753 adults from Wave III, Phase 2 of the National Health and Nutrition Examination Survey (NHANES-III) to estimate a set of conditional mean models of the demand for prescription drugs. The dependent variable of interest is the number of prescription drugs reported used in the month prior to the survey, ranging from 0 to 15 (NMEDS, sample mean 0.81). One main concern in this study is the extent to which health insurance coverage (INSURED, dummy variable, sample mean 0.77) influences the demand for prescription drugs, with particular concern about the extent to which insurance affects the utilization of individuals suffering from chronic health problems (CHRONIC, dummy variable, sample mean 0.59).

In terms of the preceding discussion, these are tantamount to concerns about interaction effects between INSURED and CHRONIC. Two estimators are used (in models that also control for other relevant exogenous covariates): a simple linear regression model; and an ECM/log-link model estimated by weighted GLM method described in Mullahy, 1999.

Table 1 summarizes the results for the main and interaction effects of these covariates. For the linear regression model, the estimated interaction effect is positive and statistically significantly different from zero at conventional levels. For

the ECM formulation, the parameters corresponding to the main effects are also significant by conventional standards, but the β_{12} parameter point estimate is actually negative albeit with a small asymptotic t-statistic.

Based on the ECM results (a model that Mullahy, 1998, argues to be more suitable for the data than the linear model), would it be proper to conclude that there are no interaction effects between insurance status and chronic disease status? Tests of the null hypothesis H_{02} as described above suggest otherwise. In table 1, this test statistic (a null-$\chi^2_{(1)}$ test statistic, computed here using Stata's *testnl* procedure) has a realization of 5.32 (p=.021), strongly recommending rejection of the null hypothesis of no interaction effect. This interaction inference is consonant with the corresponding standard linear model inferences based on $\hat{\beta}_{12}$ drawn from the results in column 1.

IX. Conclusions

Testing for interaction effects in loglinear regression models is straightforward, but entails more than simply the consideration of the parameters associated with interaction terms. The importance of a proper specification of the null hypothesis of no interaction effects was demonstrated via an

application in which such interaction effects were found to be important. Ultimately, the results in this paper should serve to alert applied analysts to one class of particular complications involved in using log-transformed models, much as has the work of Halvorsen and Palmquist, 1980, Manning, 1998, and Mullahy, 1998, in other contexts. At a minimum, the results here suggest that the strategy "Run a regression of ln(y) on **z** and test $H_0:\beta_{12}=0$" will generally not be informative about the interaction effect that probably concerns the analyst.

References

Halvorsen, R. and R. Palmquist. 1980. "The Interpretation of Dummy Variables in Semilogarithmic Equations." *American Economic Review* 70: 474-475.

Heckman, J.J. 1997. "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." *Journal of Human Resources* 32: 441-462.

Laska, E.M. et al. 1997. "Classification of the Effectiveness of Combination Treatments." *Statistics in Medicine* 16: 2211-2228.

Manning, W.G. 1998. "The Logged Dependent Variable, Heteroskedasticity, and the Retransformation Problem." *Journal of Health Economics* 17: 283-295.

Mullahy, J. 1998. "Much Ado about Two: Reconsidering Retransformation and the Two-part Model in Health Econometrics." *Journal of Health Economics* 17: 247-281.

Mullahy, J. 1999. "A Microeconometric Analysis of the Demand for Prescription Drugs." Mimeo.

Powell, J.L. 1991. "Estimation of Monotonic Regression Models under Quantile Restrictions." in W.A. Barnett, J. Powell, and G. Tauchen, eds. *Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and*

Econometrics. Cambridge: Cambridge University Press.

Wedig, G.J. 1988. "Health Status and the Demand for Health: Results on Price Elasticities." *Journal of Health Economics* 7: 151-163.

Table 1

Estimation Results: Linear and Nonlinear Models of Prescription Drug Demand
(robust asymptotic t-statistics in parentheses; other point estimates suppressed for brevity)

Variable	Linear Model	ECM Model
INSURED	.019 (0.4)	.586 (3.7)
CHRONIC	.257 (3.9)	1.086 (8.5)
INSURED×CHRONIC	.391 (4.8)	-.173 (1.0)
$\chi^2_{(1)}$ test statistic for test of H_{02}		5.32 (p=.021)