Interpretation of nonlinear difference-in-differences: the role of the parallel trends assumption

Barkowski, Scott

Clemson University

June 2021

Interpretation of nonlinear difference-in-differences: the role of the parallel trends assumption

Scott Barkowski*

June 2021

Abstract

I argue interpretation of nonlinear difference-in-differences models depends on the form of the parallel trends assumption. When they are assumed in the natural scale of the dependent variable, the treatment effect is the interaction effect (a cross-difference). If they are assumed in the transformed scale, it is a single difference. I further note that assuming parallel trends in one scale implies they do not hold in the other, except in special cases. Finally, I consider log-linear (and related) difference-in-differences models and provide a constant form of the treatment effect that is comparable across applications with different parallel trends assumptions.

Keywords: Difference-in-differences; Nonlinear Models; Model Interpretation; Identification; Probit; Logit; Log-linear; Semilogarithmic.

JEL categories: C21; C23; C25.

*John E. Walker Department of Economics, Clemson University (sbarkow@clemson.edu). I declare that I have no material financial interests related to the research in this paper. I am grateful to Joanne Song McLaughlin, Jonathan Roth, and Pedro H. C. Sant'Anna for helpful comments on this draft, and Andrew Goodman-Bacon and Edward C. Norton for constructive comments on an earlier project out of which this one grew.

1 Introduction

Difference-in-differences (DD) models are ubiquitous in modern empirical research in economics and other social sciences, and their implementation in parametric, nonlinear settings – such as in logit, probit, and semilogarithmic models – has become an important tool for empirical researchers. While the use of nonlinear DD models has become widespread, I argue the literature has not provided clarity on the issue of their interpretation. In a linear context, interpretation is straightforward, since a single parameter represents the object of interest in the model. In nonlinear DD models, the object of interest involves more than one coefficient, but there is ambiguity as to what the object of interest *actually* is. Two views have been espoused in the literature. In the first, the object of interest takes the form specified in Mullahy (1999) and Ai and Norton (2003) (jointly hereafter as M&AN): a cross-difference of nonlinear conditional expectation functions. This approach is based on treating the interaction effect between the treatment and group dummies as the object of interest. An important implication of this view is that the "true" DD result of a nonlinear model could be very different from the estimate of the interaction coefficient, and need not have the same sign.[1]

The second view critiques the first view, and is voiced most strongly by Puhani (2012), with Blundell et al. (2004), Blundell and Costa Dias (2009), and Lechner (2011) (jointly hereafter as PBL) also making parts of the argument. They argue

[1]The papers by Mullahy and Ai and Norton discuss the interpretation of interaction terms in nonlinear models, and never directly claim that the interaction term should be the object of interest in a DD model; they only mention that interpretation of interaction terms is relevant for DD models. Nevertheless, the field has widely interpreted their work – and particularly Ai and Norton's paper – as providing the correct interpretation of a nonlinear DD model. See Ryan et al. (2015) for a recent example of this.

that the object of interest specified by M&AN does not represent the treatment effect. This view contends a much simpler object, consisting of a single difference of nonlinear conditional expectation functions, represents the treatment effect and should be considered the correct object of interest in nonlinear DD models. Importantly, Puhani notes that, for nonlinear models based on monotonically increasing transformations, this approach implies that the sign of the treatment effect must be the same as that of the coefficient on the interaction term, a convenient result for empirical applications.

Whereas Puhani argued that the form he specified is only appropriate manner to interpret nonlinear DD models, I propose that the two views described above are not inconsistent. Instead, I argue both potentially represent treatment effects, with the underlying difference between the two being the form in which the parallel trends assumption is made. This leads to the view that the correct interpretation of a nonlinear DD model depends on the context of the analysis and should be driven by the form of the parallel trends assumption made by the analyst.

The rationale underlying my argument can be illustrated via example. Figure 1 presents two sets of averages for an outcome. The points connected by solid lines are represented in the natural scale of the variable while those connected by dashed lines are the same values in the log scale. The lines connect points from before treatment to those after for both treated and control groups. In the natural scale, the effect of the treatment given by DD is -630, while in the log scale it is 0.25. What should an empirical researcher conclude if confronted by such an example? Is the treatment effect positive or negative? Guided by intuition, most researchers likely conclude that the right scale is the one in which the parallel trends assumption is credible (perhaps informed by evidence from pre-trends). I formalize this intuition

for nonlinear DD models, where, as in this example, dependent variables exist in both natural and transformed scales. I show that if parallel trends are assumed to hold in the natural scale of the outcome, then the treatment effect of a nonlinear DD model takes the form specified by M&AN. Alternatively, if the parallel trends are assumed in the transformed scale, then the form PBL detailed is the treatment effect of interest. Moreover, since researchers commonly use pre-trends to provide suggestive evidence towards the validity of the parallel trends assumption, they could also be use pre-trends to suggest the proper manner to interpret a nonlinear DD model.

Aside from M&AN and PBL, the closest related work to this one is that by Kahn-Lang and Lang (2020), who argue that analysts' understanding of the nature of parallel trends in a given context (potentially via theoretical knowledge) should play a role in their choice of functional form. Additionally, when a researcher choses a nonlinear-DD functional form, it implies that he or she believes the counterfactual trend is consistent with an interpretation given by PBL. My argument goes further than theirs, as I argue that *even conditional* on choosing a nonlinear functional form, a researcher still faces the choice of basing his or her interpretation on a parallel trend assumption in either the natural or transformed scales. Thus, in contrast to Kahn-Lang and Lang, I argue it is the researchers' understanding of parallel trends that determines how to interpret the model, not the functional form he or she chooses.

The rest of this paper proceeds as follows. In the next section I derive the treatment effect in the familiar case of the linear DD model to serve as a benchmark. I then discuss the nonlinear case, deriving the treatment effect under both forms of the parallel trends assumption: the natural scale first followed by the transformed

scale. I also note that, except in special cases, when one assumes one form of parallel trends, he or she implicitly rules out the other form. This implies practitioners should take care not to inadvertently use both methods of interpretation in the same analysis (which could happen when mixing model types). Following this, I also consider the special case of log-linear DD models, noting there are multiple ways to represent the treatment effect in percentage form in these models. I suggest one that is constant whether assuming parallel trends in the natural or log-scales. Standard use of this form would make for more convenient comparison of estimates across projects. Finally, I discuss implications and conclude.

2 Linear model difference-in-differences

I introduce notation and important concepts via the familiar linear DD model,

$$y = \alpha G + \beta T + \gamma GT + \theta' X + u, \tag{2.1}$$

where y is an observed outcome of interest, T is a dummy indicating periods (0 for the first, 1 for the second), G indicates treatment (value 1) or control group (value 0), TG is the interaction of these, X is a vector of controls and a constant, and u is an error term with zero conditional mean. Additionally, though it is not necessary for a linear model, since it is common in nonlinear models that are my focus later in this paper, like probit or logit, I also assume the error term is independent of X, G, and T, and, hence, homoskedastic. Observations where both $T = 1$ and $G = 1$ (and hence $TG = 1$) are those that have received the treatment of interest. It may be noted here that I use similar notation to that of Puhani.

Equation (2.1) implies four conditional expectation functions (CEFs):

$$E[y|G=1,T=1,X] = \alpha + \beta + \gamma + \theta' X, \tag{2.2}$$

$$E[y|G=1,T=0,X] = \alpha + \theta' X, \tag{2.3}$$

$$E[y|G=0,T=1,X] = \beta + \theta' X, \text{ and} \tag{2.4}$$

$$E[y|G=0,T=0,X] = \theta' X. \tag{2.5}$$

The goal of a DD research design is to estimate the average treatment effect on the treated. For convenience, throughout this article I simply refer to the "treatment effect", but in all cases mean the average treatment effect on the treated. I define this treatment effect using the potential outcomes framework as

$$E[y^1|G=1,T=1,X] - E[y^0|G=1,T=1,X], \tag{2.6}$$

where the potential outcomes are given by y^1 (treated) and y^0 (untreated – the counterfactual). For treated observations, the observed outcome *is* the treated potential outcome. That is,

$$(y^1|G=1,T=1,X) = (y|G=1,T=1,X) = \alpha + \beta + \gamma + \theta' X + u, \tag{2.7}$$

and hence,

$$E[y^1|G=1,T=1,X] = E[y|G=1,T=1,X] = \alpha + \beta + \gamma + \theta' X, \tag{2.8}$$

For treated observations, the untreated potential outcome is unobserved, so the key assumption of a DD research design – the parallel trends assumption – specifies

how y^0 is modeled when $GT = 1$. In particular, the parallel trends assumption implies that

$$\begin{aligned}
&(y^0|G = 1, T = 1, X) \\
&\quad = E[y|G = 0, T = 1, X] + (E[y|G = 1, T = 0, X] - E[y|G = 0, T = 0, X]) + u,
\end{aligned} \tag{2.9}$$

which implies

$$\begin{aligned}
&E[y^0|G = 1, T = 1, X] \\
&\quad = E[y|G = 0, T = 1, X] + (E[y|G = 1, T = 0, X] - E[y|G = 0, T = 0, X]) \\
&\quad = \alpha + \beta + \theta' X
\end{aligned} \tag{2.10}$$

(having made use of equations 2.3 to 2.5). By modeling the potential outcome in this fashion, I am asserting that the expected counterfactual is equal to the observed average outcome of the control group plus the average pre-period group difference in the outcome.

Equations (2.8) and (2.10) allow me to show that the treatment effect in a linear

DD model is represented by the interaction term coefficient:

$$
\begin{aligned}
E[y^1|G=1,T=1,X] - E[y^0|G=1,T=1,X] & \\
&= E[y|G=1,T=1,X] - E[y|G=0,T=1,X] \\
&\quad - (E[y|G=1,T=0,X] - E[y|G=0,T=0,X]) \\
&= \alpha + \beta + \gamma + \theta' X - \alpha - \beta - \theta' X \\
&= \gamma. \qquad (2.11)
\end{aligned}
$$

3 Nonlinear difference-in-differences

In the nonlinear context, the dependent variable exists in two worlds: the natural scale, given by y, and the transformed scale, given by Y. For a researcher considering applying DD in a nonlinear scenario, there are two important choices to make. The first is the choice of transformation between the two worlds, which I indicate by $g(\cdot)$, where $g(\cdot)$ is a nonlinear, monotonic function. The second is the manner of applying the parallel trends assumption. The first choice is well understood by researchers. To illustrate the second, let $PT(\cdot)$ represent a function implementing the parallel trends assumption. The key step of modeling the natural scale counterfactual outcome (y^0) in terms of actual (i.e., observed) outcomes can then be implemented in two ways:

PTA 1: $y^0 = PT(g(Y))$, or

PTA 2: $y^0 = g(PT(Y))$.

In PTA 1, the parallel trends assumption is applied in the natural scale, and the nonlinear transformation is applied to the parallel trend assumption's components.

In the next subsection, I show that this first approach leads to a treatment effect equal to the interaction effect and takes the form specified by M&AN.

In PTA 2, the parallel trends assumption is applied in the transformed scale, and the results are then re-transformed back to the natural scale by the transformation function. This is the approach followed by PBL in developing the form of the treatment effect they propose, which I detail further below. The overall point I make here, then, is that the manner in which a nonlinear DD model should be interpreted is driven by the form in which the parallel trends assumption is made. If in the natural scale, as in PTA 1 here, then the appropriate interpretation is given by M&AN; if in the transformed scale, as in PTA 2, it is the one given by PBL.[2]

3.1 Parallel trends in the natural scale

A substantial portion of nonlinear DD models arise in situations where y is a Bernoulli variable, and so I present my discussion in the context of a standard threshold crossing model (though it could be adapted to other models, most notably semilogarithmic ones). Y is taken to be latent, with

$$y = g(Y) = \mathbb{1}\{Y \geq 0\}, \tag{3.1}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function. Standard use of DD in nonlinear contexts models the transformed dependent variable, Y, in the same way the natural scale

[2]Conceptually speaking, the two modeling approaches given here also apply in the standard linear DD model. Practically speaking, though, there is no choice to be made between the two because $g(\cdot)$ is the identity function in the linear case (as noted by Athey and Imbens, 2006), so both approaches are identical in that context.

dependent variable is modeled in the linear context:

$$Y = \tilde{\alpha}G + \tilde{\beta}T + \tilde{\gamma}GT + \tilde{\theta}'X + U, \tag{3.2}$$

where U has the same assumptions as u and the parameters are marked with tildes to distinguish them from the natural scale versions. This model and the Bernoulli nature of y imply the CEF,

$$E[y|G, T, X] = \Phi(\tilde{\alpha}G + \tilde{\beta}T + \tilde{\gamma}GT + \tilde{\theta}'X), \tag{3.3}$$

where $\Phi(\cdot)$ is a monotonically increasing, nonlinear function. This can be further broken down into

$$E[y|G = 1, T = 1, X] = \Phi(\tilde{\alpha} + \tilde{\beta} + \tilde{\gamma} + \tilde{\theta}'X), \tag{3.4}$$

$$E[y|G = 1, T = 0, X] = \Phi(\tilde{\alpha} + \tilde{\theta}'X), \tag{3.5}$$

$$E[y|G = 0, T = 1, X] = \Phi(\tilde{\beta} + \tilde{\theta}'X), \text{ and} \tag{3.6}$$

$$E[y|G = 0, T = 0, X] = \Phi(\tilde{\theta}'X). \tag{3.7}$$

It is typical in applications of threshold crossing models that $\Phi(\cdot)$ is the conditional distribution function corresponding to the distribution assumed for the error-term. However, in principle an analyst could merely assume that $\Phi(\cdot)$ is the link function between the parameters in the transformed scale and $E[y|G, T, X]$ without making an explicit assumption about the distribution of the error term. Additionally, one could notice here that $\Phi(\cdot) \neq g(\cdot)$ in standard models like probit and logit, but they could be the same in other applications – and in fact they are the same in

the semilogarithmic model (which I return to, below).

As Puhani noted, the treatment effect of interest in the nonlinear case is the same as in the linear model case, and is given by expression (2.6). As argued above, the treated potential outcome is observed when $GT = 1$, so

$$E[y^1|G = 1, T = 1, X] = E[y|G = 1, T = 1, X]. \tag{3.8}$$

I model the counterfactual outcome when $GT = 1$ by first applying PTA 1; that is, the parallel trends assumption in the natural scale. This takes the form given in equation (2.9). Taking the expectation (where the error term is still assumed to have a zero conditional mean) shows this choice leads to the following CEF for the counterfactual (the same as in the linear DD case):

$$\begin{aligned} &E[y^0|G = 1, T = 1, X] \\ &= E[y|G = 0, T = 1, X] + (E[y|G = 1, T = 0, X] - E[y|G = 0, T = 0, X]). \end{aligned} \tag{3.9}$$

Combining equations (3.8) and (3.9), the treatment effect in the nonlinear DD model is given by

$$\begin{aligned} &E[y^1|G = 1, T = 1, X] - E[y^0|G = 1, T = 1, X] \\ &\qquad = E[y|G = 1, T = 1, X] - E[y|G = 0, T = 1, X] \\ &\qquad\qquad - (E[y|G = 1, T = 0, X] - E[y|G = 0, T = 0, X]). \end{aligned} \tag{3.10}$$

Applying the nonlinear transformation (via substitution of equations 3.4 to 3.7),

this becomes

$$E[y^1|G=1,T=1,X] - E[y^0|G=1,T=1,X]$$
$$= \Phi(\tilde{\alpha}+\tilde{\beta}+\tilde{\gamma}+\tilde{\theta}'X) - \Phi(\tilde{\beta}+\tilde{\theta}'X) - [\Phi(\tilde{\alpha}+\tilde{\theta}'X) - \Phi(\tilde{\theta}'X)]. \quad (3.11)$$

So by assuming that parallel trends hold in the natural scale, I obtain a treatment effect in equation (3.11) equal to the interaction term specified by M&AN. As they note, this treatment effect is not constant and can be quite different from $\tilde{\gamma}$, including having the opposite sign.[3]

Several authors have noted that applying the parallel trends assumption in the natural scale as in equation (3.9) has a potential drawback: there is nothing restricting the conditional expectation of the potential outcome in equation (3.9) (Athey and Imbens, 2006; Blundell and Costa Dias, 2009; Lechner, 2011). In models of limited dependent variables, this could result in predictions for the counterfactual outcome that are outside the theoretical limits of the dependent variable (such as zero and one for probit or logit models). Thus, for limited dependent variables, the parallel trends assumption cannot hold in all contexts.[4] Nevertheless, it is clearly the case that it *could* be a credible assumption in a large share of empirical applica-

[3]Lechner (2011, p. 199) discusses the potential use of an estimator in the form of equation (3.11) without providing a justification for what drives the form or when one should use it. However, he also writes (on p. 198), "Thus, we conclude that estimating a DiD model with the standard specification of a nonlinear model would usually lead to an inconsistent estimator if the standard common trend assumption is upheld."

[4]This is only true to the extent that we believe the counterfactual outcome must obey the theoretical bounds of the dependent variable. Since it is never observed, this restriction is an assumption which cannot be evaluated. Further discussion of this issue would, perhaps, be quite philosophical and outside the scope of this article.

tions, to which a voluminous literature of such cases attests.[5] Moreover, the cases in which the parallel trends assumption is credible in the natural scale are precisely those when infeasible predictions for the counterfactual outcomes are least likely to be a problem, and if the bounds of the variable are unlikely to be respected in an application, the parallel trends assumption is also unlikely to be credible. Thus, the issue of whether the bounds of the dependent variable are respected in a particular application is part of evaluating whether parallel trends are credible. If they are, then researchers should not be deterred from using this method out of concern for infeasible predictions.

3.2 Parallel trends in the transformed scale

To demonstrate the second nonlinear modeling method, PTA 2, it is helpful to specify potential outcomes in the transformed scale, given by Y^0 and Y^1. These relate to the natural scale via

$$y^0 = \mathbb{1}\{Y^0 \geq 0\} \text{ and} \tag{3.12}$$

$$y^1 = \mathbb{1}\{Y^1 \geq 0\}. \tag{3.13}$$

The treated potential outcome is taken to be the same as the actual outcome in the transformed scale when $GT = 1$, just as it is in the natural scale. So,

$$(Y^1|G = 1, T = 1, X) = \tilde{\alpha} + \tilde{\beta} + \tilde{\gamma} + \tilde{\theta}'X + U, \tag{3.14}$$

[5]Most often, this literature combines DD approaches with linear probability models.

via equation (3.2). Next I apply the parallel trends assumption of equation (2.9) to Y^0. That is, in the transformed scale. This gives

$$\begin{aligned}
&(Y^0|G=1,T=1,X)\\
&= E[Y|G=0,T=1,X] + (E[Y|G=1,T=0,X] - E[Y|G=0,T=0,X]) + U\\
&= \tilde{\alpha} + \tilde{\beta} + \tilde{\theta}'X + U, \quad (3.15)
\end{aligned}$$

into which I have substituted the following CEFs implied by equation (3.2):[6]

$$E[Y|G=1,T=0,X] = \tilde{\alpha} + \tilde{\theta}'X, \tag{3.16}$$

$$E[Y|G=0,T=1,X] = \tilde{\beta} + \tilde{\theta}'X, \text{ and} \tag{3.17}$$

$$E[Y|G=0,T=0,X] = \tilde{\theta}'X. \tag{3.18}$$

Using equations (3.12) and (3.13) to re-transform the expressions for Y^1 and Y^0, given in equations (3.14) and (3.15), back to the natural scale and taking expectations gives

$$E[y^1|G=1,T=1,X] = \Phi(\tilde{\alpha} + \tilde{\beta} + \tilde{\gamma} + \tilde{\theta}'X) \text{ and} \tag{3.19}$$

$$E[y^0|G=1,T=1,X] = \Phi(\tilde{\alpha} + \tilde{\beta} + \tilde{\theta}'X). \tag{3.20}$$

The average treatment effect when assuming parallel trends in the transformed scale,

[6]It is not crucial for my argument, but the equations (3.14) and (3.15) can be used to show that the treatment effect on the latent outcome, $E[Y^1|G=1,T=1,X] - E[Y^0|G=1,T=1,X]$, is given by $\tilde{\gamma}$.

therefore, is given by

$$E[y^1|G=1,T=1,X]-E[y^0|G=1,T=1,X] = \Phi(\tilde{\alpha}+\tilde{\beta}+\tilde{\gamma}+\tilde{\theta}'X)-\Phi(\tilde{\alpha}+\tilde{\beta}+\tilde{\theta}'X), \tag{3.21}$$

which is as specified by PBL. As Puhani notes, here the sign of the treatment effect must be the same as that of $\tilde{\gamma}$, a fact implied by the monotonic and increasing nature of $\Phi(\cdot)$.

Unlike in the case of PTA 1, here the counterfactual outcome is restricted to be within the range of $\Phi(\cdot)$, so as long as $\Phi(\cdot)$ is chosen to impose the appropriate restrictions there is no possibility of the model producing infeasible predictions. But as shown in this section, this advantage is only available when parallel trends must hold in the transformed scale. Moreover, as I show next, use of PTA 2 also implies something that must be true about the natural scale version of the counterfactual. Thus, PTA 2 is only available to avoid the issue of infeasible outcomes in some contexts, and doing so does not come without a cost.

3.3 The incompatibility of the two forms of parallel trends

As the above discussion shows, the scale in which the parallel trends assumption is made implies two different models for the counterfactual outcome, y^0. Expectations implied by these are given by:

PTA 1 (Natural): $E[y^0|G=1,T=1,X] = \Phi(\tilde{\beta}+\tilde{\theta}'X) + [\Phi(\tilde{\alpha}+\tilde{\theta}'X) - \Phi(\tilde{\theta}'X)]$

PTA 2 (Transformed): $E[y^0|G=1,T=1,X] = \Phi(\tilde{\alpha}+\tilde{\beta}+\tilde{\theta}'X)$.

One of the implications of the nonlinearity of $\Phi(\cdot)$ is that these two forms for the CEF cannot be the same except in special cases. As can easily be seen,

$$\Phi(\tilde{\alpha} + \tilde{\beta} + \tilde{\theta}'X) \neq \Phi(\tilde{\beta} + \tilde{\theta}'X) + [\Phi(\tilde{\alpha} + \tilde{\theta}'X) - \Phi(\tilde{\theta}'X)], \tag{3.22}$$

unless $\tilde{\alpha} = 0$, $\tilde{\beta} = 0$, or both (Meyer, 1995; Ai and Norton, 2003; Angrist and Pischke, 2009; Lechner, 2011; Puhani, 2012; Roth and Sant'Anna, 2021).[7] Said another way, the parallel trends assumption cannot hold in both the natural and transformed scales of the dependent variable unless the treatment group in the pre-period is the same on average as the control group, or there is no change over time in the control group average, or both.

Since a parallel trends assumption is a linear function, it is quite intuitive that it would not hold in general under nonlinear transformation. Nevertheless, this result implies an important, practical trade-off for researchers: by assuming parallel trends in one scale, the practitioner is explicitly ruling out parallel trends in the other, save for the special cases mentioned. This cost is one that should be considered by researchers when they chose parallel trends assumption, including when they are trying to avoid the issue of infeasible outcomes.

This result also implies some pitfalls that researchers should be careful to avoid. One case that should be treated with care is use of linear probability models. One common scenario is a researcher using the linear probability model to implement a DD research design, and then performing some robustness checks with a probit or

[7] Roth and Sant'Anna (2021) also show that another special case exists where parallel trends could hold in both dimensions if a more general model is used that allows for mixture distributions for the error term. In that case, the time and group effects exist within separate nonlinear functions even when parallel trends are applied in the transformed scale.

logit model. In using a linear probability model there is no transformed scale,[8] so the researcher must be making his or her parallel trends assumption in the natural scale. When then turning to a nonlinear model, logical consistency would require that the researcher base interpretation off of M&AN's specification in equation (3.11) unless he or she is prepared to argue one of the special cases detailed above prevails so that parallel trends hold in both scales simultaneously. Conversely, if a researcher first uses a probit or logit model with the interpretation suggested by PBL in equation (3.21), thereby implying a transformed scale parallel trends assumption, turning next to a linear probability model would potentially be inconsistent for the same reason.

3.4 The special case of the exponential transformation

Next I consider the special case of the exponential transformation – that is, the log-linear model – which is used in a large number of empirical applications. In this model,

$$y = g(Y) = \exp(\tilde{\alpha}G + \tilde{\beta}T + \tilde{\gamma}GT + \tilde{\theta}'X + U) \text{ and} \tag{3.23}$$

$$E[y|G,T,X] = E[\Phi(Y)|G,T,X] = \exp(\tilde{\alpha}G + \tilde{\beta}T + \tilde{\gamma}GT + \tilde{\theta}'X)\eta, \tag{3.24}$$

where $\eta = E[\exp(U)|G,T,X] = E[\exp(U)]$.[9] As noted above, in this case both $g(\cdot)$ and $\Phi(\cdot)$ are the same function. Traditionally $\exp(U)$ is assumed to be log-normally distributed, but one could instead directly assume equation (3.24) without making an explicit assumption about the distribution of $\exp(U)$.

[8] More accurately, the natural and transformed scales are the same in a linear probability model, because $g(\cdot)$ is the identity function in that case.

[9] Recall that U is assumed to be independent, as is standard in nonlinear models.

A convenient feature of this model is that the treatment effect of the nonlinear DD model can be represented as a constant percentage, but there are multiple ways to do this. Here I show one particular approach results in a constant object of interest under either form of parallel trend assumption and that it is calculated relative to an expectation of an actual outcome, not a potential one.[10]

Under the exponential transformation, the treatment effect when assuming natural scale parallel trends is given by

$$\begin{aligned} &E[y^1|G=1,T=1,X] - E[y^0|G=1,T=1,X] \\ &= \exp(\tilde{\alpha}+\tilde{\beta}+\tilde{\gamma}+\tilde{\theta}'X)\eta - \exp(\tilde{\beta}+\tilde{\theta}'X)\eta - \exp(\tilde{\alpha}+\tilde{\theta}'X)\eta + \exp(\tilde{\theta}'X)\eta. \quad (3.25) \end{aligned}$$

Since in this model $E[y|G=0,T=0,X] = \exp(\tilde{\theta}'X)\eta$, a natural way to express this treatment effect as a percentage would be to divide both sides by $E[y|G=0,T=0,X]$. Since y is observed, the resulting percentage could be easily compared to a sample average of data for y to convert the treatment effect into its natural units. However, if the expectations across periods are very different, the percentage calculated relative to $E[y|G=0,T=0,X]$, a pre-period average, may give a misleading impression of the size of the treatment effect.

An alternative would be to use $E[y^0|G=1,T=1,X]$, which produces an easily understood expression for the treatment effect: a percentage in terms of the expected counterfactual outcome. However, since y^0 is not observed, a modeled

[10]Shang et al. (2018) propose an interpretation they call a "difference-in-semielasticity" for interaction terms in log-linear and related models. While their estimand is constant, like those I discuss, theirs does not represent a treatment effect. Their difference-in-semielasticity could be understood in a DD context as first calculating the percentage growth over time for both groups (relative to each group's pre-period level) and then taking the difference of these.

counterfactual mean would have to be used for benchmarking purposes as actual data would be unavailable. More importantly, since $E[y^0|G=1,T=1,X]$ in this case is comprised of a sum, the resulting percentage would not be constant and would involve ratios of estimators. These would make inference more challenging and would possibly create undesirable finite sample properties.

A better option would be to use $E[y^1|G=1,T=1,X]$, which produces

$$\frac{E[y^1|G=1,T=1,X]-E[y^0|G=1,T=1,X]}{E[y^1|G=1,T=1,X]} = 1-\exp(-\tilde{\alpha}-\tilde{\gamma})-\exp(-\tilde{\beta}-\tilde{\gamma})+\exp(-\tilde{\alpha}-\tilde{\beta}-\tilde{\gamma}). \quad (3.26)$$

This percentage has a natural interpretation as the treatment effect as a percentage of the observed outcome. additionally, it is calculated relative to an expectation on observable data, so an estimate of $E[y|G=1,T=1,X]$ could be used to convert the percentage to y's actual units. Conveniently, this expression does not depend on data, only estimates, which can easily be obtained using appropriate (consistent) estimators for the coefficients. Moreover, standard errors also could be easily obtained via the delta method.

So the expression above has several advantages, but in the context of my discussion in this article, the most attractive aspect of representing the treatment effect as in equation (3.26) is the above advantages are maintained when the treatment effect is calculated as specified by PBL according to the alternative parallel trend assumption of PTA 2:

$$\frac{E[y^1|G=1,T=1,X]-E[y^0|G=1,T=1,X]}{E[y^1|G=1,T=1,X]} = 1-\exp(-\tilde{\gamma}). \quad (3.27)$$

Under transformed scale parallel trends, the treatment effect percentage can also be conveniently represented relative to the counterfactual expectation via

$$\frac{E[y^1|G=1,T=1,X]-E[y^0|G=1,T=1,X]}{E[y^0|G=1,T=1,X]}=\exp(\tilde{\gamma})-1. \tag{3.28}$$

Given the difficulties involved with expressing this version under the M&AN form of the treatment effect, however, I argue the standard use of $(E[y^1|G=1,T=1,X]-E[y^0|G=1,T=1,X])/E[y^1|G=1,T=1,X]$ across researchers would offer an important advantage. Since parallel trends assumptions may vary across contexts, if researchers use a consistent form of the percentage treatment effect, it will be easier for estimates to be compared across projects, regardless of which form the parallel trends assumption is made in each analysis. In short, I argue that, for log-linear models, researchers should represent their treatment effects from DD models by equation (3.26) when parallel trends are assumed in the natural scale and equation (3.27) when assumed in the log-scale.

4 Conclusion

In this essay, I have attempted to reconcile the two views in the literature regarding interpretation of nonlinear DD models and show the important role the form of the parallel trends assumption plays. I argue for a simple rule: if parallel trends are assumed to hold in the natural scale of the dependent variable, then interpretation should be based on the form specified by M&AN; if they are assumed to hold in the transformed scale, then the form given by PBL should be used. Moreover, I note that when researchers decide the scale in which to make their parallel trends

assumption, they should be careful to keep in mind that doing so implies parallel trends *do not* hold in the other scale, except in special circumstances.

Given the above, empirical researchers should be careful about inadvertently being inconsistent with their parallel trends assumptions. This issue is particularly likely to arise when practitioners combine linear probability models with probit or other nonlinear versions of discrete choice models. Mixing these models is not necessarily a problem, though, researchers just need to make sure to interpret their different models according to a consistent parallel trend assumption. The above rule on interpretation can help them do that.

Additionally, while I caution researchers to keep in mind that parallel trends are unlikely to hold in both scales, this fact can also be turned around and looked at from a different perspective. It shows that when researchers encounter cases where the parallel trends assumption is more credible in one scale and not the other – perhaps due to evidence offered by pre-trends – this is not necessarily a problem. The key is to interpret the result in a manner consistent with the credible version of the parallel trends assumption. Said succinctly: a parallel trend assumption rarely holds in multiple scales for reasons unrelated to its validity, so parallel trends not holding under nonlinear transformation in some scenario is not itself an indictment of the use of a DD model in that context.

Furthermore, my argument also suggests researchers have more freedom in choosing models than would be suggested by PBL and Kahn-Lang and Lang (2020). Since the method of interpretation is tied to the form of parallel trends assumed, not the model functional form chosen, if a researcher believes parallel trends exist in the natural scale of a variable, my argument implies he or she could still chose a nonlinear-DD model. Further, the researcher is not required to assume parallel

trends exist in the transformed scale when choosing a nonlinear model. Both cases are possible as long as the analyst uses the proper method of interpretation. Thus, one could use a probit model for a binary outcome exhibiting parallel trends in the natural scale or a Poisson model for count data without assuming the parallel trend has to hold in the log-scale.

Lastly, I also note that, in the special case of the semilogarithmic models, treatment effects can be represented in convenient fashion, with the form offered by $(E[y^1|G=1,T=1,X]-E[y^0|G=1,T=1,X])/E[y^1|G=1,T=1,X]$ offering several advantages that accrue both to the individual researcher and the broader scientific community. This suggests standard use of this form of the treatment effect among researchers making use of the log-linear DD model.

References

Ai, Chunrong and Edward C. Norton, "Interaction Terms in Logit and Probit Models," *Economics Letters*, July 2003, *80* (1), 123–129. 2, 16

Angrist, Joshua D. and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, January 2009. 16

Athey, Susan and Guido W Imbens, "Identification and Inference in Nonlinear Difference-in-Differences Models," *Econometrica*, March 2006, *74* (2), 431–497. 9, 12

Blundell, Richard and Monica Costa Dias, "Alternative Approaches to Evaluation in Empirical Microeconomics," *Journal of Human Resources*, 2009, *44* (3), 565–640. 2, 12

_ , _ , Costas Meghir, and John van Reenen, "Evaluating the Employment Impact of a Mandatory Job Search Program," *Journal of the European Economic Association*, June 2004, *2* (4), 569–606. 2

Kahn-Lang, Ariella and Kevin Lang, "The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications," *Journal of Business & Economic Statistics*, July 2020, *38* (3), 613–620. 4, 21

Lechner, Michael, "The Estimation of Causal Effects by Difference-in-Difference Methods," *Foundations and Trends in Econometrics*, 2011, *4* (3), 165–224. 2, 12, 16

Meyer, Bruce D., "Natural and Quasi-Experiments in Economics," *Journal of Business & Economic Statistics*, April 1995, *13* (2), 151–161. 16

Mullahy, John, "Interaction Effects and Difference-in-Difference Estimation in Loglinear Models," NBER Technical Working Paper 0245, National Bureau of Economic Research November 1999. 2

Puhani, Patrick A., "The Treatment Effect, the Cross Difference, and the Interaction Term in Nonlinear "Difference-in-Differences" Models," *Economics Letters*, April 2012, *115* (1), 85–87. 2, 3, 5, 11, 15, 16

Roth, Jonathan and Pedro H. C. Sant'Anna, "When Is Parallel Trends Sensitive to Functional Form?," Working Paper January 2021. 16

Ryan, Andrew M., James F. Burgess, and Justin B. Dimick, "Why We Should Not Be Indifferent to Specification Choices for Difference-in-Differences," *Health Services Research*, 2015, *50* (4), 1211–1235. 2

Shang, Shengwu, Erik Nesson, and Maoyong Fan, "Interaction Terms in Poisson and Log Linear Regression Models," *Bulletin of Economic Research*, January 2018, *70* (1), E89–E96. 18

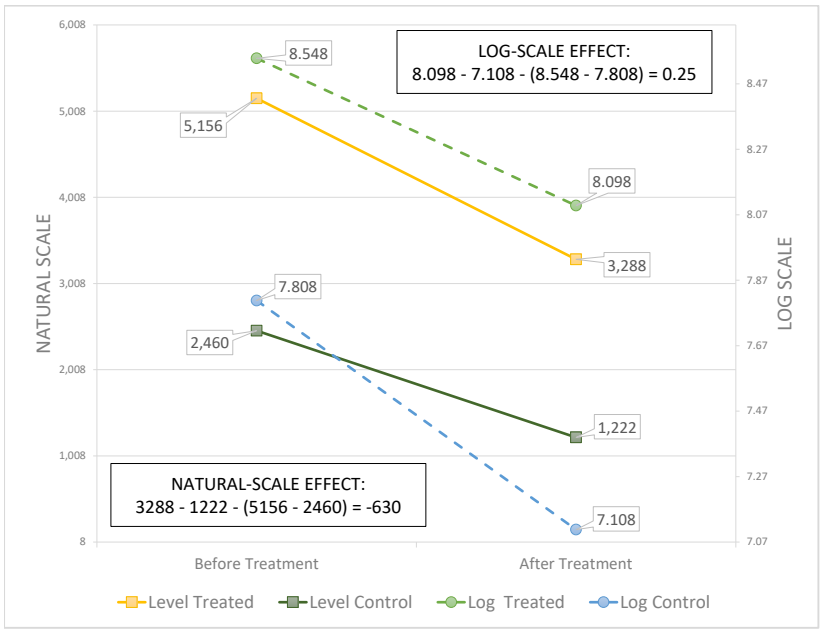

Figure 1: Hypothetical treatment and control group changes, presented in natural and log scale.