Much Ado About Not Very Much? Clarifying the Confusion about Models for Categorical Dependent Variables

Martin Elff

EPSA Conference 2022
Prague Congress Centre
23rd – 25th June 2022

Models for categorical dependent variables, such as turnout, party choice, or partisanship have eluded scholars for decades. Their parameters are often difficult to interpret and do not lend themselves easily to an intuitive understanding. Yet, the concepts and patterns of inference that work will within the framework of linear regression cannot easily transferred to models for categorical dependent variables. The paper discusses two instances where attempts to do this leads to misleading methodological recommendations.

It should be noted that in its current state the paper is quite incomplete. Its argumentation is currently rather formal-mathematical, while its practical and empirical implications need to be further fleshed out.

Models for categorical dependent variables, such as turnout, party choice, or partisanship have eluded scholars for decades. Should the coefficients be reported and interpreted, or are odds ratios, marginal effects, discrete changes in probabilities preferable? Can there be statistically significant interaction effects without the presence of product terms? How is it possible to make sense of multinomial logit models, if their coefficients depend on the choice of the response category? Several scholars have published methodological papers concerned with these questions, but not all of these have the potential to help with the confusion of

researchers who intend to apply models for categorical responses. In some instances the proposed cure is even worse than the disease. Yet much of the confusion about models for categorical dependent variables and other non-linear models can be avoided by not asking too them, by more strictly adhering to an internally consistent terminology, and heeding the distinction between model specification, estimation, and generating quantities of interest.

These issues are discussed with regards to two more or less recent papers, one already widely cited, the other published in a journal with high visibility. The first is Berry et al.'s "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?" (Berry et al. 2010). The other is Paolino's "Predicted Probabilities and Inference with Multinomial Logit" Paolino (2020). The common ground of both papers is that they try go give priority to quantities of interest that appear to have a straightforward intuitive interpretation and that generalise concepts that work well in the framework of linear regression over quantitative and concepts that are straightforward aspects of models of categorical dependent variables but are difficult to interpret. In both cases, these priorities leads to misleading recommendations with regards to inferences derived from these models.

1 Product terms and interaction effects in logit and probit models

Within the context of linear regression, interaction effects can be defined as the amount to which first differences in expected values of a dependent variable X with respect to an independent variable change with the values of another variable Z. In this interpretation, an interaction effect is identical to the cross-variable second-difference ratio:

$$\frac{\Delta^2 \mu(x,z)}{\Delta x \Delta z} = \frac{\mu(x,z) - \mu(x,z) - [\mu(x,z) - \mu(x,z)]}{(x_2 - x_1)(z_2 - z_1)}$$

where

$$\mu(x,z) = \mathrm{E}(Y|X = x, Z = z)$$

for some $x_1 < x_2$, $z_1 < z_2$. Due to the mean value theorem of calculus, the second-difference ratio with respect to x_1,x_2,z_1,z_2 is equal to the second derivative

$$\left.\frac{\partial^2 \mu(x,z)}{\partial x \partial z}\right|_{x=x^*, z=z^*}$$

for some x^* and z^* with $x_1 \leq x^* \leq x_2$ and $z_1 \leq z^* \leq zx_2$

In a linear regression model without a product term in X and Z, that is,

$$\mu(x, z) = \alpha + \beta_x x + \beta_z z + \cdots$$

it is easy to see that the first derivatives of $\mu(x, z)$ in x and z, the *marginal effects*, are constant and equal to the coefficients

$$\frac{\partial \mu(x, z)}{\partial x} = \beta_x \quad \frac{\partial \mu(x, z)}{\partial z} = \beta_z$$

while the second derivative vanishes:

$$\frac{\partial^2 \mu(x, z)}{\partial x \partial z} = 0$$

In a linear regression model with a product term in X and Z

$$\mu(x, z) = \alpha + \beta_x x + \beta_z z + \beta_{xz} xz + \cdots$$

(where $\beta_{xz} \neq 0$)the marginal effects are linear functions

$$\frac{\partial \mu(x, z)}{\partial x} = \beta_x + \beta_{xz} z \quad \frac{\partial \mu(x, z)}{\partial z} = \beta_z + \beta_{xz} x$$

while the second "cross-derivative", which henceforth is referred to as the *marginal interaction*, equals the coefficient of the product term:

$$\frac{\partial^2 \mu(x, z)}{\partial x \partial z} = \beta_{xz}.$$

This is of course well known: that the presence of a product term is a necessary and consistent condition for the presence of an interaction effect. It should be noted that a model without an interaction term in X and Z can be viewed as a special case of a model with an interaction term where the coefficient of this term is equal to zero, i.e. β_{xz}. Therefore the following discussion uses equations involving interaction terms and treats the "absence" of an interaction term as the special case β_{xz}.

The three most popular models for binary dependent variables are linear probability models, logistic regression (aka logit) models, and probit models.

A *linear probability model* with independent variables X and Z takes the form

$$\pi(x,z) = \begin{cases} \alpha + \beta_x x + \beta_z z + \beta_{xz} xz & \text{for } 0 < \alpha + \beta_x x + \beta_z z + \beta_{xz} xz < 1 \\ 0 & \text{for } \alpha + \beta_x x + \beta_z z + \beta_{xz} xz \leq 0 \\ 1 & \text{for } \alpha + \beta_x x + \beta_z z + \beta_{xz} xz \geq 1 \end{cases}$$

(again assuming $\beta_{xz} \neq 0$). Where $\pi(x,z) = \Pr(Y = 1|X = x, Z = z) = \mathrm{E}(Y|X = x, Z = z)$

It seems that the nice equivalence between the existence of interaction effects and the presence of product terms carries over from linear regression models to linear probability models. However, this applies only at most as long as $0 < \alpha + \beta_x x + \beta_z z + \beta_{xz} xz < 1$. For $\alpha + \beta_x x + \beta_z z + \beta_{xz} xz = 0$ and $\alpha + \beta_x x + \beta_z z + \beta_{xz} xz = 1$ first and second derivatives (and the corresponding difference ratios) are undefined, while for $\alpha + \beta_x x + \beta_z z + \beta_{xz} xz < 0$ and $\alpha + \beta_x x + \beta_z z + \beta_{xz} xz > 1$ all derivatives vanish (are equal to zero).

Both logit and probit models with independent variables X and Z take the form

$$\Pr(Y = 1|X = x, Z = z) = \mathrm{E}(Y|X = x, Z = z) = \pi(x,y) = f(\alpha + \beta_x x + \beta_z z + \beta_{xz} xz)$$

where

$$f(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)} = \frac{1}{1 + \exp(-\eta)}$$

for the logit model and

$$f(\eta) = \int_{-\infty}^{\eta} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t$$

for the probit model. Both functions are not only non-linear but also non-polynomial, that is for any cardinal number n

$$\frac{\mathrm{d}^n f(\eta)}{\mathrm{d}\eta^n} \neq 0$$

except for a finite set of values of η.

For the logit model we have

$$f'(\eta) = \frac{\mathrm{d}f(\eta)}{\mathrm{d}\eta} = f(\eta)[1 - f(\eta)]$$

and

$$f''(\eta) = \frac{\mathrm{d}^2 f(\eta)}{\mathrm{d}\eta^2} = f(\eta)[1 - f(\eta)][1 - 2f(\eta)]$$

while for the probit model we have

$$f'(\eta) = \frac{\mathrm{d}f(\eta)}{\mathrm{d}\eta} = \exp\left(-\frac{\eta^2}{2}\right)$$

and

$$f''(\eta) = \frac{\mathrm{d}^2 f(\eta)}{\mathrm{d}\eta^2} = -\eta \exp\left(-\frac{\eta^2}{2}\right)$$

For both models we have $f'(\eta) > 0$ for all values of η, while $f''(\eta) = 0$ if and only if $\eta = 0$.[1]

The marginal effects of X and Z are

$$\begin{aligned} \dot{\pi}_x(x,z) &:= \frac{\partial \pi(x,z)}{\partial x} = f'(\alpha + \beta_x x + \beta_z z + \beta_{xz} xz)(\beta_x + \beta_{xz} z) \\ \dot{\pi}_z(x,z) &:= \frac{\partial \pi(x,z)}{\partial z} = f'(\alpha + \beta_x x + \beta_z z + \beta_{xz} xz)(\beta_x + \beta_{xz} x) \end{aligned} \tag{1}$$

while cross-derivative of the marginal effects, the marginal interaction, is

$$\begin{aligned} \ddot{\pi}_{xz}(x,z) := \frac{\partial^2 \pi(x,y)}{\partial x \partial z} &= f'(\alpha + \beta_x x + \beta_z z + \beta_{xz} xz)\beta_{xz} \\ &\quad + f''(\alpha + \beta_x x + \beta_z z + \beta_{xz} xz)(\beta_x + \beta_{xz} z)(\beta_z + \beta_{xz} x). \end{aligned} \tag{2}$$

An immediate implication of this is that marginal effects in logit and probit models are not constant - if they are defined in terms of (partial) derivatives. If it is considered as a criterion for the absence of interaction effects that marginal effects are constant, then logit and probit models will always contain interaction effects, even in the absence of interaction terms. Yet if this is the case the assertion that a logit or probit model contains interaction effects is no longer informative. Accordingly, various authors such as Nagler (1991) or Frant (1991) have argued that non-constant marginal effects should not be interpreted as interaction effects because they are direct implications of model assumptions. Instead, an interaction effect should only be considered as a substantial aspect of a logit or probit model if the coefficient of the corresponding product term is non-zero.

There is an intuitive way one can make sense of these non-constant marginal effects: If the probability of a positive outcome is about $\frac{1}{2}$ the slope of $\pi(x,z)$ with respect to changes in x or z is maximal in both logit and probit models. If $\pi(x,z)$ is close to unity, an increase in x or z simply cannot lead to the same change in the probability of a positive outcome in the response than if $\pi(x,z)$ is close to $\frac{1}{2}$, because it cannot become greater than unity. Similarly, $\pi(x,z)$ cannot change in the same amount if it is close to zero as if it is close to $\frac{1}{2}$ if x or

[1] This is trivial for probit models. For logit models we have $f(0) = \frac{1}{2} \Rightarrow 1 - 2f(\eta) = 0 \Rightarrow f''(\eta) = 0$.

z are decreased, because a probability cannot get less than zero. This consequence of the restriction of probabilities to the range from zero to unity is what various authors (Berry et al. 2010; Rainey 2016) refer to as "compression".

While earlier authors like Nagler (1991) and Frant (1991) argue that the consequences of "compression" should not be taken as theoretically meaningful, more recent authors have come forward with contrary recommendations. Berry et al. (2010) state in a widely cited *American Journal of Political Science* article that an interaction effects are not merely artefacts of the non-linearity of logit or probit models – instead they are often of theoretical importance. Further, they argue "that a statistically significant product term is neither necessary nor sufficient for claiming interaction among independent variables in influencing" (250) the probability of a positive outcome of a binary response model. Instead, they argue that claims about interactions should not be based on the presence or significance of product terms, but on the substantive implications of the theory that one translates into a binary response model. It will be argued further below that this recommendation partially rests on an awkward, if not mistaken understanding of statistical significance.[2]

There are two main methodological arguments that Berry et al. (2010) bring forward in support of their claim:

1. a statistically significant (coefficient estimate of a) product term is not a *necessary* condition for a substantive meaningful interaction between to variables with regards to their influence on the probability of a positive outcome of the dependent variable – there are instances where such an interaction is of substantial size and statistical significant even if there is no corresponding product term present in the model (Berry et al. 2010, 257).

2. a statistically significant (coefficient estimate of a) product term is not a *sufficient* condition for a substantive meaningful interaction between to variables with regards to their influence on the probability of a positive outcome of the dependent variable – there are instances where there is no statistically significant interaction even though there *is* a statistically significant product term present in the model (Berry et al. 2010, 258).

Berry et al. (2010) however do not mean to dispute the mathematical properties presented above, they do claim however that the presence of a product term in a logit or probit model is not the same as the presence of a "substantively meaningful interaction" of two

[2]It should be noted that the formulation in the cited statement is already a bit sloppy, as it is not a product term that can be either statistically significant or not, but the *estimate of a coefficient* of such a product term.

variables in influencing a binary dependent variable. They propose two criteria for such a meaningful interaction: "(1) that the interaction is statistically significant" and "(2) the estimated magnitude of the interaction is large enough to be deemed consequential". It is important to note here that the magnitude of the interaction (in the notation of this paper) is *not*

$$\frac{[\pi(x_2, z_2) - \pi(x_1, z_2)] - [\pi(x_2, z_1) - \pi(x_1, z_1)]}{(x_2 - x_1)(z_2 - z_1)} = \frac{\Delta^2\pi(x, z)}{\Delta x \Delta z} \tag{3}$$

but

$$[\pi(x_2, z_2) - \pi(x_1, z_2)] - [\pi(x_2, z_1) - \pi(x_1, z_1)] = \Delta^2\pi(x, z). \tag{4}$$

Formalised as suggested by Berry et al. (2010), the first criterion means that the null hypothesis $\Delta^2\pi(x, z) = 0$ can be rejected with some pre-determined level of statistical significance and the alternative hypothesis $\Delta^2\pi(x, z) \neq 0$ can be accepted, while the second criterion means that $|\Delta^2\pi(x, z)| > c$ for some pre-determined c.

Due to the mean value theorem of differential calculus this means that this "Berry interaction" is equal to

$$\Delta^2\pi(x, z) = (x_2 - x_1)(z_2 - z_1)\ddot{\pi}_{xz}(x^*, z^*)$$

for some $x_1 \leq x^* \leq x_2$ and $z_1 \leq z^* \leq z_2$. Translated back into the world of linear regression, this "Berry interaction" would not be equal to β_{xz} but to $(x_2 - x_1)(z_2 - z_1)\beta_{xz}$.

From the preceding considerations it becomes obvious that the two claims cited above are mathematically trivial: $\beta_{xy} \neq 0$ is *not necessary* for the "Berry interaction" $\Delta^2\pi(x, z)$ to be statistically significant and of nontrivial magnitude, because for any $\beta_{xy} \neq 0$ one just needs to choose x_1,x_2,z_1,z_2 such that $(x_2 - x_1)(z_2 - z_1)$ is small enough that $|\Delta^2\pi(x, z)| < c$. Conversely, $\beta_{xy} \neq 0$ is *not sufficient* for the "Berry interaction" $\Delta^2\pi(x, z)$ to be statistically significant and of nontrivial magnitude, because even with $\beta_{xz} = 0$ the marginal interaction $\ddot{\pi}_{xz}(x^*, z^*)$ is different from zero for almost all of the infinitely many possible values of x^* and z^*, so that x_1,x_2,z_1,z_2 can be chosen such that $(x_2 - x_1)(z_2 - z_1)$ is large enough to make $|\Delta^2\pi(x, z)| > c$ and $|\Delta^2\pi(x, z)| > c^\dagger$ where $c^\dagger$ is a constant for which $|\Delta^2\pi(x, z)|$ can be judged to be statistically significant.

It could be argued that this "reduction to mathematical triviality" is too simple to do justice to Berry et al.' arguments and that the mathematical argument is of little practical importance. The main counterargument could be that the choice of x_1,x_2,z_1,z_2 is not just arbitrary, but these values of the independent variables are chosen on substantive grounds. Therefore, the following considerations focus on the role of the second derivatives $\ddot{\pi}_{xz}(x, z)$

for the statistical significance of the Berry interaction $\Delta^2\pi(x,z)$. Thereafter, an obviously non-arbitrary choice of x_1,x_2,z_1,z_2 is considered and what it then would mean that the Berry interaction has or has not a nontrivial magnitude.

For the sake of the argument let us assume that x_1,x_2,z_1,z_2 are chosen in a non-arbitrary, substantively meaningful fashion and let us for brevity define $a = (x_2 - x_1)(z_2 - z_1) \neq 0$. The null hypothesis $\Delta^2\pi(x,z) = 0$ then means that

$$a\ddot{\pi}_{xz}(x^*, z^*) = 0 \tag{5}$$

for appropriate values of x^* and z^* with $x_1 \leq x^* \leq x_2$ and $z_1 \leq z^* \leq z_2$.

If $\beta_x = \beta_z = \beta_{xz} = 0$ then this equation is trivially satisfied. If $\beta_{xz} = 0$, equation (5) requires that $\beta_x = 0$ or $\beta_z = 0$ or the linear equation

$$\alpha + \beta_x x^* + \beta_z z^* = 0 \tag{6}$$

is satisfied, because $f''(0) = 0$. For any given x^* and z^* there are infinitely many values of α, β_x, and β_z that solve the equation. Yet the solution set depends on x^* and z^*: If $\beta_z \neq 0$ and we substitute $z^\dagger = z^* + \epsilon$ ($\epsilon \neq 0$) for z^* in equation (6) we get

$$\alpha + \beta_x x^* + \beta_z z^\dagger = \beta_z \epsilon \neq 0$$

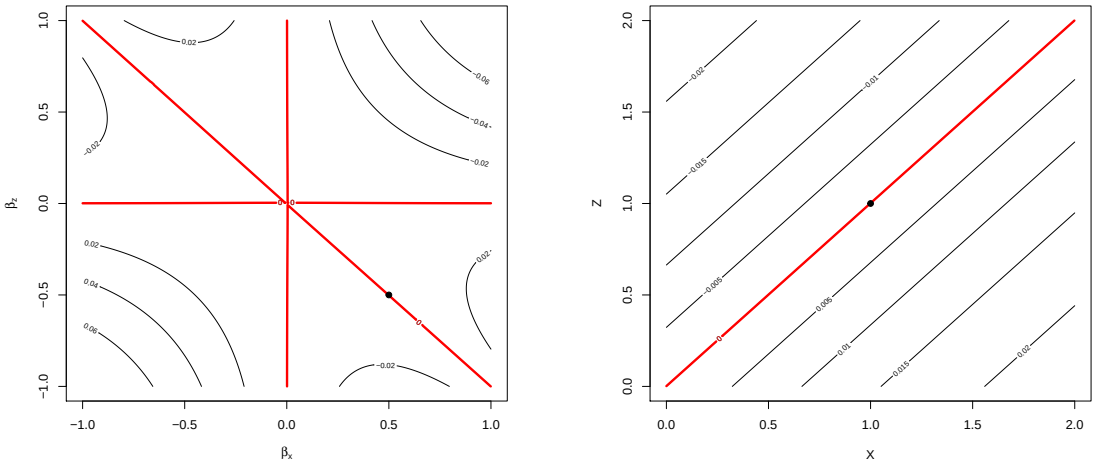
that is, the linear equation is no longer satisfied, unless we make substitutions that change α, β_x or β_z, for example $\alpha^\dagger = \alpha - \beta_z\epsilon$.

The interdependence between values of α, β_x, β_z, x and z that solve the equation $\ddot{\pi}(x,z) = 0$ is illustrated in figure 1. The left-hand sub-figure shows that for given values $x = 1$ and $z = 1$ all non-trivial solutions indeed lie on a straight line. One such solutions is $(\beta_x;\beta_z) = (\frac{1}{2};-\frac{1}{2})$, which is marked by a dot in the left-hand sub-figure. The right-hand sub-figure illustrates that $(\beta_x;\beta_z) = (\frac{1}{2};-\frac{1}{2})$ solves the equation not in general, but only for a specific subset of pairs of values of x and z, namely those for which $x = z$.

If $\beta_{xz} \neq 0$, equation (5) requires that

$$f'(\alpha+\beta_x x^*+\beta_z z^*+\beta_{xz}x^*z^*)\beta_{xz}+f''(\alpha+\beta_x x^*+\beta_z z^*+\beta_{xz}x^*z^*)(\beta_x+\beta_{xz}z^*)(\beta_z+\beta_{xz}x^*) = 0. \tag{7}$$

The idea that this equation may have non-trivial solutions leads Berry et al. (2010) to argue that $\beta_{xz} \neq 0$ is not sufficient for $\Delta^2\pi(x,z)$ to be non-zero. Rainey (2016) goes even further to argues that a product term should *always* be included in a logit or probit model, because



(a) $x = 1$ and $z = 1$, while β_x and β_z vary. Red lines mark those values for β_x and β_z, where $\ddot{\pi}(x, z) = 0$. The black dot corresponds to the values for β_x and β_z used in the right-hand sub-figure.

(b) $\beta_x = \frac{1}{2}$ and $\beta_z = -\frac{1}{2}$, while x and z vary. Red lines mark those values for x and z, where $\ddot{\pi}(x, z) = 0$. The black dot corresponds to the values for x and z used in the left-hand sub-figure.

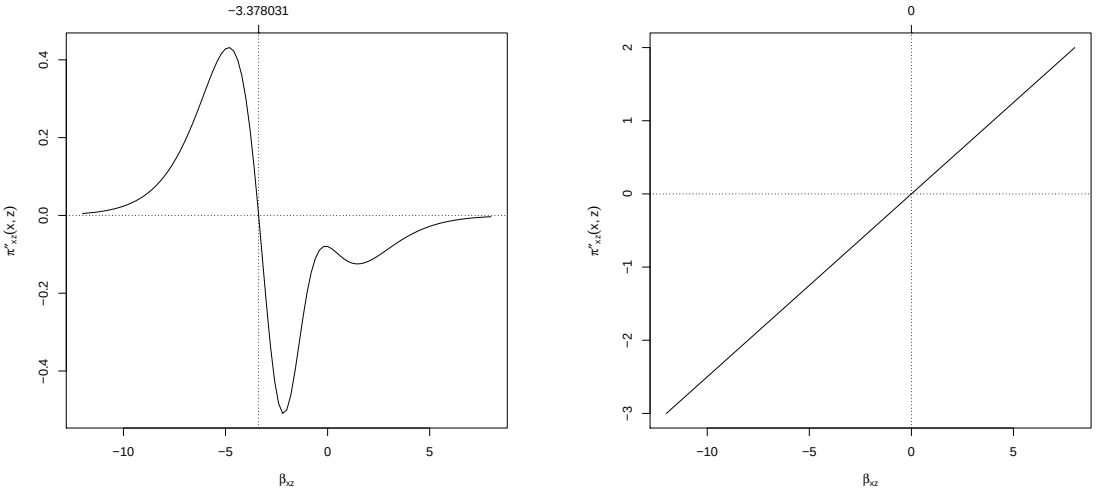
Figure 1: Illustration of the conditions for $\ddot{\pi}(x, z) = 0$, when there is no product term (with non-zero coefficient) in the model

only then it is possible to test whether there is an interaction of X and Z in their influence on Y.

Indeed for given values x^* and z^* there may be infinitely many different values for α, β_x, β_z, and β_{xz} for which equation $\ddot{\pi}_{xz}(x^*, z^*) = 0$ is satisfied. Yet the solution set will depend on the particular values x^* of z^*. For example, a set of coefficient values that satisfies $\ddot{\pi}_{xz}(x^*, z^*) = 0$ for $x^* = 1$ and $z^* = 1$ may not satisfy this equation for $x^* = 0$ and $z^* = 0$. This is illustrated in Figure 2. Both diagrams show how $\ddot{\pi}_{xz}(x, z)$ varies with β_{xz} for given settings of $\alpha = 0$, $\beta_x = 1$, $\beta_z = 1$, x, and z. In the left-hand sub-figure the settings of the independent variables are $x = 1$ and $z = 1$, and the value of β_{xz} for which $\ddot{\pi}_{xz}(x^*, z^*) = 0$ is satisfied is $-3.378031\ldots$ (the value is identified by solving the equation numerically[3]). In the right-hand sub-figure the settings of the independent variables are $x = 0$ and $z = 0$, and the value of β_{xz} for which $\ddot{\pi}_{xz}(x^*, z^*) = 0$ is a different one, namely zero. That is, the set of coefficient values for which the equation is satisfied are different for $x = 0$ and $z = 0$ from the ones for $x = 1$ and $z = 1$.

For both zero and non-zero coefficient of product terms, $\ddot{\pi}_{xz}(x, z)$ is non-zero almost everywhere for x and z, but there are also values for which $\ddot{\pi}_{xz}(x, z)$ is zero. Since $\ddot{\pi}_{xz}(x, z)$ is continuous for logit and probit models, it can be globally non-zero only if it either globally

[3]For this purpose the function `uniroot()` of the software package *R* was used (R Core Team 2020).



(a) Values of $\ddot{\pi}(x,z)$, depending on β_{xz}, with independent variables fixed at $x = 1$ and $z = 1$ and their coefficients set to $\beta_x = 1$ and $\beta_z = 1$. The vertical dotted line indicates the value of β_{xz} for which $\ddot{\pi}(x,z) = 0$.

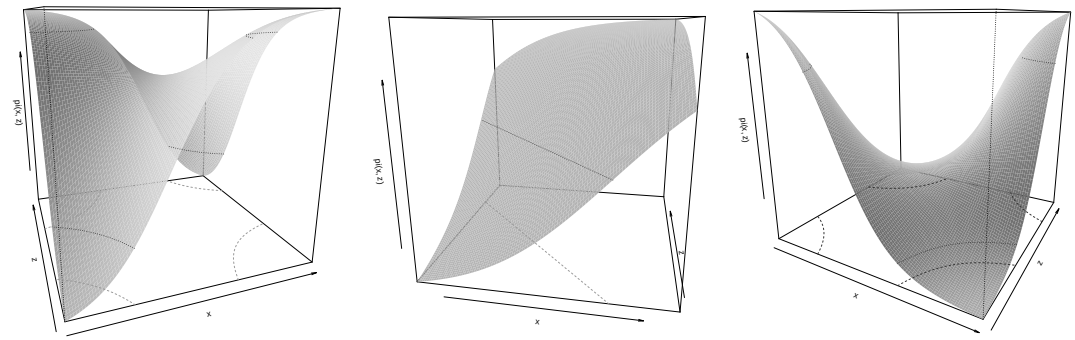
(b) Values of $\ddot{\pi}(x,z)$, depending on β_{xz}, with independent variables fixed at $x = 0$ and $z = 0$ and their coefficients set to $\beta_x = 1$ and $\beta_z = 1$. The vertical dotted line indicates the value of β_{xz} for which $\ddot{\pi}(x,z) = 0$, which in this case is $\beta_{xz} = 0$.

Figure 2: The cross-derivative $\ddot{\pi}_{xz}(x,z)$ as a function of the coefficient of a product term for different settings of x and z.

smaller or larger than zero. This, however, would require that $\pi_{xz}(x,z)$ is either globally concave or convex. Yet this is impossible, because $\pi_{xz}(x,z)$ is strictly bounded between zero and unity. This is illustrated by Figure 3 and . Figure 3 shows how the probability of a positive outcome of a dependent variable in a logit model varies with the values of the independent variables and also shows for which values of the independent variables the cross-derivatives are zero. In all three diagrams there are sets of values of x and z for which the cross-derivatives are zero, even though the diagrams are different in terms of the sign of the coefficient of the product term.

Figure 4 traces the "diagonals" of the diagrams in Figure 3 defined by $x = z$. Thus Figure 4 shows curves instead of surfaces. The dotted vertical lines indicate the values for x and z for which $\ddot{\pi}_{xz}(x,z)$ is zero (the values are identified by solving the equation numerically[4]). Here the roots of $\ddot{\pi}_{xz}(x,z)$ for $x = z$ mark the inflection points of the curves. Without the inflection points the curve of $\pi(x,z)$ in sub-figure 4a would continue downwards with $x = z$ going to $-\infty$ or $+\infty$, eventually becoming less than zero. In sub-figure 4c the curve would continue upwards with $x = z$ going to $-\infty$ or $+\infty$, eventually becoming greater than unity.

[4]For this purpose the function `uniroot()` of the software package *R* was used (R Core Team 2020)



(a) Diagram for $\alpha = 0$, $\beta_x = 1$, $\beta_z = 1$, and $\beta_{xz} = -1$; $-1 \leq x \leq +3$ and $-1 \leq z \leq +3$.

(b) Diagram for $\alpha = 0$, $\beta_x = 1$, $\beta_z = 1$, and $\beta_{xz} = 0$; $-1.5 \leq x \leq +2.5$ and $-1.5 \leq z \leq +2.5$.

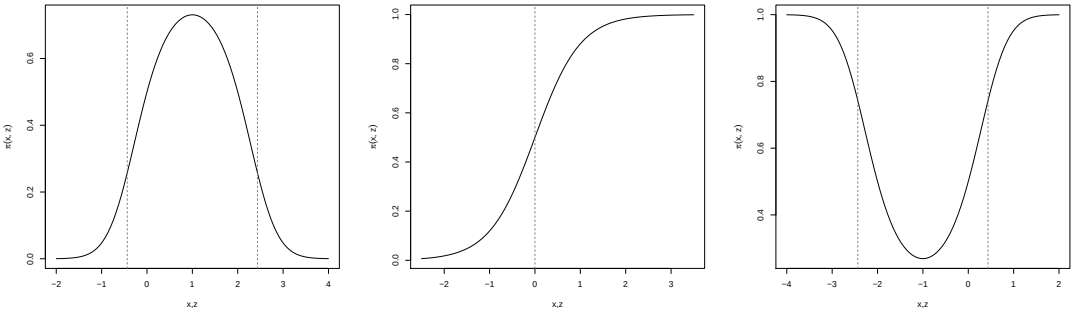
(c) Diagram for $\alpha = 0$, $\beta_x = 1$, $\beta_z = 1$, and $\beta_{xz} = +1$; $-3 \leq x \leq +1$ and $-3 \leq z \leq +1$.

Figure 3: Surface diagrams of $\pi(x, z)$ with countour lines indicating where $\ddot{\pi}_{xz}(x, z)$ is zero for logit models with negative, zero, and positive product term coefficients.

In sub-figure 4b the curve would continue downward with $x = z$ going to $-\infty$ and continue upward with $x = z$ going to $+\infty$, eventually becoming less than zero or greater than unity, respectively.

What these considerations show is, firstly, that for almost all values of the independent variables of a logit or probit model, the marginal interaction is different from zero, whether or not the coefficient of a product term is zero. Secondly, for *some* values of the independent variables of a logit or probit model, the marginal interaction equals zero, yet these values depend on the coefficient values. Except for trivial cases where one or all coefficients of the independent variables are zero, the values of the independent variables that lead to zero marginal interactions depend on the coefficient values and there are no instances where marginal interactions are zero for all values of the independent variables. Thirdly, it is not only a consequence of "compression" that marginal interactions are nonzero almost everywhere, "compression" also necessitates that marginal interactions are zero for a subset of the values of the independent variables.

Statistical hypotheses about regression models, generalised linear models, and other models of the influence of independent variables, say X and Z, on a dependent variable, say Y, are about the form that this influence takes, without any restrictions on the values of the independent variables. Therefore, if a null hypothesis cannot be true unless for very specific values of the independent variables, it should be assumed to be false. If a consistent test is used for a null hypothesis that is false, the probability that the test leads to its rejection *per definitionem* approaches 100% as the sample size approaches infinity. That is, if we reject a null hypothesis known to be false in a particular instance, then all we learn is that the sample



(a) Diagram for $\alpha = 0$, $\beta_x = 1$, $\beta_z = 1$, and $\beta_{xz} = -1$; $-1 \leq x = z \leq +3$.

(b) Diagram for $\alpha = 0$, $\beta_x = 1$, $\beta_z = 1$, and $\beta_{xz} = 0$; $-1.5 \leq x = z \leq +2.5$.

(c) Diagram for $\alpha = 0$, $\beta_x = 1$, $\beta_z = 1$, and $\beta_{xz} = +1$; $-3 \leq x = z \leq +1$.

Figure 4: Curves of $\pi(x, z)$ with lines indicating where $\ddot{\pi}_{xz}(x, z)$ is zero for $x = z$ for logit models with negative, zero, and positive product term coefficients.

size is large enough for this. Conversely, if we do not reject a null hypothesis known to be false, then what we should conclude is that our sample does not provide enough statistical power to do so.

Apart from an insufficient sample size, there are two other possibilities that may lead researchers to the conclusion that their analysis does not support a non-zero interaction. The first possibility is that the range of the values of the relevant independent variables is so small that the first and second differences are scaled down so much that they appear insubstantial in size or statistically significant. This first possibility is only relevant for the computation of a "raw" second difference or what has so far been referred to as a "Berry interaction". The second possibility is that the values of the independent variables are concentrated around the inflection points of $\pi(x, z)$, i.e. those values of x and z where the second derivatives $\ddot{\pi}_{xz}(x, z)$ happen to be equal to zero.

The lesson to learn from this discussion is that if researchers find or do not find interaction effects (understood as second differences, second difference ratios, or marginal interactions) predicted by a logit or probit model fitted to their data, it does not provide information about the data-generating process, but about the (quality of the) sample. That not withstanding, both Berry et al. (2010) and Rainey (2016) suggest, or rather pre-suppose that it may be of theoretical relevance whether such interaction effects are present and that they are not artefacts created by the use of logit, probit or other non-linear models. Rainey (2016) even recommends that one should always include product terms to allow for the absence of interactions and make the presence of an interaction a testable hypothesis. If their arguments

is to have any genuine relevance, the absence of such an interaction should theoretically and empirically make sense.

Given that interactions (understood as second differences, second difference ratios, or marginal interactions) are an intrinsic property of logit, probit, or similar models, testing hypotheses about their presence is futile unless one is willing to consider other kind of models for binary dependent variables. If scholars have strong theoretical reasons to expect that independent variables do not have any compression-induced interactions for a wider than a single point in the independent variables – which does not seem possible unless the function that links the independent variable to the probability $\pi(x, y)$ of a positive response – they should look for alternatives to linear logit or probit models. One such alternative is an additive logit or additive probit model (Hastie and Tibshirani 1990; Ruppert et al. 2003). With two independent variables an additive logit model for a binary dependent variable takes the form

$$\ln \frac{\Pr(Y = 1|X = x, Z = z)}{\Pr(Y = 0|X = x, Z = z)} = s_x(x) + s_z(z)$$

where $s_x(x)$ and $s_z(z)$ are unrestricted continuous univariate functions estimated e.g. with the help of smoothing splines. Such a model can also contain a joint term of more both independent variables:

$$\ln \frac{\Pr(Y = 1|X = x, Z = z)}{\Pr(Y = 0|X = x, Z = z)} = s_{xz}(x, z)$$

where $s_{xz}(x, z)$ is an continuous bivariate functions estimated e.g. with the help of a tensor-product or thin-plate spline (Hastie and Tibshirani 1990; Ruppert et al. 2003). Another alternative would be to use a model with a non-parametric link between a linear part of the model and the probability of a positive response (e.g. Manski 1975). Linear logistic models can be considered to be nested in such semi-parametric alternatives. Thus, while it is futile to test for marginal interactions or Berry-interactions within the framework of linear logit or linear probit models, it is possible with such non-parametric alternatives. Yet in this case, a hypothesis of suppression-induced interaction effects no longer has the role of an alternative hypothesis (as envisaged by Berry et al. 2016) but that of a null hypothesis. Finally, one could consider linear-probability models (and in fact such a linear probability model is the basis of the data generating process in Rainey's simulation study Rainey 2016). Yet in this case one has to deal with the fact that neither a linear logit or probit model is a special case of a linear probability-model or vice versa. Furthermore, unless there are sensible a-priori (and not sample-dependent) restrictions on the range of the independent variables, a linear probability-model will also exhibit "compression", which however unfortunately is non-smooth.

As already remarked on at the beginning of this section, "compression" is made inevitable by the fact that response probabilities are *per definitionem* restricted between zero and 100 percent. If a scholar's theory implies the absence of Berry-interactions or marginal interactions, then should consider alternatives to logit and probit like those just discussed. Yet if their theory implies that a probability of a positive outcome of a binary dependent variable may lie outside the range from zero to 100 percent, then something is wrong with the theory and they should look for another one. Yet rarely are social science theories elaborated to such a degree and scales of measurement so well developed that they imply any functional form for regression models or generalised regression models, not to mention the presence or absence of marginal interactions. In so far, the discussion of the presence or absence of such interaction in logit or probit models appears indeed to be much ado about nothing.

2 Baseline categories, predicted probabilities, and inference in multinomial logit models

A common problem in the interpretation of multinomial baseline logit models is the large number of coefficients they involve. This makes it difficult to interpret estimates and can even create problems for their presentation if the number of independent variables is large – not to mention if models also include product terms. This problem is exacerbated by the fact that the size, direction, and meaning of the coefficients in a multinomial baseline model depends on the choice of the baseline category. For example, the equations for a multinomial logit model for a qualitative dependent variable Y with categories A, B, and C and independent variable X has, if A is chosen as baseline category, the form

$$\ln \frac{\Pr(Y = B|X = x)}{\Pr(Y = A|X = x)} = \alpha_B + \beta_B x$$
$$\ln \frac{\Pr(Y = C|X = x)}{\Pr(Y = A|X = x)} = \alpha_C + \beta_C x$$

while if C is chosen as baseline category it has the form:

$$\ln \frac{\Pr(Y = A|X = x)}{\Pr(Y = C|X = x)} = \alpha_A^\dagger + \beta_A^\dagger x$$
$$\ln \frac{\Pr(Y = B|X = x)}{\Pr(Y = C|X = x)} = \alpha_B^\dagger + \beta_B^\dagger x$$

where $\alpha_A^\dagger = -\alpha_C$, $\beta_A^\dagger = -\beta_C$, $\alpha_B^\dagger = \alpha_B - \alpha_C$, and $\beta_B^\dagger = \beta_B - \beta_C$. This is so because

$$\ln \frac{\Pr(Y = A|X = x)}{\Pr(Y = C|X = x)} = -\ln \frac{\Pr(Y = C|X = x)}{\Pr(Y = A|X = x)}$$

and

$$\begin{aligned}\ln \frac{\Pr(Y = B|X = x)}{\Pr(Y = C|X = x)} &= \ln \frac{\Pr(Y = B|X = x)}{\Pr(Y = A|X = x)} + \ln \frac{\Pr(Y = A|X = x)}{\Pr(Y = C|X = x)} \\ &= \ln \frac{\Pr(Y = B|X = x)}{\Pr(Y = A|X = x)} - \ln \frac{\Pr(Y = C|X = x)}{\Pr(Y = A|X = x)}\end{aligned}$$

As a consequence it is possible that coefficients are non-zero in one formulation of the model but zero in another. This quite obviously is the case, for example, if $\beta_B = \beta_C$.

However, the way the probabilities are related to independent variables is independent from the choice of the baseline category. For example,

$$\begin{aligned}\Pr(Y = C|X = x) &= \frac{\exp(\alpha_C + \beta_C x)}{1 + \exp(\alpha_B + \beta_B x) + \exp(\alpha_C + \beta_C x)} \\ &= \frac{1}{1 + \exp(-\alpha_C - \beta_C x) + \exp((\alpha_B - \alpha_C) + (\beta_B - \beta_c)x)} \\ &= \frac{1}{1 + \exp(\alpha_A^\dagger + \beta_A^\dagger x) + \exp(\alpha_B^\dagger + \beta_B^\dagger x)}.\end{aligned}$$

Yet it is also straightforward to see that it is not possible within the framework of a multinomial logit model, that the conditional probability of any of the outcome categories is constant while the other probabilities are affected by the independent variables. In the previous example, the probability that $Y = C$ is independent from X is possible only if $\beta_A^\dagger = 0$ and $\beta_B^\dagger = 0$, which in turn implies that the probabilities that $Y = A$ or $Y = B$ are also independent from X. In order to allow e.g. $\Pr(Y = C|X = x)$ to be constant [i.e. $\Pr(Y = C|X = x) = \Pr(Y = C)$], while $\Pr(Y = A|X = x)$ and $\Pr(Y = B|X = x)$ vary with x one needs to give up the (linear) multinomial logit model. One alternative would be a model where the influence on the independent variables on log-odds of the outcome probabilities is non-linear, a generalisation discussed at the end of the previous section. Another alternative would be a sequential logit model as proposed by . In the previous example, a sequential logit model that would have $\Pr(Y = A|X = x)$ to be constant is

$$\begin{aligned}\ln \frac{\Pr(Y = A|X = x)}{\Pr(Y = B|X = x) + \Pr(Y = C|X = x)} &= \alpha_A \\ \ln \frac{\Pr(Y = B|X = x)}{\Pr(Y = C|X = x)} &= \alpha_B + \beta_B x\end{aligned}$$

or

$$\Pr(Y = A|X = x) = \frac{\exp(\alpha_A)}{1 + \exp(\alpha_A)}$$
$$\Pr(Y = B|X = x) = \frac{1}{1 + \exp(\alpha_A)} \frac{\exp(\alpha_B + \beta_B x)}{1 + \exp(\alpha_B + \beta_B x)}$$
$$\Pr(Y = C|X = x) = \frac{1}{1 + \exp(\alpha_A)} \frac{1}{1 + \exp(\alpha_B + \beta_B x)}.$$

While this example looks very complicated, but with the help of introducing two auxiliary variables, it can be treated by estimating two separate binary logit models.

In a recent article published by *Political Analysis*, Paolino (2020) argues that instead of focusing on the statistical significance coefficients, researchers should focus on the statistical significance of changes in the predicted probabilities depending on the values of an independent variable, because such changes can be statistical significant, even though certain coefficients are not statistically significant. The previous considerations indicate however that this recommendation can be misleading, because if one assumes a multinomial logit model, either all or none of outcome probabilities are unaffected by the values of the independent variables. In fact, inferences drawn from the (apparent) statistical significance or insignificance of probability differences for different values of an independent variable are fallacious for similar reasons as inferences drawn from Berry interactions discussed in the previous section.

Due to the mean value theorem of differential calculus we have (using the notation $\pi_A(x) = \Pr(Y = A|X = x)$)

$$\begin{aligned}\Pr(Y = A|X = x_2) - \Pr(Y = A|X = x_1) &= (x_2 - x_1)\frac{\pi_A(x_2) - \pi_A(x_1)}{x_2 - x_1} \\ &= (x_2 - x_1)\left.\frac{\partial \pi_A(x)}{\partial x}\right|_{x=x^*}\end{aligned}$$

for $x_1 \leq x^* \leq x_2$. Thus the difference in probabilities increases with the absolute value of $x_2 - x_1$.

Of course, the values of the independent variables may be non-arbitrary, e.g. correspond to the treatment and control condition in an experiment as discussed in Gelpi (2017), so that the inflation or deflation of a probability difference by the factor $x_2 - x_1$ may also be non-arbitrary. Therefore a closer inspection of the partial derivative may be in order.

In the three outcome categories example discussed previously we have (using the notation $\pi_B(x) = \Pr(Y = B|X = x)$ and $\pi_C(x) = \Pr(Y = C|X = x)$)

$$\begin{aligned}
\frac{\partial \pi_C(x)}{\partial x} &= \frac{\partial}{\partial x} \frac{\exp(\alpha_C + \beta_C x)}{1 + \exp(\alpha_B + \beta_B x) + \exp(\alpha_C + \beta_C x)} \\
&= \frac{\exp(\alpha_C + \beta_C x)}{1 + \exp(\alpha_B + \beta_B x) + \exp(\alpha_C + \beta_C x)} \beta_C \\
&\quad - \frac{\exp(\alpha_C + \beta_C x)}{[1 + \exp(\alpha_B + \beta_B x) + \exp(\alpha_C + \beta_C x)]^2} [\exp(\alpha_B + \beta_B x)\beta_B + \exp(\alpha_C + \beta_C x)\beta_C] \\
&= \pi_C(x)[1 - \pi_C(x)]\beta_C - \pi_B(x)\pi_C(x)\beta_B.
\end{aligned}$$

From this we can conclude that the marginal effect of X on the probability that $Y = C$ is non-zero in a multinomial logit model except for particular values of X, unless *both* coefficients β_B and β_C are zero. Thus, within the framework of a multinomial logit model, a marginal effect on the probability of a particular outcome will almost always be non-zero, unless the independent variable in question does not have influence on any of the outcome categories.

That an independent variable affects either all or none of the categories of the dependent variable in a multinomial logit model may seem to be a strong restriction. However, one could argue that it would be quite unusual if one specific category differs from the other categories in being affected by an independent variable, if these categories are not viewed as a pre-existing group. To draw an analogy to modelling the influence of independent variables on a numeric dependent variable, restriction would be comparable to a situation where the likelihood of values within a certain range of the dependent variable remains unaffected.

The previous considerations show that it is not a good idea to focus on predicted probability changes of individual categories of a dependent variable if a multinomial logit model is used and assumed to be correctly specified – in direct contrast to Paolino's 2020 recommendations. He does make a valid point however, if he suggest that for inference one should not focus on the statistical significance of individual coefficients. We already saw that results with respect to the size and statistical significance may depend crucially on the choice of the response category. In particular if the number of independent variables in a multinomial logit models is large, the number of coefficients not only makes a discussion of individual coefficients impractical, it also creates problems for unbiased hypothesis testing. On the one hand, nominal p-values may become anti-conservative if hypotheses about several coefficients are tested simultaneously. On the other hand, straightforward Bonferroni corrections will be incorrect, because the estimates of the logit coefficients and p-values are not stochastically independent from one another. This is problem is not unlike the problem of dummy coeffi-

cients in the context of linear regression. In that context, the best strategy to test hypotheses about the influence of independent variables is to use model comparison F-tests. In case of multinomial logit models the appropriate technique will be to compare models using likelihood ratio tests.

3 Conclusion

The present paper discusses recent articles that recommendations that advise against basing the assessment of the influence of independent variables in models for categorical dependent variables. Berry et al. (2010) argue that coefficients of product terms should not be taken as criteria for the existence or absence of interaction effects. Instead, interaction effects should be assessed based on the statistical significance of second-order differences of predicted probabilities, because even though these second-order differences are almost always non-zero due to the "compression" of probabilities into the range from zero to unity, their existence or non-existence can be of theoretical relevance. Paolino (2020) argues that individual coefficients of multinomial logit models should not be used as the basis of inference about the influence of independent variables, because their values depend on the choice of the baseline category of the dependent variable. Instead, inferences should be based on the statistical significance of differences of predicted probabilities.

Both recommendations are misleading. First or second differences of predicted probabilities are neither general aspects of binary logit or probit models nor of multinomial logit models. A more principled reason is that predicted probabilities, their differences or derivatives are derived quantities of a model but not their parameters. A more practical reason is that they are not general aspects of statistical model but depend on particular values of the independent variables.

With regards to interaction effects, the preferable recommendation, in order to avoid conceptual inconsistencies and ambiguities in particular applications, is to take the coefficients of product terms as criterion. With regards to coefficients of multinomial logit models, the preferable recommendation is neither to focus on individual coefficients nor on differences of predicted probabilities, but on appropriate multi-parameter hypothesis tests, such as likelihood ratio tests.

References

Berry, William D., Jacqueline H. R. DeMeritt and Justin Esarey, 2010. "Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?" *American Journal of Political Science* 54(1): 248–266.

Berry, William D., Jacqueline H. R. DeMeritt and Justin Esarey, 2016. "Bias and Overconfidence in Parametric Models of Interactive Processes*". *American Journal of Political Science* 60(2): 521–539.

Frant, Howard, 1991. "Specifying a Model of State Policy Innovation". *American Political Science Review* 85(2): 571–573.

Gelpi, Christopher, 2017. "The Surprising Robustness of Surprising Events: A Response to a Critique of "Performing on Cue"". *Journal of Conflict Resolution* 61(8): 1816–1834.

Hastie, T. J. and R. J. Tibshirani, 1990. *Generalized Additive Models*. Boca Raton et al.: Chapman and Hall.

Manski, Charles F., 1975. "Maximum Score Estimation of the Stochastic Utility Model of Choice". *Journal of econometrics* 3(3): 205–228.

McCullagh, P. and J.A. Nelder, 1989. *Generalized Linear Models*. Monographs on Statistics & Applied Probability. Boca Raton et al.: Chapman & Hall/CRC.

Nagler, Jonathan, 1991. "The Effect of Registration Laws and Education on U.S. Voter Turnout". *American Political Science Review* 85(4): 1393–1405.

Paolino, Philip, 2020. "Predicted Probabilities and Inference with Multinomial Logit". *Political Analysis* : 1–6.

R Core Team, 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL `https://www.R-project.org/`

Rainey, Carlisle, 2016. "Compression and Conditional Effects: A Product Term Is Essential When Using Logistic Regression to Test for Interaction". *Political Science Research and Methods* 4(3): 621–639.

Ruppert, David, M. P Wand and R. J Carrol, 2003. *Semiparametric Regression*. Cambridge: Cambridge University Press.