Not-So-Average After All:
Individual vs. Aggregate Effects in Substantive Research

MARIUS RADEAN* and ANDREAS BEGER†

ABSTRACT

In nonlinear models, the effect of a given variable cannot be gauged directly from the associated coefficient. Instead, researchers typically compute the average effect in the population to assess the substantive significance of the variable of interest. Based on the *average* response, analysts often make policy recommendations that are to be implemented at the *individual* level. Such extrapolations, however, can lead to gross generalizations or incorrect inferences. Particularly when cases carry special meaning (e.g., countries), the political and socioeconomic relevance of research findings should be assessed at the individual level. Put differently the real-world applicability of the average effect is often limited. This paper outlines the conditions under which aggregation to mean is problematic, and advocates for a case-centered approach to model evaluation. Our approach allows researchers to draw more meaningful inferences, and makes the connection between research and practical applications more realistic.

*Department of Government, University of Essex.

†Independent Scholar.

1 Introduction

In nonlinear models, the effect of a given factor cannot be gauged directly from the associated coefficient. Instead, researchers typically compute the average effect in the population to assess the substantive significance of the variable of interest. Based on the average response, analysts often make policy recommendations that are to be implemented at the individual level.[1] As an illustration, if economic development is found to have—on average—a negative effect on civil war onset, the recommendation to a country seeking to stave off civil war would be to implement sound economic policies. However, policy recommendations for *individual* cases based on the *average* response can lead to gross generalizations or incorrect inferences. Specifically, the mean may obscure a large variation in individual effects, in which case the real-world applicability of the average effect is limited. Correctly interpreting the average effect may prevent unwarranted extrapolations, but does not solve the problem of the lack of practical relevance. We outline the conditions under which aggregation to mean is not advised, and advocate for a case-centered approach to model evaluation. Particularly when cases carry special meaning (e.g., countries, supreme court justices), we advise researchers to compute and report the effect for each case in the data. Only by seeing the full spread of cases can the reader assess how well the average summarizes the population. In sum,

[1] To provide one example, Graham, Miller and Strøm (2017) explicitly link their results to policy recommendations (e.g., "our findings have policy implications for democracy promoters and international peacekeepers" p. 687), and name half a dozen countries to which the study's findings should apply. The policy advice, however, is based on the average population effect (no individual effects are reported). The authors also extrapolate to cases not included in the estimation sample (e.g., democratic Burma). For these cases, it is not even possible to estimate individual responses since the necessary information does not exist in the data (i.e., the value of relevant covariates). In the upcoming *Democratic survival in post-civil war settings* section, we discuss in more detail how closely the average value reflects the underlying cases for one of their analyses.

our approach allows researchers to draw more meaningful inferences, and makes the connection between research and practical applications more realistic.

Translating research results into meaningful practical recommendations is not straightforward. For example, Ward, Greenhill and Bakke (2010) warn about the perils of "policy by p-value," that is, a policy informed by predicted effects that are statistically significant but too small to be substantively meaningful. In this paper, we caution against the perils of policy by average effect. There are both practical and substantive concerns to automatically using the mean to summarize the population. Practically, there are distributional forms that render measures of central tendency less useful (e.g., uniform, bimodal, skewed distribution). Distributional form aside, "the concept of a 'central value' is less meaningful" when the observations are spread out, "in which case no single central value can be representative" (Gelman and Hill, 2007, 467). These conditions are common in political science research. Many social science outcomes follow non-normal distributions (e.g., the probability of civil war onset, the household income in developed countries), and the dispersion of individual responses (e.g., predicted probabilities, marginal effects) has no predetermined pattern and may vary from one analysis to the next.

Substantively, focusing solely on the average response is problematic when our cases are of immediate interest. Using Hanmer and Kalkan's (2013, 267) taxonomy, we have two types of data:

> With large n studies, such as those based on survey data, [...] any individual case is anonymous and does not carry any special meaning. With studies based on a legislative body, a set of countries, and so on, the individual cases are entities in which we might have a special interest. For example, we might want to predict what could have influenced the probability that Hillary Clinton (in her time as a senator) would have supported an immigration bill, the number of environmental regulations Ireland will enforce, which countries are likely to fall into a civil war [...].

When cases carry special meaning, the practical relevance of our findings (i.e., the social, political, or economic significance) should be assessed at the individual level. Discussions of specific cases, in turn, ought to be based on what the model says about that particular case rather than on conjectures derived from the average response.

These concerns are laid bare in nonlinear model analyses where individual cases' characteristics play a big role. Specifically, in models with a limited dependent variable (e.g., binomial, ordered, or multinomial logit, etc.), the idiosyncratic features of individual observations have a significant impact on the magnitude of estimated effects. The higher the heterogeneity of the cases in the data, the less meaningful the average effect is. Some may argue researchers are well aware of the non constant nature of effects in nonlinear models, and of the fact that extrapolations from the population average to individual cases are tenuous. This knowledge, however, is not reflected in practice. Most studies report only the average effect, which obfuscates the underlying distribution of cases. At the very least, the current practice places an undue burden on nontechnical readers who are expected to know about classes of models associated with non constant effects.[2]

Consider the real data results from a logistic regression on civil war onset, which we discuss in more detail below. The average probability of war onset is 1.6% (1.3, 1.9), and the average effect of doubling GDP per capita is -0.3 percentage points (-0.5, -0.1). The numbers in parentheses indicate the respective 95% confidence intervals (CIs). What can we infer from these quantities of interest? The low onset probability suggests that civil wars are unlikely to occur, and thus countries should not be particularly concerned. Similarly, doubling a country's GDP, which is no small feat, does not seem effective in preventing violent civil unrest. Arguably, a risk reduction of less than half of a percentage point (i.e., from 1.6% to 1.3%), does not justify acting in practice.

These conclusions are hard to square with the extensive literature on civil wars, which argues they are the most common type of conflict, are very costly, and should not to be ignored (Pettersson, Högbladh and Öberg, 2019). At the domestic level, their destructive nature is reflected in the high death toll for the combatants, as well as the high number of terrorist attacks on civilians (Lacina, 2006; Polo and Gleditsch, 2016). At the international level, peacekeeping and state-building efforts

[2] Along these lines, King, Tomz and Wittenberg (2000, 348) also argue that "students, public officials, and scholars should not need to understand phrases like 'coefficient,' 'statistically significant,' and 'the 0.05 level' to learn from the research."

entail high financial and personnel costs (Fortna, 2004; Gilligan and Stedman, 2003).

The interest of scholars and policy-makers in understanding and preventing civil wars starts to make sense only when we unpack the average response and consider individual cases. Specifically, there are countries with a probability of civil war onset higher than 50%, for which improved economic performance can reduce the risk by up to 10 percentage points. Indonesia in 1950 is one such example. The onset probability for this case is 53% (significantly higher than the 1.6% population average), while the expected decrease in the probability of civil war onset is 9 percentage points (compared to the -0.3 average value). It is precisely the countries on the brink of civil, such as 1950 Indonesia, which are the focus of international efforts. Thus, there are contexts where the interest lies not with the average response, but with cases otherwise labeled as "outliers." In these contexts, policy initiatives based on the average effect are counterproductive.

For analyses where aggregation to mean is problematic, we advise a case-centered approach. Examining the quantity of interest at the case level provides greater leverage and helps avoid two common problems. First, depicting a wide range of effects by a single value can lead to gross generalizations or incorrect inferences. By gross generalizations we mean that the magnitude of individual effects is (severely) misrepresented by at least an order of magnitude. Incorrect inferences refers to cases where individual effects have the opposite sign than that of the average effect, or when subgroups exhibit opposite trends (e.g., the effect size increases with x for some cases, but decreases for others).

Second, aggregate effects may obscure the link between research and practical applications. The aforementioned average effect of GDP/capita on civil war onset, which indiscriminately pools stable and less stable countries, is of little practical relevance. In fact, its interpretation is as simple as it is uninformative. The 0.3 reduction in the onset probability indicates that, on average, doubling the GDP of *all countries* in the world reduces the *global risk* of civil war by less than half of a percentage point. Crucially, the average does not represent a reasonable value for the change in the *individual risk* of civil war of a given country, based solely on *its own* economic

performance. Yet this is the relevant information for practical applications, since governments do not control other countries' economic policies. The limited real-world applicability of the average effect brings into question its substantive significance.

2 Technical considerations

The mean value is not necessarily representative for the observations at large or specific subgroups. This is particularly the case for marginal effects in nonlinear models, where the effect magnitude is case-specific and can vary a lot without us realizing it.[3] We illustrate this point using logistic regression results from several Monte Carlo experiments. Our generic model is $y = \beta_0 + x\beta_x + z\beta_z + (x \times z)\beta_{xz} + \epsilon$, which allows us to have model specifications with varying degrees of complexity. x is the variable of interest and z represents the other covariates (if any). x and z are randomly drawn from a normal distribution, $\mathcal{N}(0, 1)$, and are independent, $\rho(x, z) = 0$. The coefficient of the intercept, β_0, is set to 0, and the error term ϵ is drawn from the standard logistics distribution with mean 0 and standard deviation $\pi/\sqrt{3}$. β_x equals 1 is all models specifications, whereas β_z and β_{xz} take on several values (including zero). Next we discuss several scenarios, determined by different combinations of coefficient values.

A bivariate case

Table 1, Model a shows the regression results for a bivariate model where x is the sole covariate (i.e., $\beta_z = 0$ and $\beta_{xz} = 0$). Using these estimates, in Figure 1a we plot the predicted probabilities (i.e., $\Pr(y = 1|x) = \Lambda(\beta_0 + x\beta_x)$, where Λ is the logistic link function) for all our 1,000 simulated observations. Since β_x is positive, the probability increases as x increases. Besides the individual probabilities (the solid dots), we also graph the sigmoid logistic curve, whose distinct shape explains the non constant effects. Specifically, the line is almost flat close to the two extremes of the

[3] In this paper, we use the terms marginal effect and discrete effect interchangeably. Technically, however, all the effects that we report in this study are due to a discrete increase in x.

Table 1: Logistic regression results with simulated data

	(a)	(b)	(c)	(d)	(e)
	$\beta_x = 1$	$\beta_x = 1$	$\beta_x = 1$	$\beta_x = 1$	$\beta_x = 1$
	$\beta_z = 0$	$\beta_z = \frac{1}{3}$	$\beta_z = 3$	$\beta_z = 3$	$\beta_z = \frac{1}{3}$
Regressor	$\beta_{xz} = 0$	$\beta_{xz} = 0$	$\beta_{xz} = 0$	$\beta_{xz} = 0$	$\beta_{xz} = 3$
x	1.22***	1.21***	1.33***	0.97***	1.24***
	(0.09)	(0.09)	(0.12)	(0.03)	(0.13)
z	—	0.47***	3.10***	2.97***	0.41***
		(0.08)	(0.20)	(0.06)	(0.12)
$x \times z$	—	—	—	—	3.24***
					(0.24)
Constant	0.07	0.05	−0.03	0.01	0.08
	(0.07)	(0.07)	(0.10)	(0.03)	(0.08)
Log likelihood	−563.25	−555.05	−341.18	−3501.10	−428.65
Number of observations	1000	1000	1000	10000	1000

* p < 0.10; ** p < 0.05; *** p < 0.01 (two-tailed test)

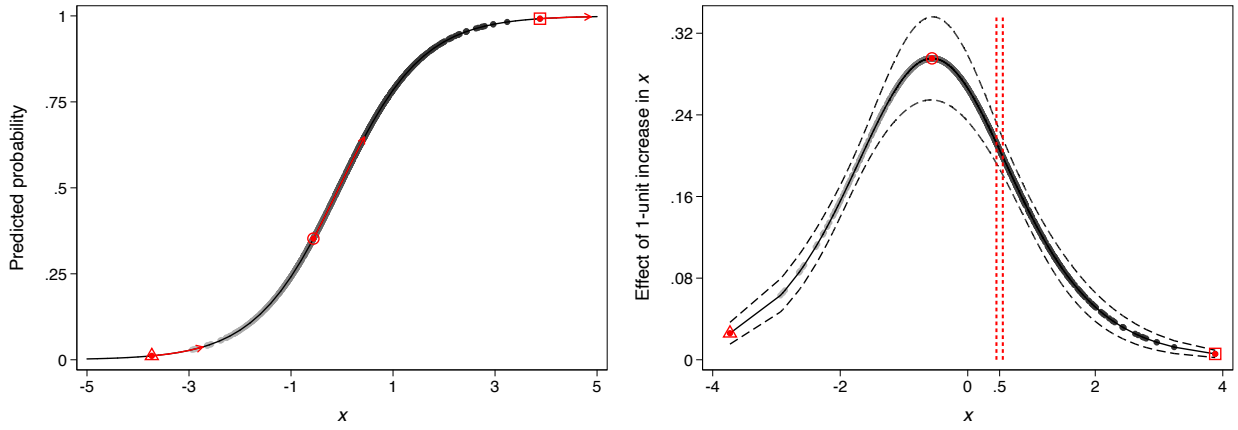
probability space (0 and 1). This in turn compresses the effects in those regions. For concreteness, consider the marginal effects associated with the three values of x highlighted in the plot: -3.73, -0.56, and 3.88. The arrows show the change in $\Pr(y)$, and the respective percentage point increase are 3, 30, and 1. Thus, conditional on the starting probability, the same unit increase can have dramatically different effects for individual cases.

Figure 1b shows the effect of a 1-unit increase in x for all observations.[4] For reference, the

[4] Following Hanmer and Kalkan's (2013) recommendation, we employ the observed-value approach to compute the effect of the counterfactual increase in x. This entails computing the effect for *each case* in the data, with the covariates set at their observed values. The mean of the individual effects is the average effect. The alternative, which until recently was the prevalent approach, is to compute the average case effect. This entails computing a *single* effect while the covariates are set at their mean value. Since it glosses over the differences between individual observations, the average case effect is often unrepresentative of many of the underlying cases (p. 268). In Online appendix A.1, we present the formulas for the two types of effect, as well as worked examples of

Figure 1: The effect of x and the impact of the logistic link function

(a) $\beta_z = 0;\ \beta_{xz} = 0$ (predicted probabilities) (b) $\beta_z = 0;\ \beta_{xz} = 0$ (marginal effects)



Note: In Figure 1a, the solid line is the S-shaped logit curve, and the solid circle marks indicate the individual predicted probabilities for all cases. In Figure 1b, the solid line indicates the average marginal effect and the dashed lines are the 95% CI. The solid circle marks indicate the individual marginal effects. The dotted vertical lines delineate the marginal effects when x is approximately 0.5. The predicted probabilities were computed using the `predict` command, whereas the marginal effects were computed using `margins, contrast` (StataCorp, 2023).

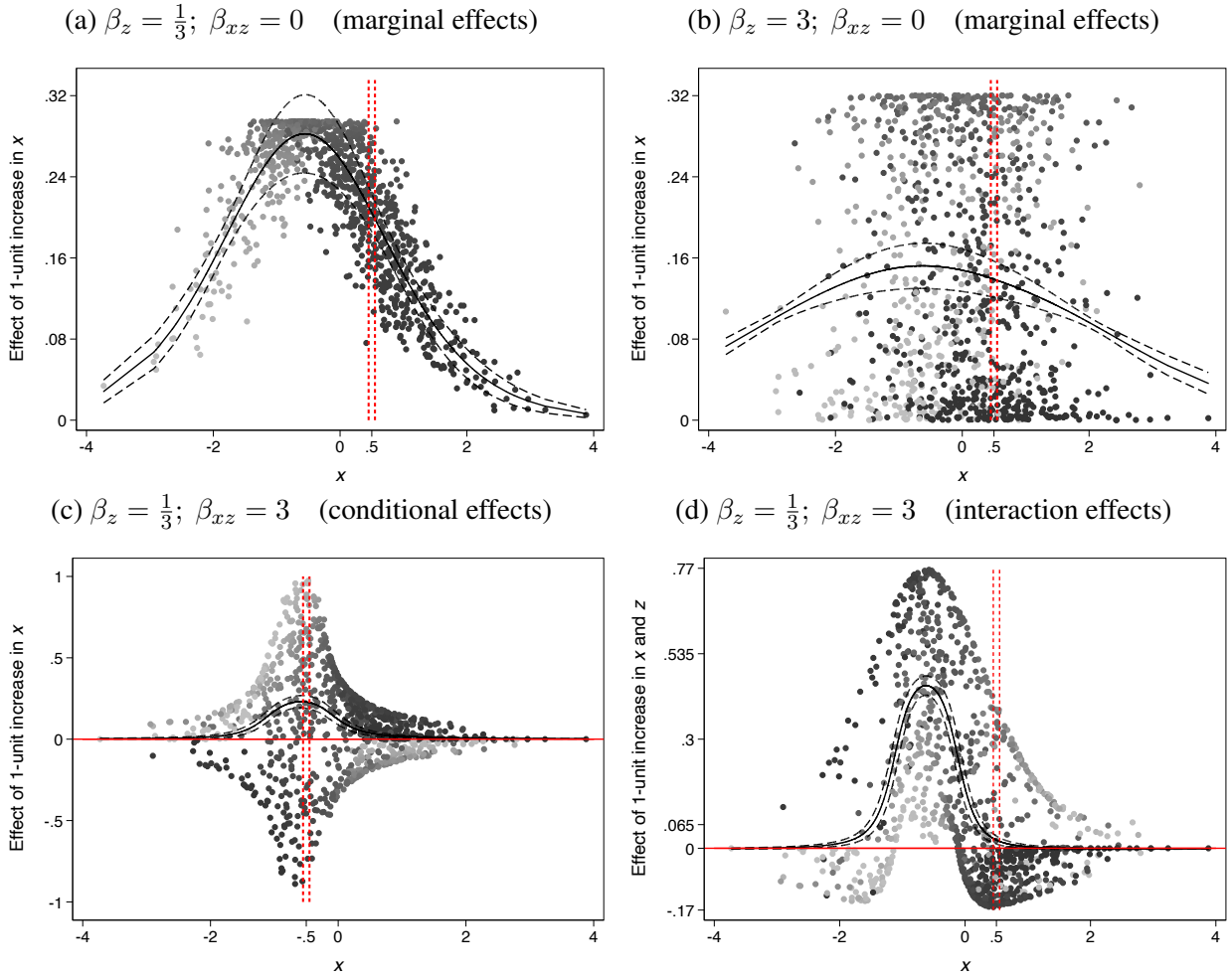
three cases discussed above are highlighted in this plot as well. Here the dots represent the 1,000 individual marginal effects. The color indicates where the baseline probability falls on the [0, 1] probability spectrum; the darker the hue, the closer to 1 the starting probability is. The solid line indicates the average effect and the dashed lines are the 95% CI. Since x is the sole covariate, there is no noise and the individual effects are nicely arranged along the average effect line. Because of compression, large marginal effects correspond to midrange initial probabilities, i.e., the gray dots. Conversely, the lighter or darker the hue, the smaller the marginal effects.

Multivariate and alternative model specifications

Model 1b shows the regression results from an additive multivariate model (β_{xz} = 0), where the input of the extra covariate is relatively small ($\beta_z = \frac{1}{3}\beta_x$). Using these estimates, Figure 2a plots the effect of a 1-unit increase in x for this scenario. Now the individual effects start to deviate from the average line, which is due to z also affecting the predicted probability. Model 1c shows the results for a scenario where the input of z is relatively large ($\beta_z = 3\beta_x$), and Figure 2b graphs the

how to compute the respective effects.

Figure 2: The effect of x conditional on the input of the other covariate z



Note: The solid lines indicate the average effect and the dashed lines are the 95% CI. The solid circle marks indicate the individual effects for all cases. The dotted vertical lines delineate the effects when x is approximately 0.5 (-0.5 in Figure 2c). The marginal and conditional effects were computed using the `margins, contrast` command (StataCorp, 2023), and the interaction effects using the `ginteff` software (Radean, 2023*a*).

associated marginal effects. In this scenario, the individual effects do not follow the average line and are dispersed without a discernible pattern. As before, small effects correspond to low or high initial probability (see the cluster of light and dark dots at the bottom of the graph). The difference is that the value of x is no longer the determinant factor for the initial probability.

The average effect deflation as β_z increases (i.e., the effect line in Figure 2b is flatter compared to the one in Figure 2a), is due to the *vertical* variation in marginal effects. Specifically, for the same x value, a 1-unit increase can lead to a change in y ranging from virtually 0 or 32

percentage points. As an illustration, let us consider the marginal effects when x is approximately 0.5, or more precisely $x \in [0.45, 0.55]$.[5] 35 observations fall within this range, which is delineated in the graphs by dotted vertical lines. Table 2 reports, by scenario, the standard deviation and range of effects for the entire sample as well as the narrow interval. The smaller the variation and range, the better the average describes the population.

When x is the only covariate ($\beta_z = 0$), the difference between the minimum (min) and maximum (max) effects in the entire sample is 28.9 percentage points (see the first row in Table 2). By contrast, the difference is roughly 0 when $x \simeq 0.5$, and so is the standard deviation. This means that the variation in marginal effects is mainly horizontal (i.e., across the range of x), with virtually no vertical variation (i.e., for a given value of x). For the $\beta_z = 3\beta_x$ scenario, however, the variation and range of effects at $x \simeq 0.5$ are virtually the same as those for the entire sample. Since the spread of effects is very similar, is makes as much sense to talk about a single effect when $x \simeq 0.5$, as would be to presume that the effect of x is constant across its entire range, [-3.73, 3.88]. Of course, assuming constant effects in nonlinear models is problematic.

What explains the vertical variation in marginal effects at $x \simeq 0.5$? Since the value of x is fixed, the variation is driven by the values of z. Recall that z represents the other model covariates, or the case-specific characteristics that discriminate between observations with the same x value. Since the average effect for a given x value conceals the vertical variation, analysts implicitly assume that cases are similar on all other characteristics. In this sense, the average effect retains some of the flaws associated with the average case approach to computing substantive effects.

So far, all our simulations had the same sample size, i.e., N = 1,000. This allows us to attribute any changes in the results to β_z taking different values, since this is the only difference across scenarios. But are the findings driven by having a relatively small sample? In other words, does the average value become more representative asymptotically? To answer this question, we

[5] We look at a narrow range rather than a fix value (i.e., $x \simeq 0.5$ vs. $x = 5$), because there are no two observations with the same value of x.

Table 2: The effect of a 1-unit increase in x on y

		Individual effects					Average vs. individual effects (%)	
Scenario	x range	Obs.	Mean	Std. Dev.	Min	Max	Within CI	Diff. Sign
$\beta_z = 0; \beta_{xz} = 0$	$[-3.73, \ 3.88]$	1,000	0.215	0.073	0.006	0.295	100	0
	$[\ 0.45, \ 0.55]$	35	0.206	0.004	0.199	0.212	100	0
$\beta_z = \frac{1}{3}; \beta_{xz} = 0$	$[-3.73, \ 3.88]$	1,000	0.211	0.074	0.005	0.295	54.30	0
	$[\ 0.45, \ 0.55]$	35	0.215	0.041	0.107	0.279	25.71	0
$\beta_z = 3; \beta_{xz} = 0$	$[-3.73, \ 3.88]$	1,000	0.141	0.113	0.000	0.320	6.00	0
	$[\ 0.45, \ 0.55]$	35	0.161	0.114	0.001	0.320	2.86	0
$\beta_z = 3; \beta_{xz} = 0$	$[-3.75, \ 4.04]$	10,000	0.107	0.084	0.000	0.239	2.87	0
	$[\ 0.45, \ 0.55]$	359	0.109	0.084	0.000	0.239	3.34	0
$\beta_z = \frac{1}{3}; \beta_{xz} = 3$	$[-3.75, \ 4.04]$	1,000	0.094	0.304	-0.890	0.974	7.30	31.80
$\left(\frac{\Delta \Pr(y)}{\Delta x}\right)$	$[-0.45, -0.55]$	29	0.296	0.524	-0.735	0.974	3.45	27.59
$\beta_z = \frac{1}{3}; \beta_{xz} = 3$	$[-3.75, \ 4.04]$	1,000	0.148	0.263	-0.162	0.766	6.60	42.80
$\left(\frac{\Delta^2 \Pr(y)}{\Delta x \Delta z}\right)$	$[\ 0.45, \ 0.55]$	35	0.063	0.196	-0.162	0.382	2.86	54.29

conduct an additional simulation for the most problematic case (i.e., $\beta_z = 3\beta_x$), while increasing the number of observations from 1,000 to 10,000. The regression results for this simulation are presented in Model 1d, and the variation in individual effects is detailed in the fourth row of Table 2. As it turns out, having more observations does not alleviate the problem. Once again, the variation and range of effects at $x \simeq 0.5$ are indistinguishable from those for the entire sample. Thus, collecting more data is not a solution to the problem we highlight in this paper.

Lastly, we consider a scenario entailing interaction effects, for which $\beta_z = \frac{1}{3}$ and $\beta_{xz} = 3$. The associated regression estimates are shown in Model 1e. When x and z interact in influencing $\Pr(y)$, there are two types of effects that may be of interest. The first is the conditional effect of x at specific levels of z. These are the estimates plotted in the standard conditional effect graph, which shows the effect of x over the range of z, $\frac{\Delta \Pr(y)}{\Delta x} = \Pr(y|x + n, z = [\min, \ldots, \max]) - \Pr(y|x, z = [\min, \ldots, \max])$. The other effect is the interaction effect associated with a discrete change in *both* x and z (Ai and Norton, 2003; Clark and Golder, 2023; Radean, 2023*a*). While

the conditional effect is the difference between two predicted probabilities (a first difference), the interaction effect is the double (or second) difference and entails four probabilities (Norton, Wang and Ai, 2004, 157), $\frac{\Delta^2 \Pr(y)}{\Delta x \Delta z} = \big[\Pr(y|x+n_x, z+n_z) - \Pr(y|x, z+n_z)\big] - \big[\Pr(y|x+n_x, z) - \Pr(y|x, z)\big]$, where n_* indicates the respective unit increase.

Figure 2c shows the conditional effect of a 1-unit increase in x over its entire range, while z is set at its observed values.[6] The first thing to note is that the individual effects occupy the entire space of plausible values, [-1, 1]. By contrast, the average effect is positive throughout and has a much narrower range, [0.003, 0.23]. This has important practical consequence in term of the average value's representativeness. Regardless whether we consider the full sample or the case when $x \simeq$ -0.5, we now have a significant share of individual effects with the opposite sign than that of the average effect (32% and 28%, respectively – see the last column of Table 2).

Figure 2d shows the effect of a 1-unit increase in both x and z, that is, the interaction effect. As with the conditional effects, the spread of individual effects is significantly larger than the one from additive models. In the full sample, 43% of individual effects have the opposite sign than that of the average effect, and a full majority when $x \simeq 0.5$, 54%. The reason why the average effect is positive despite a majority of individual effects being negative, is that the positive effects are large whereas the negative ones are small (they are all clustered just below the zero line). The fact that the average effect may have the opposite sign than the *majority* of observations, further challenges the idea that the mean value is necessarily representative for a large share of observations.

[6] We plot the conditional effect of x over its range rather than over z's, to keep the x-axis the same in all graphs. This makes the comparison between various scenarios more straightforward. For completeness, we present the conventional conditional effect plot (with z on the x-axis) in Figure B1 from Online appendix B. The figure illustrates that the reason some of the individual effects of x are negative, is because of the strong influence of z when z is negative.

Potential criticisms

Researchers often acknowledge the horizontal variation in effects (over the range of x), but not the vertical variation (the spread of effects due to the values of z). Faced with two sources of variation, some may argue that averaging on the vertical dimension is a necessary simplification. Were this the case, it is not immediately obvious why disaggregating on the horizontal dimension is a necessary complication. Technically, this information is not necessary to assess a hypothesis positing that x has a positive (negative) effect on y.[7] Moreover, as discussed above, the horizontal variation is not necessarily larger than the vertical variation. Thus, it is not a matter of prioritizing the source of the largest variation. Ultimately, if averaging is an acceptable cost of simplifying results presentation, why not take the average on both vertical *and* horizontal dimensions and report a single average effect? To clarify, we believe that researchers should acknowledge the variation in the effect of x. This reductio ad absurdum argument is to stress the need to acknowledge *all* sources of variation.

Others may point out that by using the average value as a fixed reference point, we neglect its variance. The argument would be that, while the point estimate of the mean does not match individual observations, the CI around it must contain a significant share of cases. Yet there is no theoretical reason to expect that it does. The CI captures uncertainty in the estimation of the average effect due to sampling uncertainty, not variation of cases around the mean. Other factors, such as sample size, also affect the CI width. The second to last column in Table 2 provides strong evidence for this. While all observations fall within the CI of the average effect when x is the sole covariate, that is no longer the case for the multivariate models. Specifically, the percent of

[7] Examining whether the effect magnitude changes over the range of x, is the test for a different hypothesis. Such a conditional hypothesis, however, should be driven by theory not the choice of empirical model. The following is not a theoretically derived hypothesis: "The effect of x on y varies with the values of x if we use a nonlinear model, but it is constant if we use a linear model."

observations within the CI is 54.3% when $\beta_z = \frac{1}{3}\beta_x$, and just 6% when $\beta_z = 3\beta_x$. The percent is in single digits for both the conditional and interaction effect scenarios (7.3% and 6.6%). Without the added case points in the plot, it is easy to treat the CI as describing the spread of cases.[8]

One may also argue that by de-emphasizing the average effect we place too much weight on outliers. First, thinking of observations in terms of outliers is theoretically problematic when cases are of individual relevance. Going back to the civil war example, should we dismiss countries with high risk of conflict as noise? Second, there are analyses where "outliers" are more representative for the cases at large than the mean. Using the data from the model where $\beta_z = 3\beta_x$, we calculate the percent of observations within one standard deviation below and above the mean, as well as the percent of observations within one standard deviation from the smallest and largest marginal effects. The number of observations within two standard deviations around the mean is almost six times lower than the number of cases within one standard deviation from the min and max marginal effects (5.4% vs. 29.7%). Thus, extreme values are not necessarily outliers.

3 Real data examples

Making inferences about individual cases based on the average response (under the assumption that the mean summarizes the population well), is not a theoretical problem but rather a practical issue with substantive implications. Using real data from three published studies, we showcase that reporting solely the average effect can lead to gross generalizations or incorrect inferences. The upcoming analyses are meant to exemplify some of the practical problems associated with the average effect, not to single out the respective studies.

[8] In Online appendix C, we provide additional evidence that the CI of the mean is not a good indicator of individual case variation. The appendix example employs a linear model to convey the point that this not an idiosyncratic feature of nonlinear models, but rather a more general issue.

3.1 Wealth and the risk of civil war

The first example reproduces Muchlinski et al.'s (2016) analysis of civil war onset, which we referenced in the introduction. The data consist of country-years between 1945 and 2000. The dependent variable, *Civil war onset*, is a dummy coded 1 when a civil war breaks out, and 0 otherwise. Following Muchlinski et al., we control for a host of potential determinants: $\ln$(*GDP/capita*) records countries' gross domestic product per capita, log-transformed; $\ln$(*Population*) indicates the size of the population; $\ln$(% *Mountainous terrain*) captures countries' terrain ruggedness; *Prior war* is coded 1 if a country had experienced another war in the past; *Non-contiguous territory* is a dummy variable identifying countries that have nonadjacent territories; *Oil exports*$/$*GDP* captures countries' reliance on oil as a proportion of the GDP; *New state* is a dummy identifying states that are less then two years old; *Political instability* is coded 1 if Polity records a change to 77 or 88 in the previous three years; *Democracy* is a regime dummy that equals 1 if the Polity score for a given country-year is a 6 or above; lastly, *Ethnic fractionalization* and *Religious fractionalization* are indexes that capture countries' heterogeneity on the respective dimensions.

Table 3 reports the results from a logistic regression. Using the coefficient estimates and covariates' observed values, we compute two of the most common substantive quantities of interest, that is, predicted probabilities and marginal effects (see Figure 3). Specifically, the solid line in Figure 3a graphs the average predicted probability and its 95% CI over the range of $\ln$(*GDP/capita*). Departing from the standard approach, we also plot the individual probabilities for all cases in the data. These are indicated by solid circle marks. There are a couple of points worth noting. First, the spread of individual probabilities is substantively wider than what the average value and its CI imply. While the average predicted probability is 1.6% (1.3, 1.9), the individual probabilities range from 0.1% to 53.2%. Moreover, the 95% CI of the average probability includes only 26% of the cases. Put differently, 74% of the observations lie outside of the mean's CI.

For the marginal effect plot, Figure 3b, we compute the effect of a 1-unit increase in $\ln$(*GDP/*

Table 3: The determinants of civil war onset

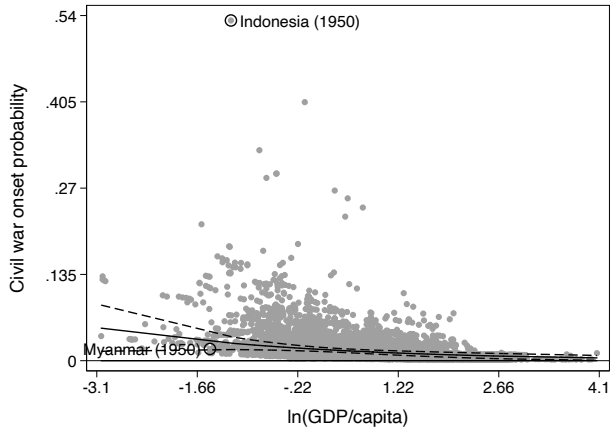
Regressor	Logistic regression
ln(GDP/capita)	-0.36 ***
	(0.12)
ln(Population)	0.20***
	(0.07)
ln(% Mountainous terrain)	0.18**
	(0.08)
Prior war	0.03
	(0.25)
Non-contiguous territory	0.22
	(0.28)
Oil exports/GDP	0.36
	(0.30)
New state	1.79***
	(0.32)
Political instability	1.30***
	(0.21)
Democracy	-0.37
	(0.28)
Ethnic fractionalization	0.44
	(0.42)
Religious fractionalization	0.58
	(0.51)
Constant	-8.42 ***
	(1.11)
Log likelihood	-528.65
Number of observations	7140

* p < 0.10; ** p < 0.05; *** p < 0.01 (two-tailed test)

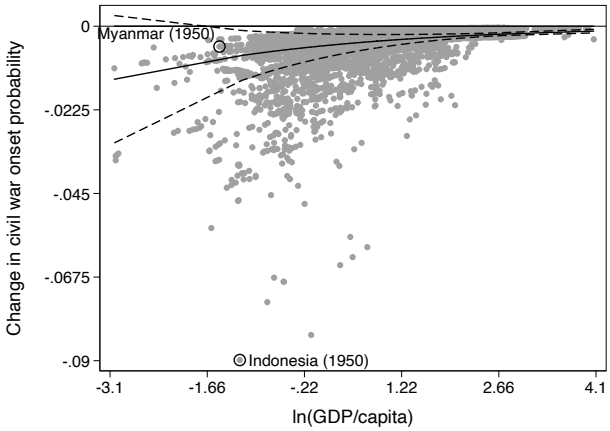
capita), while the other variables are set at their observed values. As an indicator of state capacity, GDP is one of the best predictors of civil war onset (its coefficient is large and statistically significant), and one that is actionable (in contrast to ethnic and religious fractionalization, terrain ruggedness, etc.). The average marginal effect is -0.5 percentage points (-0.7, -0.2). While statistically significant, this is not a substantively meaningful effect. Arguably, a decrease in the probability of war onset from 1.6% to 1.1%, does not warrant acting in practice. Notably, a 1-unit increase on the logged scale corresponds to tripling GDP per capita on the untransformed scale:

Figure 3: Conventional predicted probability and marginal effect plots

(a) Individual probabilities and the average probability with the 95% CI



(b) Individual marginal effects and the average marginal effect with the 95% CI



Note: The solid line indicates the average predicted probability (Figure 3a) or average marginal effect (Figure 3b), and the dashed lines are the 95% CI. The solid circle marks indicate the individual probabilities or marginal effects, respectively, for all cases.

$\ln(\textit{GDP/capita}) + 1 = \ln(\textit{GDP/capita}) + \ln(e^1) = \ln(\textit{GDP/capita} \times 2.7)$. Thus, the meager effect is not an artefact of a small increase in wealth. In terms of the number of cases within the 95% CI of the average effect, less than half of the observations are included (43%).

Substantive considerations

For any given value of wealth, there can be quite a spread in the quantity of interest; this is the vertical variation we discussed in the technical section. For instance, Indonesia in 1950 has the highest risk of civil war onset, 53%, as well as the largest marginal effect, -9 percentage points. Indonesia's high risk of civil war is not a deterministic outcome of its relative low level of wealth, but rather a consequence of its particular characteristics captured by the other covariates. Indeed, there are many cases that feature similar levels of wealth, for which the probability of war onset and the effect of wealth are near 0. For instance, in 1950, Myanmar was poorer than Indonesia. But its risk of civil war is about 0.5%, and the counterfactual 1-unit increase in $\ln$(GDP) decreases the risk by only 1.8 percentage points. These two cases are highlighted in Figure 3.

Given the wide variation in individual effects, the average marginal effect does not represent

a reasonable value for the change in the risk of civil war for any given country—not even *on average*. Specifically, it is not the case that if we were to repeatedly increase GDP in a long sequence of counterfactual experiments, the mean prediction for individual cases would be close to -0.5 (the population average). In fact, the mean prediction for our cases are the scattered values outlined in the marginal effect plot (e.g., -9 for 1950 Indonesia). An insidious problem stemming form a deflated mean is that it may lead policy makers to overlook an effective strategy simply because it is not optimal in *different* country contexts. A medical analogy would be the case of a patient forgoing a personalized treatment based on their genetic profile, because the effectiveness of that treatment in the population at large is relatively low.

The correct interpretation of the average effect may prevent unwarranted extrapolations, but does not solve the problem of the lack of practical relevance. Specifically, the worldwide propensity of civil war, which indiscriminately pools stable and less stable countries, is of little real-world relevance. For concreteness, the mean value represents the average change in the *global* risk of civil war, if *all* countries were to increase their wealth by the *same* amount. While convenient for practical calculations, such a concerted effort is not a realistic counterfactual. Moreover, it is not the case that one country can offset another country's (in)action. Specifically, the average risk would be different if two countries, A and B, increase their wealth by the same amount n, compared to when A's wealth increases by $2{\times}n$ while B's wealth stays the same. Ultimately, from the average effect we cannot infer a country's risk of civil war conditional solely on its economic performance. Yet this is the relevant information for practical applications, since governments do not have control over economic policies in other countries.

In sum, the average can obscure a large variation in the values that created that score. On the one hand, one cannot really criticize the mean for not showing the full spread of the underlying data points. On the other hand, we need to acknowledge that the mean does not always summarize the population well. This can be of significant relevance for practical case applications, especially when cases carry special meaning.

3.2 Wealth, weather shocks, and political violence

In the previous section, we showed that the mean value can lead to gross generalizations (i.e., the average response misrepresents individual effects by several orders of magnitude). More problematically, the average value can also lead to incorrect inferences (i.e., individual effects having the opposite sign than that of the average effect, or subgroups of cases exhibiting opposite trends). We illustrate this problem using a model of peace, repression, and civil war from Bagozzi et al. (2015), which in turn is a replication of Besley and Persson (2009).

The dependent variable, *Political Violence*, is a three-category ordinal variable. Specifically, it equals 0 when the level of violence is low, 1 when the government engages in violent repression, and 2 when both the government and rebels resort to violence (i.e., civil war). The explanatory variables are the same as in Bagozzi et al. (2015): $\ln$(*GDP/capita*) is the standard indicator for economic development; *Weather shock* is a count of the number of floods and heat waves that countries experience each year; *Parliamentary democracy* captures the institutional setup; *Large primary exporter* equals 1 if, in a given year, more than 10% of a country's GDP is generated by product exports, and 0 otherwise; lastly, *Large oil exporter* equals 1 if more than 10% of a country's GDP is generated by oil exports.

Table 4 reports the results from an ordered logistic analysis. A key finding in Besley and Persson (2009) is that wealth decreases the probability of violence. This finding is illustrated by the average effect plots in Figure 4, which show the effect of a 1-unit increase in $\ln$(*GDP/capita*) while holding the other covariates at their observed value. More specifically, the solid lines outline the mean effect and the dashed lines are the 95% CI. Specifically, Figure 4a indicates that, on average, wealth has a positive effect on the probability of peace. Conversely, increasing GDP decreases the probability of both repression, Figure 4b, and civil war, Figure 4c. In all three scenarios, the magnitude of the effect decreases as wealth increases.[9]

[9] The average effects in Figure 4a-4c are similar to those reported by Bagozzi et al. (2015, 742) in the first row of Figure 2. To keep things concise, we do not discuss their ZiOP results. However,

Table 4: The determinants of political violence

Regressor	Ordered logit
ln(GDP/capita)	-0.33 ***
	(0.06)
Weather shock	0.39***
	(0.05)
Parliamentary democracy	-1.02 ***
	(0.19)
Large oil exporter	1.54***
	(0.62)
Large primary exporter	-0.68
	(0.45)
τ_1	-1.59 ***
	(0.44)
τ_2	-0.09
	(0.44)
Log likelihood	-1435.44
Number of observations	1984

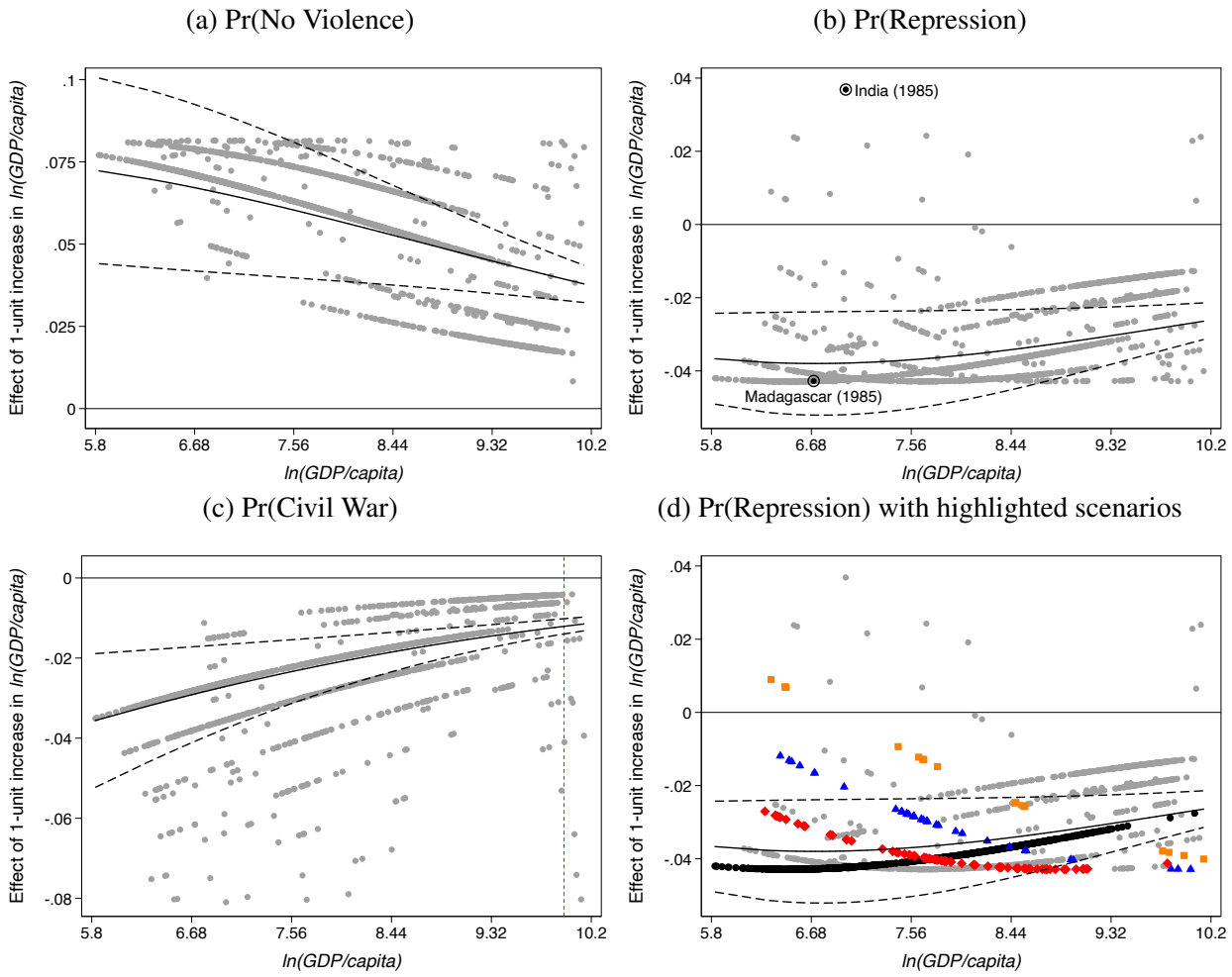
* p < 0.10; ** p < 0.05; *** p < 0.01 (two-tailed test)

Besides the average effect, we also plot the effect for all cases in the data (the gray dots). The wide variation in individual effects provides a more nuanced picture of the effect of wealth on political violence. As an illustration, let us consider the effect of wealth on the probability of civil war for ln(*GDP/capita*) values higher than 9.96. This is the area to the right of the vertical dotted line in Figure 4c. We focus on this wealth interval because the CI is at its narrowest, which would suggest that the uncertainty around the average value is low. The average effect for the 13 cases within this interval is -3.3 (-4.0, -2.5). But the observations are spread out (the range is [-8, -0.4] percentage points), and only one is contained by the average effect's 95% CI. What is more, the subsample's standard deviation is twice the size of the one for the entire sample, 2.58 vs 1.25. Importantly, we would not be aware of the magnitude of the vertical variation in individual effects, if we had plotted only the average effect.

The discrepancy between the average and individual effects is most striking in the repression

since the ZiOP estimates are still average effects, the issues highlighted here apply nonetheless.

Figure 4: The effect of ln(*GDP/capita*) on the probability of political violence



Note: The solid line indicates the average marginal effect and the dashed lines are the 95% CI. The solid marks indicate the individual effects for all cases in the data.

scenario, Figure 4b. While the average effect is always *negative*, there are cases with *positive* marginal effects at both low and high values of wealth. For a concrete example, consider India and Madagascar which had similar GDP/capita levels in 1985. Because of country specific attributes, increasing wealth *decreases* the risk of government repression in Madagascar by 4.3 percentage points. By contrast, the same economic boost *increases* the risk in India by 3.7 points. These diametrically different effects are both statistically significant ($p < 0.01$). More generally, for the level of wealth India had in 1985, the average effect is -3.8 percentage points. Given the opposite

direction, it is ill-advised to make policy recommendation for India based on the average effect.[10]

Opposite effects aside, there are subgroups of cases with a *different effect pattern* than the average trend. Figure 4d highlights four such subgroups. The cases in the respective subgroups differ solely in their experience with natural disasters (i.e., they have identical values on all other attributes). Recall that *Weather shock* is a count of the number of floods and heat waves that countries experience in a given year. As a reference point, the black dots indicate the cases that were spared by extreme weather events. Since these observations represent about 40% of the data, the average effect line closely follows their trajectory. For this subgroup, increasing GDP/capita reduces the probability of repression, and the magnitude of the effect *decreases* as wealth increases.

The average pattern, however, is not representative for the subgroup of countries that have experienced natural disasters. Specifically, in red with diamond marks are the countries that experienced 2 weather shocks; in blue with triangle marks are the countries with a score of 3; and, in orange with square marks are the countries with a score of 4. In contrast to the diminishing

[10] Ordered models allow for both monotonic and disparate marginal effects. The counterfactual increase in $\ln$(*GDP/capita*) has a monotonic effect with respect to the probability of the end categories. Specifically, it decreases the probability of the highest category (Civil War), and increases the probability of the lowest category (No Violence) for all cases. But the effect is not monotonic for the intermediary category (Repression). The probability for the interior category is a function of the threshold parameters delineating that category, and the cases' individual characteristics. Since the threshold parameters are fixed, it is the case-specific attributes that explain why the effect on repression is negative for Madagascar but positive for India. For completeness, let us look at the three scenario-specific effects for both countries. For Madagascar, the effect of increasing wealth on Pr(No Violence), Pr(Repression), and Pr(Civil War), are 0.071 (0.045 0.097), -0.043 (-0.058 -0.027), and -0.028 (-0.039 -0.017). For India, the respective effects are 0.044 (0.019 0.070), 0.037 (0.011, 0.063), and -0.081 (-0.108, -0.054). Notably, the three probabilities are linked, with the change in one probability being reflected in the probability of the remaining categories.

magnitude of the average effect, the size of the effect *increases* with the levels of $\ln$(*GDP/capita*) in countries affected by natural disasters. Moreover, the rate of change (i.e., the slope of the effect line) becomes steeper with the frequency of weather shocks. By looking solely at the average effect, we would be aware of neither opposite effects nor opposite trends in our data.

3.3 Democratic survival in post-civil war settings

In this section, we replicate an analysis from Graham, Miller and Strøm (2017) on the impact of inclusive institutions on democratic survival, *conditional* on states' experience with civil conflict. The dependent variable, *Democracy*, equals one if the executive and legislature are fairly elected, and a majority of adults have the right to vote. It is coded zero otherwise. Besides a number of control variables (*Ethno-linguistic fractionalization*, *Regional polity*, *GDP/capita*, *GDP growth*, *Fuel dependence*, *Population*, *Past democratic breakdowns*, and *Democracy age*), the analysis includes three two-way interactions between *Post-civil war* and *Inclusive*, *Dispersive*, and *Constraining powersharing*. The estimation model is a probit, and the regression results are presented in Table 4, Model 1 (p. 699). According to the authors, the findings show that constraining powersharing helps democratic survival generally, inclusive powersharing only in post-civil war settings, whereas dispersive powersharing may be detrimental for democracy.

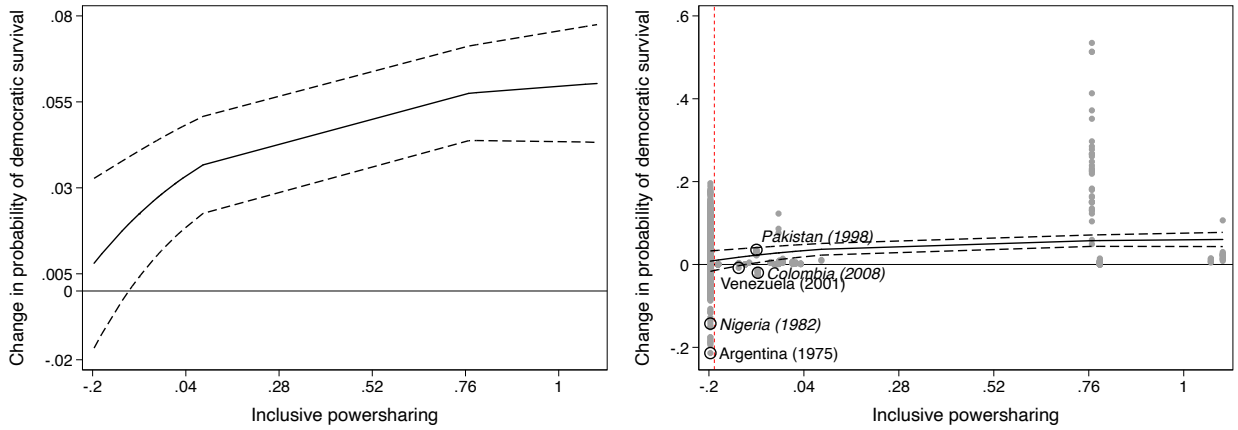
Figure 5a shows the average effect of postconflict status on the probability of democratic survival across the range of inclusive powersharing (from the 5th to the 95th percentile). More specifically, this figure shows the difference in the predicted probability of democratic survival for countries with and without a recently ended civil war, which the authors report in the top plot of Figure 3 (p. 700).[11] The difference in probability allows us to directly assess at what levels of inclusiveness the conditional effect is statistically significant. When examining the individual

[11] We were able to replicate Graham, Miller and Strøm's regression results perfectly (Table 4, Model 1, p. 699), but not the predicted probability estimates (Figure 3, p. 700). However, our results are close to theirs and the substantive inferences are the same.

Figure 5: The conditional effect of past civil war experience on democratic survival

(a) The average effect

(b) The average and individual effects



Note: The solid line indicates the average conditional effect and the dashed lines are the 95% CI. The solid marks in Figure 5b indicate the individual effects for the data cases. The the dotted vertical line delineates the below and above average inclusiveness levels.

probabilities side by side, we cannot tell whether the two are distinct when their CIs overlap (Gelman and Stern, 2006; Radean, 2023*b*). The average conditional effect is small and statistically insignificant at low levels of inclusive powersharing, but increases quickly with inclusiveness.

Figure 5b shows the same effect with the individual effects superimposed. As a reference point, we identify the Pakistan case that the authors reference. The dotted vertical line delineates the below and above average inclusiveness levels.[12] There are a couple of things worth noting. First, with the highest value of 6 percentage points, the average effect absconds substantively large effects in the population (min=-21, max=53). In fact, the putative null effect at low levels of inclusiveness (with a nominal level of less than 1 percentage point) is an artefact of large positive and negative effects (some as high as $|21|$ percentage points) canceling each other out. Second, 30% of individual effects have the opposite sign than that of the average. We highlight four such

[12] Taking the average level as a useful heuristic, the authors note that "within our sample, only two post-civil war democracies with an above average level of inclusive powersharing have broken down (Lebanon 1976, Pakistan 1999)" (p. 699). As a side note, the authors use the term 'average level' to indicate the median value not the mean.

observations, representing conflict-free and postconflict states (names in italics) at below and above average level of inclusive powersharing. Given the high share of observations that buck the trend, policy recommendations based on the average effect are tenuous.

On the basis of the average response in the population, the authors make policy recommendations for both domestic institution-builders and international organizations. More specifically, they name half a dozen countries to which the study's findings should apply, but no individual effects are reported. This includes out of sample cases, for which it is not even possible to estimate individual responses since we do not have the necessary information (i.e., the value of model covariates). As an example, the authors argue that, considering the efforts to "stabilize democratic transition in Burma," international actors should rethink their push for inclusive powersharing and decentralized institutions (p. 702). Yet Burma is a dictatorship for the entire timeframe of the study (1975 to 2010). Given the available information, we cannot use the model estimates to assess democratic Burma's odds of survival. While some country attributes are not affected by the democratization process (e.g., ethno-linguistic fractionalization), others are bound to change. At the very least, the constraining dimension of powersharing (i.e., the extent institutions constrain political leaders) would change as Burma transitions form dictatorship to democracy. Of course, one may impute not-yet-observed values using valuations of country experts, or by employing some other technique. But this would then be a forecasting exercise, which is not the case here.

4 Conclusion

To convey the substantive importance of their research findings, political scientists commonly compute the average effect in the population. There are instances, however, when a single value is too generic to be meaningful when applied to specific cases. In particular, the mean can obscure a large variation in individual effects, and may lead to gross generalizations or incorrect inferences. Crucially, we cannot tell a priori whether individual responses are close to the mean or dispersed, and often used heuristics can be misleading (e.g., the CI width of the average effect).

To improve empirical practice, we make two recommendations to substantive researchers:

1. **Compute the desired quantity of interest for all cases and report the full distribution.** Canned commands typically compute the mean response, making it easy to skip this step. But it is a necessary one to notice any discrepancies. In some situations discussing the average response may suffice. In others, a more personalized approached may be suited to highlight features obfuscated by the average trend. Ultimately, unless readers see the spread of cases, they cannot assess how well the average summarizes the data.[13]
2. **Do not extrapolate from the average response to individual cases.** While others have warned about the perils of policy by p-value (Ward, Greenhill and Bakke, 2010), we caution against the perils of policy by average effect. Because of case-specific characteristics, individual effects can be substantively different than the average effect. For instance, the magnitude of individual effects can differ by an order of magnitude or more. More problematically, case effects may exhibit different trends or have the opposite sign than that of the average effect. Indeed, we can have opposite effects in a variety of models and model specifications (e.g., binomial logit with interaction, ordered logit). Plotting the individual effects helps the analyst identify all such instances, without requiring prior knowledge about the specific features of various models.

The problem we highlight in this paper is not a small sample issue, and collecting more data would not necessarily help. Similarly, correctly interpreting the average effect may prevent unwarranted

[13] To help with the implementation, in Online appendix A we present the formulas for the average and individual-level quantities of interest discussed in the paper (i.e., predicted probabilities, marginal effects, conditional effects, and interaction effects). The companion replication file contains all the commands we used to compute these quantities of interest. It also includes the code for all figures, thus illustrating how one may graph these estimates. Since it contains the step-by-step command sequence (from model estimation to computing and graphing the aforementioned quantities of interest), the replication code is a template that other researchers may also use.

extrapolations, but does not solve the problem of the lack of practical relevance. Case in point, what practical recommendations can we draw from the finding that doubling the GDP of *all countries* in the world reduces the *global risk* of civil war by less than half of a percentage point?

Our recommendations are particularly relevant for analyses where cases carry special meaning. In such instances, discussions about practical or substantive significance ought to take place at the individual level. This type of analyses is common in political science. In international relations, the canonical data comprise countries extant at a given time. In American state politics, the unit of analysis are the fifty U.S. states. This is also the case in comparative politics research that focuses on party politics or legislative behaviour.

A good rule of thumb when deciding the appropriate level for model evaluation, is to think about the level at which policy recommendations make sense (e.g., individual, subgroup, population). If the study is about the effect of an initiative to increase voter turnout, we cannot and should not focus on the idiosyncratic responses of individual voters. In this case, it is the average response that is of immediate interest. Conversely, if we assess the effect of wealth on the likelihood of civil war onset, we are interested in individual responses—particularly those of countries on the brink of civil war. In this case, the average response is not particularly informative.

Lastly, we address a couple of possible concerns. One potential criticism is that researchers are well aware of the limitations of the average effect. If that were the case, making policy recommendations based on the average response would be the exception; yet it is the norm. Another concern is that case-level estimates are less reliable, since they are more sensitive to model specifications than the average value. That is a valid concern. However, using a robust but disparate average value, possibly with an opposite sign, is not the solution. The better approach is to conduct sensitivity analyses checking whether the case effect of interest is robust, and we urge researchers to do so. The appropriate robustness test for practical relevance is the effect strength stability, not the standard significance-robustness test. This means testing whether the estimated effect size is substantively different across various model permutations, rather than whether the estimated effect

keeps its sign and significance status (Neumayer and Plümper, 2017).

Some may also point out that not all researchers are interested in discussing specific cases. But analysts who present evaluations of substantive effects or policy recommendations, implicitly demonstrate an interest in making a connection between research and practical applications. Our recommendations help assess the distance between theory and practice, and make applications to specific cases more realistic. While rooted in technical considerations, our concerns are substantive in nature. In analyses where the mean is not a representative central value, its practical significance is brought into question. Due to a lack of substantive relevance, researchers have been advised against making big claims about small and practically inconsequential effects—even if they are statistically significant (Esarey and Danneman, 2015; Gross, 2015; Rainey, 2014). Pushing the argument further, we argue that the substantive significance concern extends to situations where researchers focus on non actionable average effects.

References

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

Bagozzi, Benjamin E., Daniel W. Hill, Jr., Will H. Moore and Bumba Mukherjee. 2015. "Modeling Two Types of Peace: The Zero-inflated Ordered Probit (ZiOP) Model in Conflict Research." *Journal of Conflict Resolution* 59(4):728–752.

Besley, Timothy and Torsten Persson. 2009. "Repression or Civil War?" *American Economic Review* 99(2):292–97.

Clark, William Roberts and Matt Golder. 2023. *Interaction Models: Specification and Interpretation*. New York, NY: Cambridge University Press.

Esarey, Justin and Nathan Danneman. 2015. "A Quantitative Method for Substantive Robustness Assessment." *Political Science Research and Methods* 3(1):95–111.

Fortna, Virginia Page. 2004. "Does Peacekeeping Keep Peace? International Intervention and the Duration of Peace After Civil War." *International Studies Quarterly* 48(2):269–292.

Gelman, Andrew and Hal Stern. 2006. "The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant." *The American Statistician* 60(4):328–331.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Gilligan, Michael and Stephen John Stedman. 2003. "Where Do the Peacekeepers Go?" *International Studies Review* 5(4):37–54.

Graham, Benjamin A.T., Michael K. Miller and Kaare W. Strøm. 2017. "Safeguarding Democracy: Powersharing and Democratic Survival." *American Political Science Review* 111(4):686–704.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman and Douglas G. Altman. 2016. "Statistical Tests, P Values, Confidence Intervals, and Power: a Guide to Misinterpretations." *European Journal of Epidemiology* pp. 1–14.

Gross, Justin H. 2015. "Testing What Matters (If You Must Test at All): A Context-Driven Approach to Substantive and Statistical Significance." *American Journal of Political Science* 59(3):775–788.

Hanmer, Michael J. and Kerem Ozan Kalkan. 2013. "Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models." *American Journal of Political Science* 57(1):263–277.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* pp. 347–361.

Lacina, Bethany. 2006. "Explaining the Severity of Civil Wars." *Journal of Conflict Resolution* 50(2):276.

Muchlinski, David, David Siroky, Jingrui He and Matthew Kocher. 2016. "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data." *Political Analysis* 24(1):87–103.

Neumayer, Eric and Thomas Plümper. 2017. *Robustness Tests for Quantitative Research*. Cambridge: Cambridge University Press.

Norton, Edward C., Hua Wang and Chunrong Ai. 2004. "Computing Interaction Effects and Standard Errors in Logit and Probit Models." *The Stata Journal* 4(2):154–167.

Pettersson, Therése, Stina Högbladh and Magnus Öberg. 2019. "Organized Violence, 1989?2018 and Peace Agreements." *Journal of Peace Research* 56(4):589–603.

Polo, Sara M.T. and Kristian Skrede Gleditsch. 2016. "Twisting Arms and Sending Messages: Terrorist Tactics in Civil War." *Journal of Peace Research* 53(6):815–829.

Radean, Marius. 2023*a*. "ginteff: A Generalized Command for Computing Interaction Effects." *The Stata Journal* 23(2):301–335.

Radean, Marius. 2023*b*. "The Significance of Differences Interval: Assessing the Statistical and Substantive Difference between Two Quantities of Interest." *The Journal of Politics* 85(3):969–983.

Rainey, Carlisle. 2014. "Arguing for a Negligible Effect." *American Journal of Political Science* 58(4):1083–1091.

Soyer, Emre and Robin M Hogarth. 2012. "The Illusion of Predictability: How Regression Statistics Mislead Experts." *International Journal of Forecasting* 28(3):695–711.

StataCorp. 2023. *Stata 18 Base Reference Manual*. College Station, TX: Stata Press.

Ward, Michael D., Brian D. Greenhill and Kristin M. Bakke. 2010. "The Perils of Policy by p-value: Predicting Civil Conflict." *Journal of Peace Research* 47(4):363–375.

Appendix A : Technical appendix

In the main text, we encourage researchers to report both the average and individual values of their quantity of interest, whatever that may be. In what follows, we review what this formally means for the four most common quantities of interest in political science applications (i.e., predicted probabilities, discrete (marginal) effects, conditional effects, and interaction effects). In the upcoming formulas, N is the number of observations, F is the link function, and $\mathbf{X}$ is a set of independent variables including the constant term. For researchers interested in how to compute these quantities of interest in practice, the companion replication file contains all the commands we used to compute and graph each of these estimates. In brief, we use the `predict` command to compute predicted probabilities, `margins,contrast` to compute marginal and conditional effects (StataCorp, 2023), and the `ginteff` software to compute interaction effects (Radean, 2023*a*).

Predicted probabilities

Formally, the individual and average predicted probabilities are defined as follows:

$$\text{IndPr}_i = F(\mathbf{X}_i\beta) \qquad \forall \quad i \in N \tag{A1}$$

$$\text{AvgPr} = \frac{\sum_{i=1}^{N} \text{IndPr}_i}{N}. \tag{A2}$$

Discrete (marginal) effects

The individual and average discrete effects are defined as follows:

$$\text{IndEff}_i = F(x_c\beta_x + \mathbf{X}_i\beta) - F(x_b\beta_x + \mathbf{X}_i\beta) \qquad \forall \quad i \in N \tag{A3}$$

$$\text{AvgEff} = \frac{\sum_{i=1}^{N} \text{IndEff}_i}{N}, \tag{A4}$$

where x_b is the base level of x, and x_c is the counterfactual value. The base level may be x's observed values (x_i), its mean ($\bar{x}$), or some other theoretically relevant level. $x_c = x_b + n_x$, with n_x being the respective n-unit increase in x.

Conditional effects

The individual and average conditional effects are defined as follows:

$$\text{IndCondEff}_i = F(x_c\beta_x + z_i\beta_z + \mathbf{X}_i\beta) - F(x_b\beta_x + z_i\beta_z + \mathbf{X}_i\beta) \qquad \forall \quad i \in N \tag{A5}$$

$$\text{AvgCondEff} = \frac{\sum_{i=1}^{N} \text{IndCondEff}_i}{N}. \tag{A6}$$

Interaction effects

The individual and average interaction effects are defined as follows:

$$\begin{aligned}\text{IndIntEff}_i = &\left[F(x_c\beta_x + z_c\beta_z + \mathbf{X}_i\beta) - F(x_b\beta_x + z_c\beta_z + \mathbf{X}_i\beta)\right] \\ &- \left[F(x_c\beta_x + z_b\beta_z + \mathbf{X}_i\beta) - F(x_b\beta_x + z_b\beta_z + \mathbf{X}_i\beta)\right] \qquad \forall \quad i \in N\end{aligned} \tag{A7}$$

$$\text{AvgIntEff} = \frac{\sum_{i=1}^{N} \text{IndIntEff}_i}{N}, \tag{A8}$$

where z_b is the base level of z, and $z_c = z_b + n_z$ is the counterfactual value.

A.1 The average case effect and the observed-value average effect

Discrete effects are simply differences in predicted probabilities. Formally, the average case probability and the observed-value average probability are defined as

$$\text{AvgCasePr} = F(\overline{\mathbf{X}}_i\beta) \tag{A9}$$

$$\text{AvgPr} = \frac{\sum_{i=1}^{N} F(\mathbf{X}_i\beta)}{N}, \tag{A10}$$

where $\overline{\mathbf{X}}$ indicates the average value of the respective variables, $\frac{\sum_{i=1}^{N} \mathbf{X}_i}{N}$.

For a practical example, let us say we have the following logistic regression: $y = \beta_0 + k\beta_1 + v\beta_2$. Thus, besides the constant, $\mathbf{X}$ contains two additional covariates, k and v. Given the chosen empirical estimator, the link function F is the cumulative standard logistic distribution $\Lambda(\beta\mathbf{X}) = \frac{1}{1+e^{-\beta\mathbf{X}}}$. Finally, let us assume that $N = 2$, $\beta_0 = 0.3$, $\beta_1 = 0.2$, and $\beta_2 = 0.4$. The independent variables' values for the two observations are: $k_1 = 1$ and $v_1 = 2$, and $k_2 = 3$ and $v_2 = 5$. Now we have all the information we need to compute the average case probability (AvgCasePr), and the observed-value average probability (AvgPr).

$$\begin{aligned}\text{AvgCasePr} &= F(\overline{\mathbf{X}}_i\beta) \\ &= \Lambda(\beta_0 + \bar{k}_i\beta_1 + \bar{v}_i\beta_2)\end{aligned}$$

$$
\begin{aligned}
&= \Lambda\left(\beta_0 + \frac{\sum_{i=1}^{2} k_i}{2}\beta_1 + \frac{\sum_{i=1}^{2} v_i}{2}\beta_2\right) \\
&= \Lambda\left(0.3 + \frac{1+3}{2} \times 0.2 + \frac{2+5}{2} \times 0.4\right) \\
&= \Lambda(0.3 + 2 \times 0.2 + 3.5 \times 0.4) \\
&= \Lambda(2.1) \\
&= 0.891
\end{aligned} \tag{A11}
$$

$$
\begin{aligned}
\text{AvgPr} &= \frac{\sum_{i=1}^{N} F(\mathbf{X}_i\beta)}{N} \\
&= \frac{\sum_{i=1}^{2} \Lambda(\mathbf{X}_i\beta)}{2} \\
&= \frac{\Lambda(\beta_0 + k_1\beta_1 + v_1\beta_2) + \Lambda(\beta_0 + k_2\beta_1 + v_2\beta_2)}{2} \\
&= \frac{\Lambda(0.3 + 1 \times 0.2 + 2 \times 0.4) + \Lambda(0.3 + 3 \times 0.2 + 5 \times 0.4)}{2} \\
&= \frac{\Lambda(1.3) + \Lambda(2.9)}{2} \\
&= \frac{0.786 + 0.948}{2} \\
&= 0.867
\end{aligned} \tag{A12}
$$

Thus, the average case probability and the observed-value average probability do not have the same value. This difference is reflected in all quantities of interest that are computed using predicted probabilities. As an illustration, let us compute the average case effect and the observed-value average effect, associated with a 1-unit increase in k from its observed values. Since discrete effects are differences in predicted probabilities, we can compute the two effects as follows:

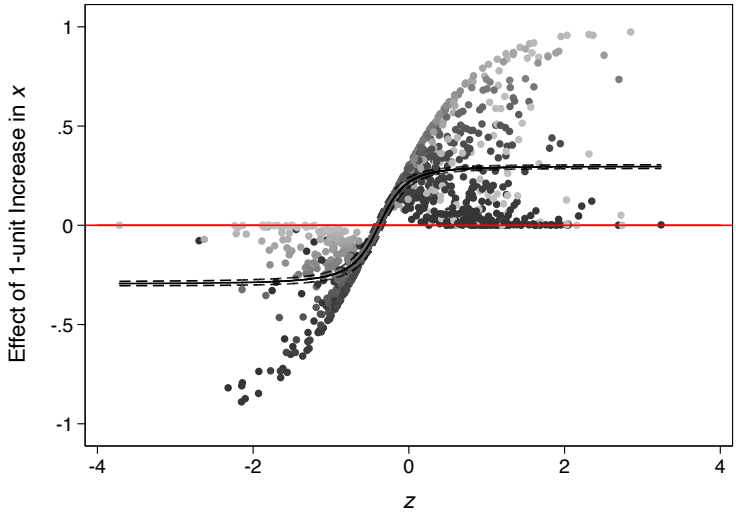
$$
\begin{aligned}
\text{AvgCaseEff} &= F((\bar{k}_i + 1)\beta_k + \overline{\mathbf{X}}_i\beta) - F(\bar{k}_i\beta_k + \overline{\mathbf{X}}_i\beta) \\
&= F((\bar{k}_i + 1)\beta_k + \beta_0 + \bar{v}_i\beta_v) - F(\bar{k}_i\beta_k + \beta_0 + \bar{v}_i\beta_v) \\
&= \Lambda(2.3) - \Lambda(2.1) \\
&= 0.909 - 0.891 \\
&= 0.018
\end{aligned} \tag{A13}
$$

$$
\begin{aligned}
\text{AvgEff} &= \frac{\sum_{i=1}^{N}[F((k_i+1)\beta_k + \mathbf{X}_i\beta) - F(k_i\beta_k + \mathbf{X}_i\beta)]}{N} \\
&= \frac{\sum_{i=1}^{2}[F((k_i+1)\beta_k + \beta_0 + v_i\beta_v) - F(k_i\beta_k + \beta_0 + v_i\beta_v)]}{2} \\
&= \Big\{[F((k_1+1)\beta_k + \beta_0 + v_1\beta_v) - F(k_1\beta_k + \beta_0 + v_1\beta_v)] + \\
&\qquad [F((k_2+1)\beta_k + \beta_0 + v_2\beta_v) - F(k_2\beta_k + \beta_0 + v_2\beta_v)]\Big\}/2 \\
&= \frac{[\Lambda(1.5) - \Lambda(1.3)] + [\Lambda(3.1) - \Lambda(2.9)]}{2} \\
&= \frac{0.032 + 0.009}{2} \\
&= 0.021
\end{aligned}
\tag{A14}
$$

Appendix B : The conditional effect of x on y over the range of z

In the main text, Figure 2c shows the conditional effect of x over its range (rather than over z's), to keep the x-axis the same in all graphs of Figure 2. Here we present the conventional plot, with Figure B1 showing the effect of a 1-unit increase in x over the range of z. While we hinted at this in Footnote 6, now it is easy to see that the sign of the individual conditional effects is largely determined by z (i.e., the effect of x is typically negative when z is negative, and vice versa).

Figure B1: The conditional effect of x over the range of z



Note: The solid line indicates the average conditional effect and the dashed lines are the 95% CI. The solid circle marks indicate the individual conditional effects for all cases. The color indicates where the baseline probability falls on the [0, 1] probability spectrum; the darker the hue, the closer to 1 the starting probability is.

Notably, the average effects reported in Figure 2c and B1 are not the same, even though they are both observed-value effects due to a 1-unit increase in x. The difference stems from the fact that we have different x-axis variables, each with a distinct range of values. An example may help illustrate the root of the discrepancy. In Figure 2c, the average effect when x is at its minimum, is the mean of the difference between the predicted probability when $x = \min$ and $x = (\min + 1)$, across the 1,000 observed values of z. In Figure B1, the average effect when z is at its minimum, is the mean of the difference between the predicted probability across the 1,000 observed values of x, and that associated with an increase of 1 in all 1,000 values of x, while $z = \min$ in both cases.

But the individual conditional effects are the same in both Figure 2c and B1. Consequently, the discussion from the main text about the significant spread of cases around the mean (as measured by the standard deviation and the difference between the min and max values), still applies. Lastly, while there are no individual effects with opposite sign, it is still the case that only a minority number of cases (22.6%) fall within the average effect's CI.

Appendix C : Variation in cases around the mean

In the main text, we emphasize the fact that the confidence interval of the average response is not necessarily a good indicator of individual case variation. Theoretically, there is no reason to expect that it would be. The CI captures the uncertainty in the estimation of the average response due to sampling uncertainty, not variation of cases around the mean. We illustrate this point further with a linear regression example. The example entails two alternative datasets which differ on two accounts: (i) the number of observations (i.e., N_1 = 500 and N_2 = 100), and (ii) the spread of the dependent variable, such that the outcome variable in the larger sample has a higher variance. Specifically, y_1 and y_2 are continuous variables with mean 15, but the standard deviation of y_1 is roughly twice as large as the one of y_2 (5 vs. 2.6). Both x_1 and x_2 are continuous variables drawn from a uniform distribution within the [-5, 5] range, and β_{x_1} = β_{x_2} = 0.5. The results of the two bivariate regressions are presented in Table C1.

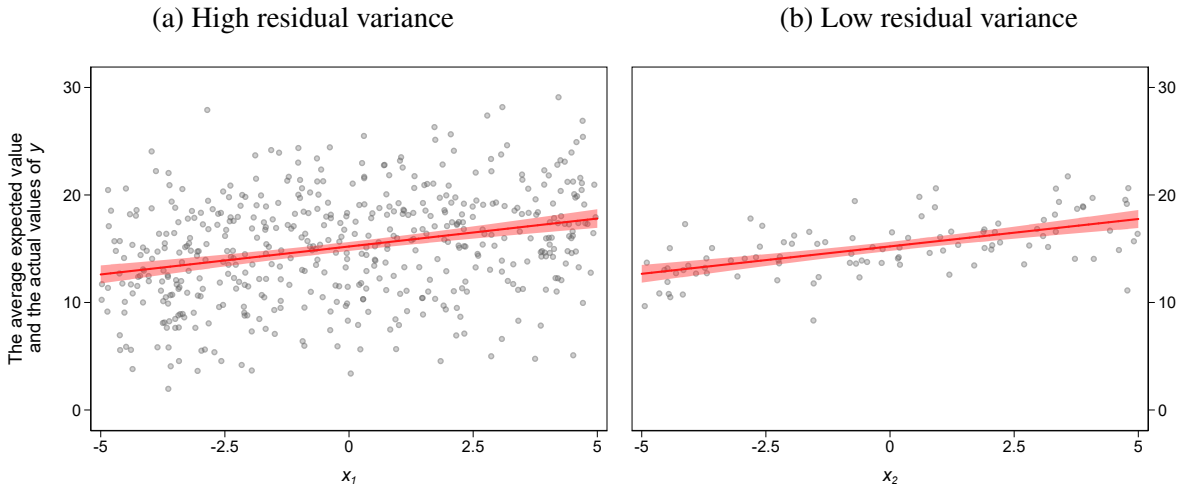
Table C1: Linear Regression Results with Simulated Data

	Scenario	
Regressor	y_1	y_2
x_1	0.52***	—
	(0.08)	
x_2	—	0.51***
		(0.07)
Constant	15.21***	15.22***
	(0.21)	(0.21)
Number of observations	500	100

* p < 0.10; ** p < 0.05; *** p < 0.01 (two-tailed test)

Using these estimates, in Figure C1a and C1b, we plot the average expected value (the solid line) with the 95% CI, as well as the actual y values for the individual cases (the gray dots). While the respective means and CIs are similar in both the high and low variance scenarios, the residual variation around the two group means is very different. Thus, inferring case variation from the

Figure C1: Residual variance and the CI width



Note: The solid line indicates the average expected value with the 95% CI. The dots indicate the actual values of y.

CI width of the average response is not warranted. In practice, however, the CI of the mean is often wrongly assumed to contain some large fraction of cases. Soyer and Hogarth (2012) show that even experts (i.e., academic economists) are prone to this kind of misinterpretation (see also Greenland et al., 2016). One can only assume that non-experts are even more likely to interpret the CI in this manner.