Five Shortcomings of Quantitative Peace Research

Most findings in quantitative peace scholarship are probably not real. Point estimates from quantitative studies often fail to convey the desired information researchers are seeking to estimate with a high degree of accuracy. Subsequent interpretations of point estimates are rarely given the interpretation they require. Oftentimes, neither the estimation nor interpretation of a statistical estimate is reflective of the population parameter a researcher is interested in evaluating. This is not a conclusion that is particularly damning for peace research, given that similar sentiments have been expressed for broader scientific scholarship (Ionnidis 2005, Colquhoun 2014). However, works such as these primarily discussed the presence of poor statistical power and the high false discovery rates of "effects" in scientific research. Undoubtedly, these problems are persistent in quantitative peace studies. However, issues of statistical power and the consequences of low power are well-established. While power analyses are rarely seen in published works within quantitative peace research, scholars are generally aware that small sample sizes create statistical issues that dramatically inhibit inferences from being drawn in quantitative studies. In many cases, researchers have little control over sample size matters, although disaggregated data sets are increasingly becoming more common, aiding high-powered peace research.

Unfortunately, low statistical power is far from the only methodological issue that plagues quantitative peace scholarship. Even if most quantitative peace studies were sufficiently powered, this chapter maintains that the results from such research are still probably not real. While some studies may estimate and interpret their parameter of interest with the necessary methodology (either by chance or an in-depth understanding of the necessary methodology), existing surveys and critiques of quantitative peace research and other social scientific literatures

suggest troubling patterns that call many "findings" and interpretations of statistical estimates into question (Ionnidis 2005, Colquhoun 2014, Aronow and Samii 2016, Keele et al. 2019, Hunermund and Louw 2020, Lundberg et al. 2021, Wysocki et al. 2022, Dworschak 2023). This article follows prior works critiquing methodology in quantitative peace research such as Achen (2005) and Schrodt (2014). In particular, five shortcomings are discussed, with associated solutions featured.

First, researchers tend to make causal claims (naturally extending from their developed causal theories) without supporting causal methodology. Findings are often interpreted as "effects", "explanations", or "influences" without the warranted pre- and post-estimation steps required to elevate the interpretation of purely statistical estimates to causal effects. Second, researchers often engage in the practice of mutual adjustment, where multiple hypotheses are tested within the same model with a "standard set of controls". Such a practice will likely yield estimates that do not sufficiently examine *any* hypothesis of interest. Third, given that most quantitative peace scholars work with multiple cross-sectional units over time, it is surprising that the biasing properties of panel data are often not accounted for beyond clustered robust standard errors. Fourth, researchers are incredibly reliant on regression adjustment for estimation, often without demonstrating sufficient understanding the serious limitations the method features. While regression adjustment may be necessary in many situations, methods that are better able to estimate the effects researchers are interested in are available at the researcher's discretion. Despite this availability, regression adjustment is almost habitually selected over alternative methods that may preform much better. Lastly, researchers often frame the quality and importance of their findings around small p-values or 90-95% confidence intervals that fail to overlap with zero. However, the dominant statistical inferential paradigm in the social sciences

from which p-values and confidence intervals derive – frequentist inference - is fundamentally not applicable to most areas of research within quantitative peace studies. Such a disconnect begs for an improvement in the current hypothesis testing paradigm in quantitative peace research.

Causal Claims Without Causal Methodology

Conflict management and peace studies overwhelmingly concern themselves with causal questions. What causes conflict? What are the consequences of conflict? How can conflict be managed and prevented from reoccurring? Causal questions are not simply limited towards questions that explicitly use the word "cause". Instead, a causal question is *any* question that implies a causal relationship. If a researcher is curious how a variable *affects* another, this implies causality. Researchers from this discipline will generate theories that are clearly causal in nature, seeking answers to the above questions. However, subsequent hypotheses typically use non-causal language such as "associated" and findings are interpreted under the assumption that any statistical estimate is correlative and not representative of a causal effect. Such a disconnect is odd and not particularly helpful for transparency or for answering questions of substantive importance. If a researcher is willing to elaborate a novel causal theory, any subsequent analysis should be conducted with the aim of *directly testing this theory*. Absent a direct test of the researcher's hypothesis, subsequent quantitative methods are, at best, falling short and, at worse, misleading. Researchers currently have the tools to rigorously develop their research designs to directly test the causal theories they generate. Instead, descriptive findings are naturally paired with results that *should not be interpreted causally*. Without an explicit causal research design, descriptive results are simply descriptive. No amount of additional adjustment for seemingly-important covariates makes a finding "more causal" unless this adjustment is paired with tools and methodology from the causal inference literature. Even if the researcher is not being

intentionally misleading, policymakers, the public, and perhaps other academics may not notice these nuances and interpret these findings as causal.

In fairness, researchers are aware of this disconnect, and will justify this practice on the grounds that causal inference is a near-impossible task when working with observational data. As a result, descriptive or "correlational" analysis represent the next best approach. However, this paradigm if flawed for numerous reasons. First, if a researcher is simply interested in a correlational analysis (hereafter referred to as "descriptive" since correlation describes how variables move together and not whether this movement is driven by a causal relationship), a simple bivariable correlation between two variables of interest (X and Y) is all that is required (Hernán 2018). However, adjustment for covariates implies that a researcher is attempting to remove variation between X and Y that is explained by a set of covariates (Z_k) *in an attempt to isolate the effect of X on Y*. Even with the adjustment of covariates, a researcher may fall short of being able to make causal claims in their analysis. For example, a researcher may be unable to control for all relevant (confounding) covariates or issues of reverse causality may be present. Still, such analyses are fundamentally causal in nature, regardless of the degree to which results can be interpreted causally or as "merely descriptive" (Gerring 2012).

Transparency regarding the goal of a research design addresses the second issue with the dominant approach where, despite the word "cause" or "causal" being avoided, causal language such as "effect", "increases/decreases", "impact", and others are used liberally. If a researcher claims their results are descriptive, such language is inappropriate, and terms such as "is associated", "correlates" and "moves together" represent the correct language. However, it is not uncommon to witness researchers strictly avoiding causal language in their hypotheses while interpreting their results causally. For example, a researcher may offer a hypothesis that, as the

number of U.N. military personnel decreases, the number of battle-related deaths will decrease. Such a hypothesis is descriptive because it simply posits negative correlation between U.N. military personnel and battle deaths. However, the same researcher may then interpret a subsequent coefficient as an effect of U.N. military personnel on battle-related deaths despite the fact that "effect" implies a causal relationship which was not stated in the hypothesis nor supported by the data analysis. Such practice creates a natural line of defense for the researcher. Any claim that their findings cannot be interpreted causally can be met by the retort that the hypothesis in question was never causal. Such vague language ("stealth causal inference") creates scenarios where findings are difficult to falsify due to the ambiguity of interpretation (Grosz et al. 2020). If a methodology is developed to isolate a causal effect (generally speaking, adjusting for covariates within a regression represents such a methodology), it should be critiqued by this same standard.

Third, the dominant approach operates under an assumption that causal inference with observational data is *de facto* impossible, so any intentional causal interpretation of regression output ought to be avoided. For decades, randomized access to treatment under a controlled experiment was viewed as the near sole mechanism for isolating causal effects, generating a norm that, in the absence of randomized treatments, causal inference was impossible. However, randomization reflects merely *one type of an identification strategy* to estimate causal effects (Keele 2015). An identification strategy refers to a set of assumptions and goals that, if attained, warrant a causal interpretation of a statistical estimate. Randomized treatments, for example, are a type of an identification strategy because, given randomized access to treatment, all forms of confounding are removed and any difference in means in the outcome can be explained by treatment alone. Many identification strategies tools *have* been employed in peace studies, such

as instrumental variables, difference-in-differences, and selection on observables (regression adjustment, matching, etc.). However, these methods are often employed in isolation to the broader picture of causal inference. For example, while it is true that matching can address selection bias concerns (which itself is a task that any researcher seeking to make causal inferences will need to address), it also serves a valuable purpose as an alternative to conventional regression adjustment and, when paired with the correct tools and diagnostics, can be used to estimate causal effects. However, its usage in peace studies is rarely contextualized as such, partially because the necessary tools to estimate causal effects from matching (or regression adjustment), such as sensitivity analyses and transparent causal models (such as a directed acyclic graphs) are not paired with most matching/regression designs.

Such necessary features for an appropriate causal research design (hereafter referred to as an identification strategy in reference to the process of identifying causal effects from statistical estimates) includes 1) an explicit statement of the effect to be estimated, 2) a rigorous causal model, 3) a method designed to estimate the statistical estimate, and 4) some form of sensitivity analysis to test crucial identification assumptions.

In reference to the first criterion, not all causal effects are the same. For example, randomized experiments have the capacity to estimate the average treatment effect (ATE) which represents the difference between two identical populations where one is entirely exposed to treatment while the other is entirely unexposed. However, in most observational studies, estimating the ATE is infeasible, so effect estimates such as the average treatment effect on the treated (ATT - the effect of treatment for units who received treatment) and/or untreated (ATU – the effect of withholding treatment for units who did not receive treatment) may be more appropriate. Ultimately, the decision of which treatment effect to estimate will be determined by

data constraints and the identification strategy.[1] Still, clearly identifying the effect of interest is important given that an estimated effect will not be applicable to all units within a data set. For example, a researcher using a data set with 170 countries will likely not discover a treatment effect that is applicable to all countries within the data set because each treated/non-treated unit is unlikely to have a clear counterfactual (the outcome for a given unit if its treatment status were reversed) to be compared to. For this very reason, matching designs often offer greater transparency regarding which units a causal effect applies to given that only units where a counterfactual match can be found are kept within the data set. In contrast, regression adjustment masks the process of weighting each unit "under the hood", presenting coefficients that are mistakenly interpreted for all units within a data set (Aronow and Samii 2016).

Second, a researcher must construct a rigorous theoretical causal model that explains their assumptions regarding the causal relationships between X, Y, and Z_k. Causal inference is not possible without prior understanding of causal relationships. Statistics alone can tell us that the sun rising and the rooster crowing covary, but logic, theory, and an informed background in the topic motivate an understanding of the causal direction between the two. Further, only an informed background in a given area of research initially informs researchers of the variables that are important to adjust for. Any number of variables could be included in a model regressing civil war on GDP per capita. A researcher *could* "control" for democracy, ethnic diversity, colonial history, natural resource wealth, etc. but *should* they? Statistical software will not stop a researcher from creating a "garbage can model" where everything under the kitchen sink is adjusted for. Rather, prior theoretical knowledge of the causal dynamics complicating the

[1] For example, under identification strategies such as instrumental variables or regression discontinuity designs, the local average treatment effect (LATE) is estimated. For matching designs where treated units are dropped for a lack of finding a match/matches, the sample average treatment effect on the treated (SATT) is often estimated.

relationship between X and Y guides helpful and insightful adjustment strategies. Loosely-informed control variables are inevitably highly likely to damage causal inferences due to the threat of post-treatment bias (Dworschak 2023) and the myriad of possibilities in which "bad controls" can arise (Cinelli et al. 2022). Researchers can directly assess whether the inclusion of a control variable may be needed for a causal interpretation or may damage causal interpretation through the use of tools such as directed acyclic graphs (DAGs) where causal assumptions between X, Y, and Z_k are non-parametrically specified through the simplicity of nodes and arrows. Given that most causal research in quantitative peace studies relies and likely will continue to rely on selection on observables strategies such as regression adjustment and matching, the importance of identifying good and bad control variables cannot be understated.

Third, a method must be implemented to obtain a statistical estimate that can be given a causal interpretation. Familiarly, regression adjustment *can* be used for this, although, the method has legitimate drawbacks (discussed later in this chapter). Depending on the availability of data and the structure of a causal model, researchers can explore alternatives such as matching, inverse probability weighting, regression discontinuity designs, instrumental variables, difference-in-differences, synthetic controls, and more. Lastly, some form of sensitivity analysis is generally a good idea to test the strength of key identification assumptions that, when satisfied, help justify causal interpretation of a statistical estimate. For example, under a selection on observables identification strategy, the researcher adjusts for theoretically *all* confounding variables (variables that cause some change in both treatment and outcome where, if left unadjusted, a spurious association can be transmitted between treatment and outcome) so that treatment assignment is considered random conditional on the adjusted observable confounders. However, this assumption of no confounding is a difficult assumption to satisfy and is impossible

to directly test. Despite this, causal inference under the selection on observables identification strategy still rests on the satisfaction of this assumption. Sensitivity analyses, while not perfect, can be of great assistance in such scenarios that peace scholars often find themselves in. Under designs created by Cinelli and Hazlett (2020) and McGowan (2022), users can specify the strength of a hypothetical unobserved confounder and examine how much their observed estimate trends towards 0 or flips in the opposite direction when confronted with varying specifications of unobserved confounding. Such tools, while not directly testing the assumption of no unobserved confounding, are near-mandatory for selection on observables strategies since they allow the researcher to directly pre-empt and evaluate the inevitable critique that they may have failed to control for some key variable that is missing from the analysis. The skepticism from such critiques is often warranted and sensitivity analyses allow researchers to quantitatively address such skepticism.

In many ways, a call towards overt causal research using observational data in peace studies is not radical. Causal inference has been the primary goal of peace research even if the dominant approach has not overtly acknowledged this. Theories, methods, and interpretations of results in various disciplines clearly illustrate the causal goals of most contributions, even if the word "cause" itself is self-screened and appropriate methodology for evaluating causality is not fully realized. It is important to note that one can still formulate a causal identification strategy *even if* the final results cannot be interpreted causally. Indeed, this is not a rare occurrence as making causal inferences, especially in disciplines such as peace studies plagued by measurement error, missing data, endogeneity, etc., is very difficult. Under such circumstances, a researcher can make it clear that their goal is to estimate a causal effect, but their design may fail to do so, suggesting that their subsequent effect can be interpreted descriptively, without any

obvious causal interpretation (Gerring 2012). Even if a researcher is unable to identify a causal effect due to too many violations of assumptions within an identification strategy, at the very least, the result can simply be interpreted as the standard is now - a descriptive estimate.

Another consequence of ignoring the methodological contributions of the causal inference literature is the generation of "pseudo-facts" (Samii 2016). Absent a clearly defined causal model where researchers understand the causal ordering between X, Y, and Z_k, misspecification in a regression equation, and the implementation of *harmful* control variables are an inevitability. In a recent article, Dworschak (2023) quantified the gravity of this issue in conflict management and peace studies by finding that 75% of articles published in the Journal of Peace Research (JPR) and the Journal of Conflict Resolution (JCR) between January 2018 and May 2021 suffered from the implementation of "bad" control variables that potentially bias their results. Concerningly, only a small number of articles acknowledged the problematic nature of their control variables in the first place. Such patterns are also observed in other social scientific disciplines (Wysocki et al. 2022). Overall, in the interest of evaluating the theories that are of substantive interest to most researchers and the generation of meaningful causal estimates that can appropriately inform the public and policymakers, crucial aspects of a causal research design such as causal models, identification strategies, and sensitivity analyses should be regular features of quantitative peace studies.

Mutual Adjustment

Often, a single research paper is devoted towards the evaluation of more than one hypotheses. In these contexts, the hypotheses often concern the effects of different independent variables on a constant dependent variable. In situations such as these, two flaws become readily apparent. First, researchers may jointly evaluate all of their hypotheses within a single regression

model (mutual adjustment) and mistakenly interpret the coefficient for each independent variable as the total effect of that independent variable on the outcome. Second, even if a researcher specifies separate models to evaluate each hypothesis, a "standard set of controls" for each model is likewise inappropriate.

To consider the problem of mutual adjustment, recall that confounding (when a variable causes some change in both treatment and outcome: X ← Z → Y) represents a fundamental stumbling block for making causal inferences. Left unadjusted, any estimated effect of X → Y may be explained by Z. Consider a hypothetical "Hypothesis 1" where it is stated that X_1 has a causal effect on Y. Next, consider a hypothetical "Hypothesis 2" where it is stated that X_2 has a causal effect on Y. When adjusting for both in the same model, problems quickly arise. For example, the variables confounding the relationship between X_1 and Y are likely not identical to the variables confounding the relationship between X_2 and Y. While there may be some overlap, the set of relevant confounding variables to help identify the causal effects, $X_1 \rightarrow$ Y and $X_2 \rightarrow$ Y cannot be identical because the treatments themselves are not identical. One may object that one can simply adjust for *all* confounding relationships for each hypothesis within the same model. Such an objection would ignore the concerns of previously-referenced, "bad" post-treatment controls. A confounder for one causal relationship may be a type of post-treatment control for another. Adjusting for both within the same model simultaneously aids the improvement of causal inference for one hypothesis while damaging the other. Further, a confounder for one hypothesis may be entirely irrelevant for the evaluation of another hypothesis. While situations may exist where one can provide causal interpretation to several hypotheses within the same model, this situation is likely exceedingly rare given the complexity of the phenomena researchers try to model (Keele et al. 2019).

Second, even if a researcher evaluates each hypothesis with a separate model, a "standard set of controls" may be used for each model specification. While it is likely that each control variable will have a theoretical causal effect on the outcome, it may be inappropriate to include each control in each model given that a confounder needs to effect *both* the treatment *and* outcome. Common controls variables such as population, GDP per capita, democracy, etc. may have an impact on the outcome, but are not causally related to the treatment or vice versa. In such cases, their inclusion is not necessarily appropriate.[2] As a result, each model should feature control variables that are intentionally relevant for addressing confounding between the specific causal hypothesis being evaluated within a certain model. Additionally, while a researcher ought to explain the motivation behind including such control variables, including the coefficients in a regression table is not necessary given that valid interpretations cannot be given to control variables (Westreich and Greenland 2013, Keele et al. 2019, Hünermund and Louw 2022, Dworschak 2023). Coined by Westreich and Greenland (2013) as the "Table 2 Fallacy", control variables cannot be given a causal interpretation because the control variables themselves suffer from omitted variable bias. While inclusion of the controls is necessary for adjustment to analyze the effect of a given treatment on outcome, sufficient adjustment to identify the causal effect of a control variable on outcome is not implemented within a given model. To do so, one would have to control for the confounding between each control variable and the outcome simultaneously, which would quickly create an overly complex model with completely uninterpretable output. As a result, the sign and significance of a control variable is meaningless to inform the researcher

[2] While predictors of outcome are shown to increase statistical precision of causal estimates and predictors of treatment are shown to do the opposite (Cinelli et al. 2022), a researcher should still carefully consider whether the inclusion of a predictor of outcome is appropriate to include in their research design given issues concerning common support and the curse of dimensionality. For example, a researcher cognizant of the fact that each additional covariate creates counterfactuals that are more complicated to produce should always consider whether their data might support the inclusion of such an additional covariate.

whether their model is behaving "well", because it is known pre-estimation that the coefficients for the control variables cannot be interpreted as causal effects (Hünermund and Louw 2022, Dworschak 2023).

Solutions for this issue are quite easy in implementation. First, researchers should carefully model the causal relationships between their independent variables, dependent variables, and covariates. Given that each hypothesis provides a statements on a unique treatment or outcome of interest, the adjustment set should vary from model to model. A DAG is an easy-to-use and intuitive tool that allows the researcher to specify assumed causal relationships and identify variables that *should* and *should not* be adjusted for in the evaluation of separate hypotheses (Rohrer 2018, Cinelli et al. 2022). Lastly, much to the benefit of researchers fighting for every word they get in can given word count limits, reporting coefficients for control variables is not required in the presentation of results. In fact, given their lack of interpretive value, reporting coefficients on control variables should probably be actively avoided. Not only will this make results simply more readable, it may further encourage a norm of more intuitive regression reporting such as plotting marginal effects results.

Agnosticism Towards Panel Data Complications

Given that conflicts, their consequences, and conflict management programs vary across time and space, researchers studying such phenomena collect and analyze data that likewise varies spatially and temporally. Such panel data (time series cross-sectional/longitudinal) creates statistical inferential issues, as widely recognized with the implementation of clustered robust standard errors. Less recognized, although perhaps more important, are the complications that panel data creates for *causal* inference.

One clear consequence of data that varies across time is that an effect of interest may not be constant. That is, the fixed effect of a coefficient from a regression output (not to be confused with the fixed effects estimator) may misrepresent the dynamic nature of a given effect. Is the potentially pacifying estimated effect of a peace agreement constant across time? Is foreign aid effective only in the short-term? A coefficient for the treatment of interest does not reflect the entire range of effects across time that researchers may be substantively interested in. Another consequence of panel data are the complications that emerge in the construction of a counterfactual. Under a "single-shot" causal inference method, two similar units along a set of pre-treatment values are compared to each other that differ only in treatment status at a single point in time (Blackwell 2012). However, panel data observations are inherently more complicated than the single-shot scenario because units differ along both pre-treatment values *and pre-treatment histories*. Such pre-treatment histories force the researcher to consider whether the current treatment and/or the history of treatment impact a contemporary outcome (Blackwell and Glynn 2018).

Lastly, time-varying treatments and outcomes naturally feature time-varying covariates. Such time-varying covariates may quickly reveal themselves as problematic given the oftentimes ambiguous nature of their causal ordering. For example, consider a research project examining the effect of peacekeeping operations (PKOs) on one-sided violence (OSV). A host of confounding relationships may come to mind in this example, such as the effect of government military capacity or conflict duration on both treatment (PKOs) and outcome (OSV). Of course, in this scenario, lagged versions of the treatment and outcome *also* represent confounding effects ($PKO_t \leftarrow PKO_{t-1} \rightarrow OSV_t$ *and* $PKO_t \leftarrow OSV_{t-1} \rightarrow OSV_t$). A PKO in 1996 is causally related to a PKO in 1997. Further, a PKO in 1996 also likely has an effect on the levels of one-sided violence

in 1997 since past treatments often impact current outcomes (the inverse is also true). One can adjust for lagged versions of treatment or outcome (although, strategies such as these are not above critique), but one would find that the problem if history still persists since PKO_{t-k} and OSV_{t-k} are likewise confounding effects. As the following paragraph demonstrates, popular approaches for dealing with such issues are generally insufficient to identify a causal effect.

While regression adjustment and matching remain the most popular methods in quantitative peace research, such methods on their own do not account for the unique dynamics within panel data. Under such methods, counterfactuals are generated without the acknowledgement of pre-treatment histories. Further, researchers face a "damned-if-you-do and damned-if-you-don't" situation. On the one hand, one must adjust for a time-varying confounder to remove the bias introduced by this confounder. One cannot isolate a causal effect of military victories on peace without adjusting for a time-varying confounder such as the degree of support from an external actor. On the other hand, this time-varying confounder may also be changed by the treatment *and/or* outcome itself, resulting in post-treatment bias when it is controlled for (Blackwell 2012). After all, the degree of external support for a warring party will inevitably be impacted by whether conflict has remained ongoing. As a result, in the panel data context, a concept such as levels of external support is simultaneously a confounder (a good control) and a post-treatment variable (a bad control). A popular approach to solving this issue is to include lagged treatment values to account for the dynamic nature of panel data. However, such approaches are proven to also introduce post-treatment bias when lagged treatment values are causally related to time-varying covariates (Blackwell and Glynn 2018). When working with panel data, the fixed effects (FE) estimator is also a popular choice as it accounts for all time-invariant confounding. However, recent research strongly suggests that the FE estimator is

generally inappropriate for most panel data settings (Bell and Jones 2015, Blackwell and Glynn 2018, Imai and Kim 2019, Plümper and Troeger 2019). In particular, the FE estimator is incredibly sensitive to a number of dynamic assumptions including non-allowance for scenarios where past treatments directly affect current outcomes *and* where past outcomes directly affect current treatments (Imai and Kim 2019). Such a stringent assumption is unlikely to be met in peace research where past outcomes directly inform the onset, termination, and level of variation within conflict management programs. For example, levels of foreign aid to a post-conflict country will respond to past outcomes since donors will evaluate their aid packages dependent on the past trend in outcome.

Due to the dominance of panel data in peace research, a number of existing methods should be added as a regularly employed tool to account for the dynamic nature of panel data. Generalized methods (g-methods) such as marginal structural models and structural nested mean models along with panel data matching are existing methods that extend the familiar regression adjustment and matching framework to specifically account for the complications introduced by panel data (Blackwell and Glynn 2018, Imai et al. 2021). Under a more rigorous set of assumptions, researchers can adopt a difference-in-differences design. In cases where treatment is rarely distributed and the parallel trends assumption does not hold, the synthetic control method presents an appealing (and shockingly underutilized in peace research) method for estimating a causal effect (Abadie et al. 2015, Abadie 2021). Regardless of the methods listed above, modern research demonstrates that ignoring the dynamic nature of panel data biases estimates. Further, once-standard approaches such as fixed effects or including lagged values in the right-hand side of a regression equation have likewise proven generally incapable of resolving the complications introduced by panel data. As such, quantitative peace scholars should

consider making the shift towards methods that were designed specifically for the data they most often employ.

The Crutch of Regression Adjustment

By far, the predominant methodological tool in quantitative peace studies is regression. In particular, regression adjustment - the practice of specifying covariates in the right-hand side of the equation along with treatment - is the *de facto* default tool to estimate causal effects (whether a researcher knowingly or unknowingly is attempting to do so).[3] Despite the popularity of regression adjustment, this method features serious limitations for making causal inferences. First, an alleged comparative advantage of regression adjustment over alternatives such as matching or randomized experiments is the method's ability to examine trends on a larger, generalizable population. For example, where a randomized experiment may be "limited" to examining effects of foreign aid in a handful of local communities, advocates of regression adjustment may claim that their models' capabilities of examining the effects of foreign aid in *all* developing countries represents a comparative strength of regression adjustment. However, as demonstrated by Aronow and Samii (2016), this "strength" is often a masquerade, hiding behind unreported regression weights. Regression weights may be allocated highly unequally between units, leading to estimates that are *heavily* influenced by an "effective sample" (Aronow and Samii 2016). This critique does not imply that alternative strategies do not face issues with external validity. Rather, regression adjustment does not necessarily boast increased external validity as a comparative strength. While regression adjustment may nominally utilize all if the data in a provided data set (assuming no missing values), the method *effectively* is often weighted

[3] This article makes a distinction between regression and regression adjustment because regression itself is utilized as a core feature for estimation under a variety of causal inference estimation techniques such as difference-in-differences, matching, inverse probability weighting, instrumental variables, etc.

heavily towards a subset of units that the method does not directly report. Further, alternative strategies make their limitations in external validity clear whereas it is uncommon to report the distribution of weights in a design using regression adjustment.

Second, the strong functional form assumptions of regression adjustment place serious challenges on the researcher to correctly specify the relationship between treatment and covariates on outcome. While technical solutions are readily avaliable (log-transformations, polynomials, interactions, etc.), such adjustments do not naturally inform the researcher which variables should be transformed to satisfy functional form assumptions. Any incorrect specification within the regression equation is likely to bias a potentially causal estimate. While this second point is unlikely novel to the reader, it is a worthwhile critique of regression adjustment, especially given readily avaliable designs that rely on much weaker functional form assumptions, such as matching. While it may be convention to transform certain variables (log-transformations for right-skewed variables, polynomial transformations for curvilinear relationships), it should not be understated that researchers are making strong assumptions that are consequential to their regression output when either including or excluding non-linear terms.

Third, the use of regression adjustment for causal inference becomes particularly strained when considering its reliance on extrapolation to generate counterfactuals beyond the support of the data. To contextualize the complications of this feature, consider plugging any unit of interest (a country for example) at a given time in a regression equation. Holding all covariates at their mean/mode, a researcher may examine the difference in outcome when treatment is set to one versus when treatment is set to zero. Such an approach will yield an estimate. However, much like the issue of unreported regression weights, the burden is on the researcher to determine whether such an estimate is supported by the *data used for the model* and not the *model itself*. As

King and Zeng (2006) demonstrated, where covariates fail to overlap between treated and non-treated groups (meaning treated and non-treated groups fundamentally differ in some observable way), a valid counterfactual cannot be constructed from the data. Regardless, regression allows for the construction of a counterfactual, although, this counterfactual is not generated from the data itself, rather, from highly model-dependent predictions. Such predictions place strong assumptions on the validity of the model that are often not sufficiently examined in regression adjustment studies. This complication obviously becomes problematic on the fringes of counterfactual questions. For example, a researcher is unlikely to get a valid counterfactual estimate if one wishes to understand what the economic effects of a civil war in Denmark in 2018 would be because there are no countries similar to Denmark in 2018 that *are* experiencing civil war. Nonetheless, regression adjustment would supply a counterfactual for Denmark that extrapolates beyond the support of the data. Concerningly, King and Zeng (2006) suggest that extrapolation may not be an issue merely at the fringes, implying concerns for the validity of average estimates generated from regression adjustment in studies where the distribution between treated and non-treated units differs in fundamental ways. Again, there are ready solutions for this problem beyond regression adjustment, such as matching which explicitly discards observations where no counterfactual can be observed.

In this section, matching has been suggested a number of times to address some of the limitations of regression adjustment. However, in this final critique, matching no longer serves as a valid alternative. As mentioned earlier, regression adjustment relies on the selection on observables (SoO) identification strategy to estimate causal effects. The issue with SoO is that is it the *weakest* of identification strategies due to the incredibly difficult ignorability assumption to satisfy. After all, it will be a hard sell for many to justify random treatment assignment

conditional on a set of observed covariates. Even accompanied with supportive results from a sensitivity analysis (which should be mandatory in any design built off of SoO), the unverifiable threat of an unobserved confounder still remains present. As such, it is worthwhile to consider a reorientation in the basic framework that most quantitative peace studies research designs operate within. Rather than leap to regression adjustment as the default (or a matching or inverse probability strategy which also rely on the SoO identification strategy), researchers should consider and exhaust the possibility of stronger identification strategies. Is experimentation feasible, practical, and ethical? Does a discontinuity in treatment assignment exist that may support a regression discontinuity design? Are current instruments valid for an instrumental variables analysis? Could a stronger instrument be implemented? Such investigations to justify stronger identification strategies take more time than simply jumping to regression adjustment, and such questions may naturally lead researchers back to SoO anyways. Nonetheless, in the interest of accumulating *reliable knowledge* in a field where *real human suffering* is often the outcome of interest, such due diligence is warranted.

Frequentist Inference as Default Statistical Inference

Overwhelmingly, the data used in quantitative peace studies consists of a population of some sort (*all* countries, *all* warring dyads, etc.) over a certain set of years. Such data may be referred to as an "apparent population" (Berk et al. 1995). While it may be inappropriate to refer to such data as representative of the true population due to small gaps in spatial coverage and measurement error, the data is certainly not representative of a random sample from the population as conventional statistical methods assume. Nonetheless, statistical inference in quantitative peace research is solidly conducted under the frequentist paradigm. This represents something of a puzzle in the literature given how seemingly uninterpretable frequentist p-values

and confidence intervals *are* with most conflict data. Nonetheless, low p-values and 95% confidence intervals that fail to overlap with zero are considered strong evidence that an effect of interest exists. In particular, the reporting of confidence intervals has been considered an alternative to the oft-criticized practice of placing asterisks next to estimates. While confidence intervals may seemingly report more information, it is worth considering what they tell us. If a 95% confidence interval fails to overlap with 0, we can claim that, with repeated sampling, the parameter of interest would be covered within the range of the confidence interval 95% of the time and that. Given this, it is unlikely that an effect would be 0.

It should be glaringly obvious why confidence intervals are uninterpretable for apparent population data. There is no repeated sampling in apparent population data. Unlike data collected as a sample from the population, downloading data from the Correlates of War or the Uppsala Conflict Data Program will result in the same data every time (until updated versions of the data are released). In contrast, drawing data as a sample from the population will result in an entirely unique data set every time. It is under these circumstances that conventional frequentist inferential tools such as p-values and confidence intervals were developed. Many scholars have noted that the social sciences feature various sub-fields where frequentist methods are generally not applicable for this very reason (Western and Jackman 1994, Berk et al. 1995). It is simply conceptually inapplicable to apply frequentist methods under such contexts when the units of interest can all (or almost all) be entirely observed. Again, this stands in stark contrast to situations where frequentist inference is applicable, such as American political behavior polling where we cannot observe *all* of the voting-age American population and uncertainty is naturally a feature of each unique sample. Of course, this does not imply that there is *no* uncertainty in point estimates derived from apparent population data and that apparent population data *is itself*

population data and that the endeavor of statistical inference itself is therefore meaningless. Uncertainty abounds in measurement error, missing data, operationalization of key concepts, etc. Rather, researchers ought to acknowledge that if long-run repeated sampling is not the context in which their data is generated, conventional frequentist practices are not applicable to their research and will generate statistics that are uninterpretable.

Of course, frequentist methods have also been heavily critiqued in the fields that they *are* conceptually applicable to. One large critique levies criticism concerning the interpretable value of frequentist null hypothesis significance testing (NHST), arguing that, in most academic contexts, the NHST generates information that is contrary to what most scholars wish to estimate. Researchers conclude their theoretical sections with a hypothesis (or hypotheses) of interest and delve into the research design, presumably assuming that such research designs are able to *test* their hypothesis of interest. Under the frequentist paradigm, this does not occur. NHST evaluates the probability of observing given data *under the assumption that the null hypothesis is correct* (P(D|H)). If the data on-hand seems impractical under such an assumption, most researchers will say that the null has been rejected, offering evidence for their hypothesis of interest.

Many scholars have noted that this approach is heavily unsatisfactory. For one, such an approach neither directly tests the hypothesis of interest *nor* the null hypothesis. A small p-value is not an indicator of either the probability of a certain hypothesis being true nor the probability of the null hypothesis being false (Gill 1995, Imbens 2021). The p-value is not an indicator of the probability of *any* hypothesis being true because frequentist NHST examines the probability of *data* under a fixed null hypothesis rather than the probability of a hypothesis under fixed data. However, echoing similar concerns with the practice of causally interpreting non-causal

regression estimates, applied interpretation of p-values and confidence intervals tends towards interpretation that are not accurate. As Gill (1995) argues, any framework that tests the inverse, the probability of a certain hypothesis being true given data (P(H|D)), has more desirable properties. Under such an approach, the data may be assumed as fixed and informs which hypothesis is most likely. An application such as this is certainly more interpretable for important consumers of conflict management research (such as policymakers), and it is more aligned with the intent of conflict management research itself where a hypothesis (or several) is proposed and the researcher seeks to evaluate the strength of said hypothesis.

Bayesian inference serves as a alternative for statistical inferential purposes that, given the limitations of frequentist NHST mentioned above, is highly applicable for quantitative peace research. Under Bayesian inference, probability is not defined as an estimate of the long-run frequency of a repeated event, so interpretability problems rooted under such unrealistic assumptions are not an issue under the Bayesian framework. Rather, Bayesian inference views probability as a subjective degree of belief in the plausibility of some statistic. Such a conception of probability likewise leads to hypothesis testing that directly tests hypotheses of interest rather than testing the probability of data under an assumed fixed null hypothesis. Not only does Bayesian hypothesis testing offer a more intuitive approach to hypothesis testing, statistics for interpreting the probability of hypotheses are more interpretable. As an example, consider the interpretability of frequentist 95% confidence intervals. Plotting confidence intervals will demonstrate a point estimate and an interval covering values above and below the point estimate (assuming the estimation of a two-sided confidence interval). Not only are such confidence intervals uninterpretable in most peace research contexts, confidence intervals fail to report the information many researchers implicitly assume they do. For example, the point estimate in a

confidence interval is no more likely to represent the statistic of interest than any of the other values that are covered within the 95% confidence interval range. The 95% confidence interval is not assigning confidence to the point estimate of interest. That point estimate is simply a single value within the interval that is equally likely to be correct as any other value that might fall along the interval. Again, such an interpretation is fairly counter-intuitive and uninteresting. However, Bayesian 95% credible intervals offer an easily-interpretable and applicable method for evaluating a hypothesis. Under a Bayesian 95% credible interval, a researcher can state that the estimate of interest lies within a given range with 95% probability *and*, in contrast to the frequentist confidence interval, a researcher can report the probability of any point within the credible interval being the estimate itself. That is, a researcher may report that estimates near the ends of both sides of the 95% credible interval are unlikely while estimates nearing the interior of the credible interval are more likely with a precise percentage estimate.

Of course, Bayesian inference as an alternative to the problems introduced by frequentist NHST is not without its critiques. On a practical front, Bayesian inference is (despite dramatic growth) a less populous enterprise. As a result, entry-level teaching resources, online support, and coverage of methods are less popularly supported. In addition, Bayesian methodology will often take more time to execute, both in estimation (computational time, although this has been dramatically improved) and pre-estimation (time spent thinking about priors). On its popular methodological critiques, perhaps none is more well-known than the reliance on subjective priors. The Bayesian posterior estimate is a "compromise" between prior knowledge (quantified and specified by the user) and current data. Many have expressed discomfort with the user-specification of Bayesian priors, assuming that such subjectivity easily allows for subjective biases to taint "objective" scientific estimates and intentional "hacking" of results. First, any

paper employing Bayesian methods should report their priors, rendering priors publicly available for academic review and scrutiny, as any other portion of a research design incorporating subjective decisions is. Relatedly, frequentist methods allow for a number of equally subjective methodological decisions, such as operationalization decisions, the inclusion/exclusion of control variables, a threshold by which statistical significance is "achieved", etc. Third, as more data accumulates, the influence of the prior on the posterior estimate naturally diminishes, alleviating concerns of a prior biasing a study. Fourth, rather than viewing the specification of priors as an inconvenient feature of Bayesian inference, one could also view prior specification as a powerful tool. Assume that a researcher does their due diligence and sufficiently surveys existing research so that their specification of a prior is truly reflective of what extant literature suggests. Bayesian methodology would be able to formally incorporate prior theory, empirical findings, domain knowledge, and relevant history that, under a frequentist method, would simply be ignored if it was not somehow present in the current data set. Indeed, it is not uncommon for quantitative peace scholars to find themselves with small data sets but a wealth of prior information. For situations such as these, Bayesian methods thrive. While specifying a prior may be considered as subjective, it is also subjective to arbitrarily leave out important information from an analysis because it is not present in a given data set.

Lastly, subjective choices in frequentist NHST can easily "hack" results (intentionally or unintentionally). Lenz and Sahn (2021) demonstrate this point well by examining how the introduction of control variables may artificially increase statistical significance. The authors find that many bivariable correlations between key independent variables of interest and dependent variables are both never reported nor reach statistical significance. Instead, the authors found that over 30% of observational studies published in the American Journal of Political

Science between 2013 and 2015 achieve statistical significance solely through the introduction of control variables. Among these studies, none reported this information. Importantly, as the authors note, achieving statistical significance following the incorporation of control variable is not inherently indicative of a problem. Certain control variables are necessary to include in the analysis (although, determining necessity of controls highly depends on formally presenting assumed causal relationships, justifying the use of DAGs all the more). However, given the highly subjective process of selecting control variables in quantitative peace studies, any critique levied against Bayesian methods for subjective priors likewise applies to frequentist methods where core inferential findings can depend entirely on subjective decisions made by the researcher(s). Overall, despite the level of importance attached to "statistically significant" findings, the extant dominant approach of NHST for quantitative peace research is both generally inapplicable and comparatively uninteresting. While a shift towards Bayesian inference is simply one alternative, it is an alternative that directly addresses the various problems introduced by the application of NHST towards quantitative peace research.

Suggestions and Moving Forward

As it currently stands, a consumer of the typical quantitative peace research scholarship is likely to read an article that insufficiently examines the validity of a causal theory, interpreting the results as causal (even accidentally) when no such interpretation is warranted. If such an article interprets more than one hypothesis, it is likely that this article mutually adjusts for both hypotheses within the same model, or employs a "standard set of controls" for all models if separate hypotheses are examined with separate models, further damaging any causal interpretation. Complexities of prior treatment and outcome values are likely either ignored or outdated methods to deal with such a problem are used, again, further damaging causal

interpretation. Regression adjustment likely produces the key estimate of interest, although some type of matching may be utilized as a robustness check to regression adjustment output. Regardless, neither approach accounts for the difficulties imposed by using panel data and both approaches represent some of the weakest identification strategies. Lastly, the degree to which a result may be deemed important or even publishable is likely conditioned on an arbitrary threshold of significance under an inferential paradigm that is not even applicable to the data used in the typical quantitative peace research article. In short, despite the honest technical efforts made my scholars, one can place great confidence in the statement that the fields that make up quantitative peace research produce largely pseudo-factual research (Samii 2016). Causal interpretation is given to results that have no business being interpreted causally, and the methodology used to examine whether these pseudo-factual estimates are statistically different from zero is fundamentally not applicable nor policy-relevant.

Ready solutions in booming literatures exist to easily address these shortcomings. Researchers can use DAGs to identify possible identification strategies to pursue and, as will most often be the case, should a researcher find themselves employing a selection on observables identification strategy, a DAG will guide the researcher in selecting a set of controls that do not harm a potential causal interpretation of an estimate. Under selection on observables, a variety of sensitivity analyses are avaliable to assess the threat of unobserved confounding. When researchers are interested in competing/two or more hypotheses, DAGs are also helpful for informing the unique set of control variables that each evaluation of separate hypotheses will require. While the literature and methodology on causal inference with panel data certainly has more room to grow, rich literatures on g-methods, difference-in-differences, and synthetic controls can guide researchers moving away from the regression adjustment paradigm. In this

transition away from the comfortable use of regression adjustment, researchers can consider what stronger identification opportunities may exist in their respective fields. While novel natural experimental designs are certainly harder to execute than specifying a regression model, further attention should be paid towards the existence of discontinuities in peace research, and more scrutiny and commentary should be devoted towards the use of popular instruments commonly employed in instrumental variables designs. Finally, most (although, not all) scholars in quantitative peace studies should consider making the transition to Bayesian inference. Not only is the underlying logic of Bayesian inference much more applicable than frequentist NHST, the interpretability of inferential findings under Bayesian inference is wholly more interesting and digestible for both technical and non-technical audiences.

A helpful observation in *any* quantitative discipline is that no project will be perfect. Uncertainty abounds in empirical research and better data and more nuanced methods are always around the corner. Simply put, a valuable project that contributes to cumulative knowledge in such fields can be *good enough* to qualify as meaningful research. The conclusions drawn from this article should not be that research must be perfect for it to be valuable. Instead, this article argues that researcher goals should be married with the appropriate methodology and interpretation. Unfortunately, many of the dominant conventions in quantitative peace management research do not satisfy this latter criterion. The good news is that existing methodology is generally sufficient to dramatically increase the quality of scientific output within quantitative peace studies.

Finally, one may erroneously conclude that, due to the standards of newer methodological research, the contribution of past research using dated or incorrect methods is not informative. While it is true that the results from the empirics may be up for serious question, one cannot

deny the contribution of theory that such works provide. To varying degrees, both causal inference and Bayesian inference are greatly assisted by the existence of prior theoretical work that informs the identification of valid control variables and the specification of priors. Absent such theory, causal estimation and Bayesian inference would be starting from ground-zero.

References

Abadie, Alberto. 2021. "Using Synthetic Controls: Feasibility, Data Requirements, and Methodological Aspects." *Journal of Economic Literature* 59(2): 391–425.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2015. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59(2): 495–510.

Achen, Christopher H. 2005. "Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong." *Conflict Management and Peace Science* 22: 327–39.

Aronow, Peter M., and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?" *American Journal of Political Science* 60(1): 250–67.

Bell, Andrew, and Kelvyn Jones. 2015. "Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data." *Political Science Research and Methods* 3(1): 133–53.

Berk, Richard A., Bruce Western, and Robert E. Weiss. 1995. "Statistical Inference for Apparent Populations." *Sociological Methodology* 25: 421–58.

Blackwell, Matthew. 2012. "A Framework for Dynamic Causal Inference in Political Science." *American Journal of Political Science* 57(2): 504–20.

Blackwell, Matthew, and Adam Glynn. 2018. "How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables." *V-Dem Working Paper Series* 67.

Cinelli, Carlos, Andrew Forney, and Judea Pearl. 2022. "A Crash Course in Good and Bad Controls." *Sociological Methods & Research* 0(0): 1–34.

Cinelli, Carlos, and Chad Hazlett. 2020. "Making Sense of Sensitivity: Extending Omitted Variable Bias." *Journal of the Royal Statistical Society* 82(1): 39–67.

Colquhoun, David. 2014. "An Investigation of the False Discovery Rate and the Misinterpretation of P-Values." *Royal Society Open Science* 1: 1–16.

Dworschak, Christoph. 2023. "Bias Mitigation in Empirical Peace and Conflict Studies: A Short Primer on Posttreatment Variables." *Journal of Peace Research* 0(0): 1–15.

Gerring, John. 2012. “Mere Description.” *British Journal of Political Science* 42(4): 721–46.

Gill, Jeff. 1999. “The Insignificance of Null Hypothesis Significance Testing.” *Political Research Quarterly* 52(3): 647–74.

Grosz, Michael P., Julia M. Rohrer, and Felix Thoemmes. 2020. “The Taboo Against Explicit Causal Inference in Nonexperimental Psychology.” *Perspectives on Psychological Science* 15(5): 1243–55.

Hernán, Miguel A. 2018. “The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data.” *American Journal of Public Health* 108(5): 616–19.

Hünermund, Paul, and Beyers Louw. 2022. “On the Nuisance of Control Variables in Regression Analysis.” http://arxiv.org/abs/2005.10314 (June 11, 2023).

Imai, Kosuke, and In Song Kim. 2019. “When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?” *American Journal of Political Science* 63(2): 467–90.

Imai, Kosuke, In Song Kim, and Erik H. Wang. 2021. “Matching Methods for Causal Inference with Time-Series Cross-Sectional Data.” *American Journal of Political Science* 0(0): 1–19.

Imbens, Guido W. 2021. “Statistical Significance, *p* -Values, and the Reporting of Uncertainty.” *Journal of Economic Perspectives* 35(3): 157–74.

Ioannidis, John P. A. 2005. “Why Most Published Research Findings Are False.” *PLOS Medicine* 19(8): 0696–0701.

Keele, Luke. 2015. “The Statistics of Causal Inference: A View from Political Methodology.” *Political Analysis* 23(3): 313–35.

Keele, Luke, Randolph T. Stevenson, and Felix Elwert. 2020. “The Causal Interpretation of Estimated Associations in Regression Models.” *Political Science Research and Methods* 8(1): 1–13.

King, Gary, and Langche Zeng. 2006. “The Dangers of Extreme Counterfactuals.” *Political Analysis* 14(2): 131–59.

Lenz, Gabriel S., and Alexander Sahn. 2021. “Achieving Statistical Significance with Control Variables and Without Transparency.” *Political Analysis* 29(3): 356–69.

Lundberg, Ian, Rebecca Johnson, and Brandon M Stewart. 2021. “What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory.” *American Sociological Review* 86(3): 532–65.

McGowan, Lucy D'Agostino. 2022. "Tipr: An R Package for Sensitivity Analyses for Unmeasured Confounders." *Journal of Open Source Software* 7(77): 1–6.

Plümper, Thomas, and Vera E. Troeger. 2019. "Not So Harmless After All: The Fixed-Effects Model." *Political Analysis* 27(1): 21–45.

Rohrer, Julia M. 2018. "Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data." *Advances in Methods and Practices in Psychological Science* 1(1): 27–42.

Samii, Cyrus. 2016. "Causal Empiricism in Quantitative Research." *The Journal of Politics* 78(3): 941–55.

Schrodt, Philip A. 2014. "Seven Deadly Sins of Contemporary Quantitative Political Analysis." *Journal of Peace Research* 51(2): 287–300.

Western, Bruce, and Simon Jackman. 1994. "Bayesian Inference for Comparative Research." *American Political Science Review* 88(2): 412–23.

Westreich, D., and S. Greenland. 2013. "The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients." *American Journal of Epidemiology* 177(4): 292–98.

Wysocki, Anna C, Katherine M Lawson, and Mijke Rhemtulla. 2022. "Statistical Control Requires Causal Justification." *Advances in Methods and Practices in Psychological Science* 5(2): 1–19.