Assessing Robustness of Post-Estimation Quantities of Interest*

Ian Delabie and Max Goplerud

February 18, 2025

Abstract

Applied research in political science often produces post-estimation quantities of interest, such as marginal effects or predicted probabilities, to report the impact of a key variable of interest. An important question is the robustness of these results to small changes in the underlying data, which is closely connected to some notions of external validity. Building on recent work in econometrics, we focus on examining the impact on a reported quantity of interest after removing a small set (e.g., 0.5%-1% of the estimation data) of highly influential observations. If the results change considerably, this suggests a non-robust result, and our paper outlines various strategies to improve the robustness of the original model.

Identifying highly influential sets of observations is challenging, and we provide a computationally efficient framework for doing so that can be applied to generalized linear models, random effects models, and/or models with clustered standard errors. We also derive analytical expressions for the approximate impact of deleting this set of observations on key quantities of interest such as predicted probability curves. We re-examine a number of existing studies and find that some reported results are non-robust to removing even quite small sets of observations; in many cases, robustness can be improved by using more flexible models.

*We thank Matt Blackwell, Melody Huang, Tara Slough, and participants at PolMeth 2024 and APSA 2024 for their helpful comments on earlier drafts. All remaining errors are our own. An `R` package to implement these methods will be available at `https://github.com/mgoplerud/ApproxInfluence`.

1 Introduction

A key step in the workflow of much applied research, after estimating some regression model, is to produce a more interpretable quantity of interest (King, Tomz and Wittenberg, 2000). For example, instead of reporting coefficients from the model directly, researchers often produce quantities such as average marginal effects, first differences, or predicted probability curves to clearly illustrate the impact of a key variable. A further step is examining the robustness of the reported results. Current practice often includes many different specification tests in an appendix to demonstrate the robustness of key results to changes in the underlying model. Our paper provides an easy-to-implement test that focuses on a different sort of robustness: How sensitive is the stated result to small changes in the underlying data?

Using our methodology, the researcher can quickly identify a small fraction of the data (e.g., 0.5% of the original sample) whose removal is likely to result in substantial changes from the result (e.g., a confidence interval that now contains zero; a predicted probability curve that is flat, etc.). If this is the case, then the stated result is likely not robust, and the researcher should engage in further diagnostics. We stress that this measure of influence is distinct from other commonly used metrics of "global" influence, e.g. Cook's distance, in that it measures impact on a specific quantity of interest versus impact on the entire model.

As we demonstrate in a stylized example and empirical applications, non-robustness to small removals of data is often—although not necessarily—due to model mis-specification or unmodelled heterogeneity, and thus our method indicating a non-robust result encourages the researcher to re-examine their existing model (see King and Roberts 2015 for a related philosophy). Separate from concerns about internal validity, we also note that this type of non-robustness is closely connected to some notions of external validity, as it shows the existence of nearby populations that report a considerably different effect than the one associated with the current sample (Findley, Kikuta and Denly, 2021; Egami and Hartman, 2023; Devaux and Egami, 2022).

The key methodological challenge, however, is identifying which sets of observations are likely to lead to substantial changes in the original results. While identifying the influence of a single observation is a well-studied problem (see, e.g., Chatterjee and Hadi 1986), it is much harder to identify a *set* of observations that might be jointly influential. Naively searching over all possible

sets of observations is infeasible; for example, assume that one had a dataset with 2,000 observations, 10 observations constitutes a set of 0.5%; there are over 10^{26} sets that would need to be checked to exhaustively find the one with the greatest impact.

Thus, research typically relies on approximate methods; our paper uses on an approach based on influence functions that approximately—but quickly—identifies a set that likely to be highly influential if removed jointly (Broderick, Giordano and Meager, 2023). This methodology has been used in some existing papers in political science (e.g., Martinez 2022; Eubank and Fresh 2022), although we suggest adapting the workflow in Broderick, Giordano and Meager (2023) to only focus on examining sets of a small size (e.g., no greater than 2%) where the approximate method is likely to perform well.

Beyond providing detailed guidance on how to interpret and address evidence of a non-robust result, we also substantially methodologically extend Broderick, Giordano and Meager (2023)'s approach to cover key quantities of interest and methods considered in applied social science. First, we derive exact analytical formulae to compute the robustness of coefficients and standard errors that come from generalized additive models—including random effect models and spline models as important special cases. Our accompanying software implements these formulae for Gaussian and non-Gaussian outcomes, extending beyond linear regression and instrumental variables available in existing software. Second, we derive formulae and software to compute the robustness of key quantities of interest such as average marginal effects, first differences, and expected value curves—in addition to the existing focus on a regression coefficient and its confidence interval.

Critically, however, assessing the robustness of those quantities of interest must not be done in a "naive" fashion as doing so imposes enormous computational costs. The problem arises because many quantities and their standard errors (e.g., Wald statistics, average marginal effects, etc.) depend on the full *covariance* matrix of the estimated parameters. A naive way of computing the robustness of those quantities involves computational operations whose cost—in number of observations N and coefficients p—scale as badly as Np^3 and require storing new matrices with an order of Np^2 elements.[1] Thus, in even modestly sized problems, this becomes enormously expensive and, often, beyond the memory capacity of many standard laptops.

[1] For comparison, simple linear regression has matrix computations whose cost scale as $Np^2 + p^3$ and whose storage requirements are around p^2.

A key methodological contribution of this paper, therefore, is showing that careful manipulation of the underlying matrix algebra and calculus means that for all relevant quantities of interest, the approximate sets can be computed without ever forming a matrix of Np^2 and with a computational cost that grows on the order of Np.

The remainder of the paper proceeds as follows; Section 2 provides a detailed discussion of our notion of "robustness" and how to proceed if lack of robustness is detected. Section 3 formally derives our approximate influence scores for coefficients and standard errors in generalized linear models and then generalized additive models. Section 4 considers robustness for a suite of post-estimation quantities of interest. Section 5 applies our methodology to two empirical applications, and Section 6 concludes.

2 Conceptualization and Implications

We begin with a stylized example to illustrate our methodology and its incorporation into existing research workflows. Adapting an example from King and Roberts (2015), consider a model whose true data generating process is quadratic with respect to a covariate x_i,

$$y_i \sim N(-x_i + 6x_i^2, \sigma_\epsilon^2); \quad x_i \sim N(0,1); \quad \sigma_\epsilon^2 = 1, \tag{1}$$

but for which a linear model, $E[y_i|x_i] = \hat{\beta}_0 + \hat{\beta}_1 x_i$, is incorrectly used. If one were interested in the impact of deleting a single observation, one could simply remove an observation x_i and compare the impact on the estimates. The difference between the original $\hat{\beta}_1$ and the estimate resulting after deleting observation i—or other closely related quantities—has been studied extensively (e.g., DFBETA, "sample influence curve", or "sensitivity curve"; Chatterjee and Hadi 1986). For our purposes, it is closely related to the "influence" that observation i has on $\hat{\beta}_1$; see Section 3.

In a least squares setting, there is an exact closed form for this quantity that is a function of (a) its leverage, i.e., observations with covariate values that are far away where most of the data is located, and (b) its residual, i.e., the difference between the observed outcome and the predicted value (Chatterjee and Hadi, 1986). The intuition is that influential points on $\hat{\beta}_1$ are ones that are not well fit by the model and/or are quite extreme in their covariate values (Chatterjee and Hadi, 1986; Broderick, Giordano and Meager, 2023). We note that this measure is focused on the

impact on a specific quantity of interest (e.g., $\hat{\beta}_1$), and thus differs fundamentally from quantities such as Cook's distance that provide a measure of "global" influence on the fit of *entire* model. Appendix F.1 provides an in-depth discussion of this point showing that Cook's distance is closely connected to the sum of the *squared* influences across all coefficients, e.g., $\hat{\beta}_0$ and $\hat{\beta}_1$ (Chatterjee and Hadi, 1986).

There is a large literature that studies the influence of a single observation; however, Broderick, Giordano and Meager (2023) tackle the more difficult challenge of finding a set of α observations that are, jointly, the most influential on $\hat{\beta}_1$. Computing this exactly is computationally infeasible on most problems as it would require searching through an enormous number of sets of observations.

Thus, the key idea of Broderick, Giordano and Meager (2023) is to inexpensively create an *approximate* most influential set (AMIS) $\mathcal{S}_\alpha$ by collecting the α observations with the largest individual influences. They extensive theoretical analysis of this quantity and suggest that for a small α, e.g., below 2.5% of the data, the approximation works well—in that the approximate impact of deleting $\mathcal{S}_\alpha$ is close to the actual impact of doing so—although it begins to break down for larger α (p. 18). Once $\mathcal{S}_\alpha$ has been identified, a next step is to delete those observations from the dataset, re-estimate the model and examine the impact on the quantities of interest. In our stylized example, we can examine how $\hat{\beta}_1$ behaves after deleting this set of observations.

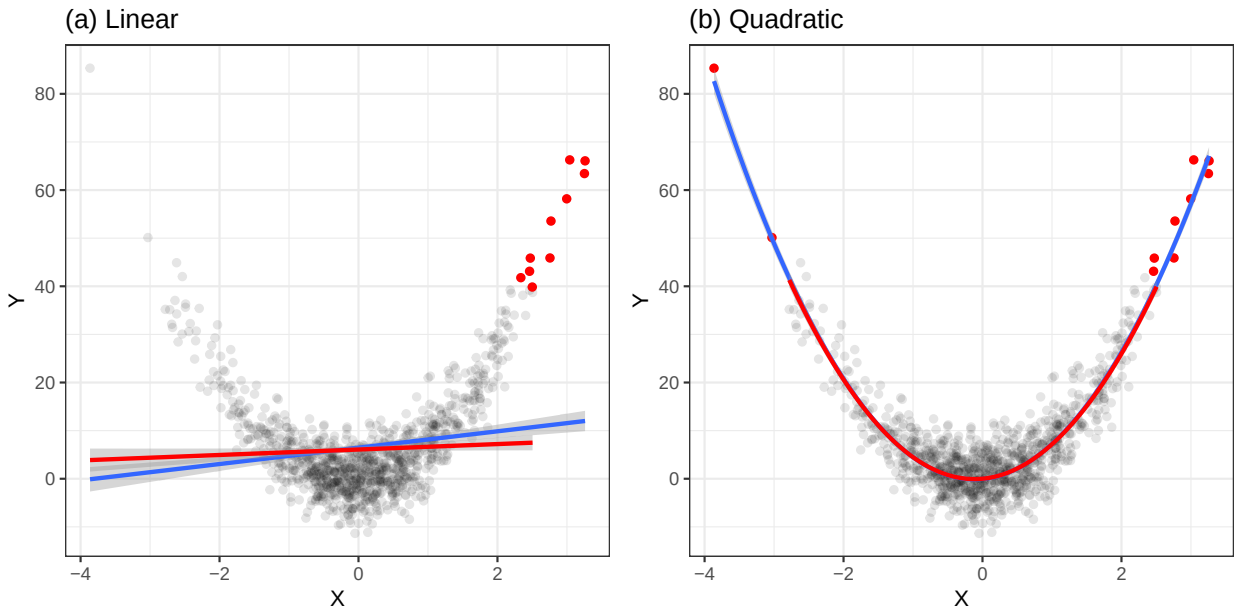
Figure 1a illustrates this by drawing $N = 1000$ observations from Equation 1 and fitting a linear model shown in blue. Setting $\alpha = 0.01N$, we identify $\mathcal{S}_\alpha$ and then refit the model after removing those observations. The removed points are colored in red and the refit linear model is shown in red.

The removed observations have large values of x_i and large residuals; their removal considerably changes the shape of the estimated regression line. This is not a problem of a small sample size and does not vanish asymptotically as N grows for $\alpha = 0.01N$.

Consider, however, fitting a quadratic model, i.e. $E[y_i|x_i] = \hat{\gamma}_0 + \hat{\gamma}_1 x_i + \hat{\gamma}_2 x_i^2$, computing an AMIS with the same α, chosen by the impact on the Wald statistic of $[\hat{\gamma}_1, \hat{\gamma}_2]^T = \mathbf{0}$ (see Section 4). The right panel of Figure 1 plots the original and refit quadratic curves after removing its AMIS; the shape of the curve is virtually unchanged.

While this example is highly stylized, it illustrates the point that non-robustness of an existing result can sometimes be attributed to model mis-specification. The crux of this paper, therefore, is

Figure 1: Impact of Removing Influential Observations



that non-robustness after removing a small AMIS should encourage the researcher to examine their model and data. One would hope that if a model could be specified such that no points are unduly influential, then removing a small AMIS should hopefully have limited impact on the quantities of interest to the researcher.

However, it is not the case that *all* non-robustness can be attributed to mis-specification. Broderick, Giordano and Meager (2023, p. 4) note that even for a correctly specified model, there may be non-robustness for a small α if there is a low signal-to-noise ratio (i.e., between the true value of the quantity of interest and the noise in the data). To make this more concrete, consider the linear regression setting: The "signal" would be true value of a regression coefficient, γ_1, and the noise is the standard deviation of the error variance $\sqrt{\sigma_\epsilon^2}$ divided by the standard deviation of the covariate $\{x_i\}_{i=1}^N$. This noise term does *not* go to zero as N goes to infinity—even though the standard error on $\hat{\gamma}_1$ does. Thus, if γ_1 is sufficiently small relative to a given α, for a large N, then it will be highly likely that a non-robust result will be detected *even if* $\hat{\gamma}_1$ is highly statistically significant. In the stylized case above, note that this low signal-to-noise ratio would manifest as a very low predictive power of the model (e.g., a low R^2), but a highly statistically significant coefficient.

2.1 Conceptualization of Robustness and Links to External Validity

Our notion of robustness in this paper is focused around a "worst case" scenario, insofar as we have identified a set of observations of size α that are designed to be *maximally* harmful to the original result. Deleting a randomly selected set of α would generally have a much smaller impact on the result. Thus, a non-robust result for a small α does not indicate that any—or even most!—perturbations in the data will undermine the stated result, but it indicates instead that there are "nearby" populations where the stated result does not hold. As a limiting case, if deleting a single observation were to make the confidence interval on a quantity of interest contain zero, it is hard to see how the study can accomplish the important goal of learning about more general populations (Findley, Kikuta and Denly, 2021). There are thus clear connections between our methodology finding "non-robustness" and certain conceptions of external validity (e.g., X-validity or population validity [Egami and Hartman 2023] or unit validity [Findley, Kikuta and Denly 2021]).

This connection is most evident by comparing our methodology and workflow to a recent paper by Devaux and Egami (2022). Their quantity of interest is the target population average treatment effect, and they focus on finding the closest set of sample weights that can "explain away" an estimated effect, i.e. cause the point estimate to become zero or change sign. They argue that if their measure finds that the effect is robust to a large amount of reweighting, then it has has high external robustness as it means there are many different populations for which the conclusion holds (Devaux and Egami, 2022, p. 8). By contrast, if a small amount of reweighting can change the result, then the result can only be generalized to quite similar populations to the original sample.

Our methodology has a similar philosophy and interpretation, although we focus on removing observations (i.e., setting weights to zero) versus more elaborate forms of weighting. Importantly, we also encourage researchers to be less focused on the amount of removal needed to "explain away" an existing effect but rather seeing how small perturbations affect the existing results.[2] Our main focus is also not on the robustness of specific causal quantities—although our approach could be used in experimental designs to assess external robustness (see also Broderick, Giordano and Meager 2023 for applications to experimental studies).

Rather, in the remainder of the paper, we use our method in observational studies to improve

[2]Our accompanying software allows researchers to look the amount of deletion needed to "explain away" an effect if desired.

robustness and illustrate the importance of model building and testing and leave more detailed exploration of the conceptual connections to external validity to future research. In the cases we examine, our methodology is aimed at encouraging researchers to dig into their model and see whether improvements can be made to address underlying issues that are revealed by non-robustness (see King and Roberts 2015 for a similar approach).

2.2 Addressing a Finding of Non-Robustness

A key question is what should a researcher do if they find that their result is non-robust to removing a small number of observations. Following advice from Broderick, Giordano and Meager (2023, p. 43), we *do not* suggest that one should exclude the observations in the AMIS from the reported results. Rather, in the same spirit as King and Roberts (2015), we suspect that non-robust results typically indicate some important type of unmodelled heterogeneity or model mis-specification that can be addressed by extending the proposed model. While the exact method for assessing non-robustness should depend on the specific analysis at hand and be guided by a researcher's substantive knowledge, we suggest three steps for exploring non-robustness: (i) characterizing the AMIS, (ii) relaxing key functional form assumptions, and (iii) estimating heterogeneous effects.

Characterizing the AMIS: The first step, characterizing the AMIS, notes that if there is some clear pattern as to which observations are influential, it is often possible to easily diagnose and address the source of the non-robustness. For example, in the stylized figure above, the fact that all points in the AMIS had large values of x_1 was a tell-tale sign of functional form mis-specification.

In a more complex setting, we recommend using simple visual inspections (e.g., is the distribution of key covariates rather different for those in the AMIS) alongside a simple and interpretable machine learning model (e.g., a single [shallow] decision tree). It is often helpful to include the residuals from the original model in the decision tree as these are often strong predictors of which observations are in the AMIS.

Relaxing Key Functional Form Assumptions: In some cases, however, characterizing the AMIS does not return a clear answer as to the source of non-robustness. In this case, our second suggestion is to estimate a model that is considerably more flexible but regularized to prevent over-fitting and maintains interpretability. While many such options exist, if one's initial model was a (generalized) linear model or a model with random effects, we suggest using generalized

additive models (Wood, 2017) for this initial robustification. As the source of non-robustness is often assuming a linear effect of a predictor—where the true effect is non-linear, we suggest using penalized splines on each continuous covariate. This will allow a non-linear effect to be estimated as well as increasing uncertainty in regions with limited data. If interactions are included in the original model, these can be also incorporated using factor-by-curve smooth terms or bivariate smoothers (Ruppert, Wand and Carroll 2003; Wood 2017).

Estimating Heterogeneous Effects: Even with the robustified model, it may still be the case that the main quantity of interest is non-robust. In that case, the final diagnostic test that we recommend is estimating heterogeneous effects. Put another way, there may be *interactions* between the key covariates of interest and other covariates that are being missed in the original model and are the source of the non-robustness. For this paper, we focus on how to use generalized additive models to estimate heterogeneous effects (see Section 5.2).

Overall, even these three steps may not lead to clear conclusions about the source of non-robustness. While there could be other model specification issues at play (e.g., distributional assumptions or stochastic components; King and Roberts 2015), one should not iterate on this process indefinitely. After performing a reasonable set of tests to assess the cause of non-robustness and if no clear cause appears, we recommend reporting both the original results, the results after removing the AMIS, and a discussion of the steps taken to adjust the initial model.

2.3 Existing and Best Practices for Assessing Non-Robustness

Before turning to new empirical applications of this method, it is useful to examine how it has been used in existing research to illustrate common practice and where some improvements could be made. We focus on two examples of recent papers in political science that have used the existing work by Broderick, Giordano and Meager (2023) to provide examples of how this might be used in practice. First, Martinez (2022, p. 2749-2950) says that

> I show in [appendix] table D16 that one must drop a subset of highly influential observations equivalent to 5.1% of the baseline sample in order for $\hat{\sigma}$ to become negligible and insignificant. This indicates that the results are not driven by a minuscule fraction of the sample.

In their paper $\hat{\sigma}$ is a non-linear transformation of a regression parameter; in the appendix, Martinez (2022) notes that they look at the AMIS to make the relevant regression parameter insignificant versus directly targeting $\hat{\sigma}$; the ability to target non-linear quantities of the regression parameters is an extension that our paper and accompanying software facilitates. In their appendix, Martinez (2022) also conducts some important additional explorations of (i) examining which observations are in the AMIS, (ii) noting a potential factor that might explain the AMIS's composition, and (iii) re-estimating the model addressing that concern and showing the results remain robust.

By contrast, Eubank and Fresh (2022, p. 801) find that their results lack robustness and contextualize this by noting that

> Appendix N presents data on the robustness of our conclusions to maximally influential perturbations (Broderick, Giordano, and Meager 2021). It finds that results do tend to fall below statistical significance with the removal of a relatively small number of the most influential observations (unsurprising given our sample size and significance levels), but we would need to selectively remove at least 19% of our data for the sign of our estimates to change, and no amount of data removal could generate statistically significant results in the other direction.

This discussion is careful and nuanced, although we would typically suggest against relying on the approximation method beyond around 2.5%. If the approximate method suggests an $\alpha > 0.025N$, we would recommend saying that there was not evidence of non-robustness, consistent with our focus less on "explaining away" and more on the impact of small perturbations. We would further suggest that the authors might examine whether some (reasonable) adjustment to the model would improve its robustness.

Finally, we note that care should be taken to note that there *could* be sets of observations that generate statistically significant results in the other direction, although they cannot be identified by the approximate heuristic used in Broderick, Giordano and Meager (2023) and this paper. The methodology of identifying approximately most influential sets and then seeing the robustness of their model to their exclusion is best thought of as quickly identifying a set of observations that might render a result non-robust (tested via refitting the model excluding those observations), but

it does *not* establish that no such set could exist as that would require either exhaustively searching over all sets or using other methodology.

3 Approximately Finding the Most Influential Set

We now formally outline our methodology and start by considering generalized linear models estimated using maximum likelihood. For concreteness, we focus here on logistic regression but Appendix B provides results for the general case. Equation 2 provides, under standard regularity conditions, the definition of the maximum likelihood estimate in a logistic regression with N observations whose outcomes y_i are modelled using covariates $\boldsymbol{x}_i$ and parameters $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{i=1}^{N} y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - \ln(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})). \tag{2}$$

To quantify uncertainty, it is standard to use the following variance-covariance matrix that uses the inverse of the negative Hessian evaluated at $\hat{\boldsymbol{\beta}}$,

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T \hat{p}_i (1 - \hat{p}_i) \right]^{-1} ; \quad \hat{p}_i = \frac{\exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})}{1 + \exp(\boldsymbol{x}_i^T \hat{\boldsymbol{\beta}})}. \tag{3}$$

Beyond simply examining $\hat{\boldsymbol{\beta}}$ and $\widehat{\operatorname{Var}}(\boldsymbol{\beta})$, researchers usually consider post-regression quantities of interest ϕ that depend on $\hat{\boldsymbol{\beta}}$ and $\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}})$, e.g. a confidence interval where $\phi = \hat{\boldsymbol{\beta}}_j \pm 1.96\sqrt{\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_j)}$.

The key focus of this paper is the influence of a set of observations $\mathcal{S} \subseteq \{1, \cdots, N\}$ on ϕ. The impact of removing those observations can be exactly quantified below, where the subscript $-\mathcal{S}$ denotes the estimates found without those observations in $\mathcal{S}$:

$$\operatorname{infl}(\phi, \mathcal{S}) = \phi\left(\hat{\boldsymbol{\beta}}_{-\mathcal{S}}, \widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_{-\mathcal{S}})\right) - \phi\left(\hat{\boldsymbol{\beta}}, \widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}})\right). \tag{4}$$

Outside of very special scenarios, computing the influence requires refitting the model. Finding the set of maximum influence, i.e. $\mathcal{S}^* = \sup_{\mathcal{S}} \operatorname{infl}(\phi, \boldsymbol{S})$, is clearly computationally infeasible for almost all realistic settings. To address this, Broderick, Giordano and Meager (2023) consider the *approximate* influence that can be computed after fitting the model on the entire dataset. They first augment the original maximum likelihood problem with a vector of observation weights $\boldsymbol{w}$,

$$\hat{\boldsymbol{\beta}}(\boldsymbol{w}) = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{i=1}^{N} w_i \left[y_i \boldsymbol{x}_i^T \boldsymbol{\beta} - \ln\left[1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta})\right]\right]; \widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}(\boldsymbol{w})) = \left[\sum_{i=1}^{N} w_i \boldsymbol{x}_i \boldsymbol{x}_i^T p_i(1-p_i)\right]^{-1}. \tag{5}$$

If all $w_i = 1$, i.e. $\boldsymbol{w} = \mathbf{1}$, then we recover the original estimator. For notational simplicity, we often write $\hat{\boldsymbol{\beta}}(\mathbf{1}) = \hat{\boldsymbol{\beta}}$ when doing so is not ambiguous, and similarly for $\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}})$ and other quantities. Setting $w_i = 0$ is equivalent to deleting the observation from the model; thus, one can think of $\hat{\boldsymbol{\beta}}_{-\mathcal{S}}$ in terms of $\hat{\boldsymbol{\beta}}(\boldsymbol{w}_{\mathcal{S}})$ where $\boldsymbol{w}_{\mathcal{S}}$ equals zero for all $i \in \mathcal{S}$ and one otherwise. Thus, the influence can be rewritten as follows

$$\operatorname{infl}(\phi, \mathcal{S}) = \phi\left(\hat{\boldsymbol{\beta}}(\boldsymbol{w}_{\mathcal{S}}), \widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}(\boldsymbol{w}_{\mathcal{S}}))\right) - \phi\left(\hat{\boldsymbol{\beta}}(\mathbf{1}), \widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}(\mathbf{1}))\right). \tag{6}$$

Taking the first order Taylor expansion of this around the observed data, i.e. $\boldsymbol{w} = \mathbf{1}$, yields an approximation to the influence function (Broderick, Giordano and Meager, 2023, p. 8)

$$\operatorname{infl}(\phi, \mathcal{S}) \approx \sum_{i=1}^{N} \left([\boldsymbol{w}_{\mathcal{S}}]_i - 1\right) \left[\frac{\partial \phi\left(\hat{\boldsymbol{\beta}}(\boldsymbol{w}), \widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}(\boldsymbol{w}))\right)}{\partial w_i}\right]_{\boldsymbol{w}=\mathbf{1}}. \tag{7}$$

To find the *approximate* most influential set of size α (i.e., where $\mathcal{S}$ contains exactly α observations), Broderick, Giordano and Meager (2023) maximize Equation 7 over $\mathcal{S}$ or, equivalently, over $\boldsymbol{w}$ where $w_i \in \{0, 1\}$ and $\sum_i w_i = \alpha$. It is straightforwardly shown that there is an exact solution for this approximate most influential set $\mathcal{S}_\alpha$ that collects the α observations with the most negative $\partial\phi/\partial w_i$ evaluated on the observed data.[3] They refer to $\partial\phi/\partial w_i$ as "influence scores" as $N\partial\phi/\partial w_i$ equals the empirical influence function for ϕ evaluated at $(\boldsymbol{x}_i, y_i)$ (Broderick, Giordano and Meager, 2023, p. 23). Other ways of improving the approximation could be explored in future research (e.g., Kuschnig, Zens and Cuaresma 2021).[4]

Thus, the approximate most influential set $\mathcal{S}_\alpha$ can be computed by (i) estimating the model

[3]The goal is to maximize Equation 7; $\boldsymbol{w} = 1$ has a value of zero. Setting $w_i = 0$ adds $-\partial\phi/\partial w_i$ to the objective. Thus, to maximize the objective given exactly α w_i must be zero (the rest must be one), one selects the α w_i with teh most negative $\partial\phi/\partial w_i$. Note that if fewer than α $\partial\phi/\partial w_i$ are negative, by convention $\mathcal{S}_\alpha$ contains fewer than α elements.

[4]For example, Kuschnig, Zens and Cuaresma (2021) suggests an alternative procedure: Find the AMIS for α/K. Then, remove this set and find the AMIS of the *refitted* model; remove an AMIS of α/K. Repeat this procedure until α observations are removed. This requires fitting the model K times but it is likely to improve the identification of influential sets.

on the full dataset, (ii) computing the derivative of the quantity of interest ϕ with respect to the observation weights $\boldsymbol{w}$, evaluated at $\boldsymbol{w} = \mathbf{1}$, and (iii) selecting the α observations with the most negative derivatives to create $\mathcal{S}_\alpha$. If multiple ϕ are of interest, one can repeat this procedure without re-estimating the original model and performing steps (ii) and (iii).

The key to this method is the tractability of $\partial\phi/\partial w_i$, expressed using the chain rule below,

$$\frac{\partial \phi\left(\hat{\boldsymbol{\beta}}(\boldsymbol{w}), \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}(\boldsymbol{w}))\right)}{\partial w_i} = \frac{\partial \phi}{\partial \hat{\boldsymbol{\beta}}}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial w_i} + \mathrm{tr}\left[\frac{\partial \phi}{\partial \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})}\frac{\partial \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})}{\partial w_i}\right]. \tag{8}$$

This illustrates that for many ϕ the building blocks are the derivatives of the estimator and its covariance matrix with respect to w_i alongside a ϕ-specific derivative for the quantities of interest under consideration. Our accompanying software is structured to first find the derivatives with respect to these key building blocks and then allow the user to re-use these when estimating the derivatives of many quantities of interest, e.g., the confidence interval for all $\hat{\boldsymbol{\beta}}$. The following sub-sections derive these building blocks while the following section tackles different choices of ϕ.

3.1 Derivatives of Point Estimates and Variance Matrix

Consider first the derivative of the point estimates with respect to some observation w_i. Even though $\hat{\boldsymbol{\beta}}$ has no closed form—i.e. it can only be computed using an iterative procedure, the implicit function theorem provides a closed-form expression of its derivative with respect to w_i; Appendix B provides a detailed discussion. In the logistic regression setting, it can be expressed as

$$\frac{\partial \hat{\boldsymbol{\beta}}}{\partial w_i} = \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{x}_i\left[y_i - p_i\right]. \tag{9}$$

The derivative of $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ can be computed as follows, noting that the derivative of $\boldsymbol{A}(x)^{-1}$ with respect to x is $-\boldsymbol{A}(x)^{-1}\frac{\partial \boldsymbol{A}(x)}{\partial x}\boldsymbol{A}(x)^{-1}$ (Magnus and Neudecker, 2019),

$$\frac{\partial \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})}{\partial w_i} = -\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\left[\boldsymbol{H}_{i,1} + \boldsymbol{H}_{i,2}\right]\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) \tag{10}$$

$$\boldsymbol{H}_{i,1} = \boldsymbol{x}_i^T p_i(1-p_i)\boldsymbol{x}_i^T; \quad \boldsymbol{H}_{i,2} = \sum_{i'=1}^{N}\boldsymbol{x}_{i'}\left[p_{i'}(1-p_{i'})(1-2p_{i'})\boldsymbol{x}_{i'}^T\frac{\partial \hat{\boldsymbol{\beta}}}{\partial w_i}\right]\boldsymbol{x}_{i'}^T. \tag{11}$$

This reveals that the impact of the variance depends on both a direct effect of moving observation w_i on the Hessian but also the fact that changing w_i affects *every* observation's contribution to the (negative) Hessian, $\boldsymbol{x}_{i'} p_{i'}(1-p_{i'})\boldsymbol{x}_{i'}^T$, through its impact on $\hat{\boldsymbol{\beta}}$.

Naively applying this formula in downstream quantities, however, can result in an algorithm that is extremely computationally intensive. For example, if the model has p covariates, this requires forming a possibly enormous $p^2 \times N$ matrix to hold all relevant partial derivatives and has a computational cost that grows on the order of Np^3.

Fortunately, Appendix B notes that we rarely require the full variance matrix itself but focus on the diagonal elements (e.g., to compute standard errors) or quadratic forms with this matrix, i.e. $\boldsymbol{a}^T\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{a}$. It provides a critical result (Proposition B.2) establishing that the derivative of those quantities can be computed with a computational complexity that scales linearly in N and p. Indeed, all of the quantities of interest discussed in this paper can be evaluated without ever explicitly forming a matrix of size $p^2 \times N$.

3.2 Extensions Beyond the Standard Regression Model

The framework we describe above can be extended in many ways; we consider two, below, that are implemented in the accompanying software and are popular in many political science applications.

Clustered Standard Errors: It is common to use an alternative estimator for $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$ to account for clustering and/or heteroskedasticity in the data. The above formula can be straightforwardly adapted to the case of (multiway) clustered standard errors, see Appendix C.

Random Effects: It is common to add random effects or other hierarchical terms (e.g., splines) to a generalized linear model to account for unobserved heterogeneity or non-linear functional forms of certain covariates. Appendix D treats the most general form of this in detail (i.e., generalized additive models; Wood 2017), but we discuss this in the case of a single random effect.

Assume that we add a random intercept to Equation 2, i.e., each observation i belongs to some group g where a coefficient α_g for each group g is added to the model where $\alpha_{g[i]}$ selects the corresponding α_g for each i. Collecting these into $\boldsymbol{\alpha}$, the distinguishing feature of a random intercept model is that there is some amount λ of regularization (or, equivalently, a prior) on all $\boldsymbol{\alpha}$ (see Hazlett and Wainstein 2022 and Chang and Goplerud 2024 for recent discussions in political science). As λ grows to infinity, the regularization disappears and $\boldsymbol{\alpha}$ become traditional "fixed

effects"; as λ approaches zero, then all $\hat{\boldsymbol{\alpha}}$ will be estimated to be zero. Formally, and mirroring Chang and Goplerud (2024, p. 160), we can write the (penalized) maximum likelihood estimates of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ as a function of $\boldsymbol{w}$ and λ,

$$\hat{\boldsymbol{\beta}}_\lambda(\boldsymbol{w}), \hat{\boldsymbol{\alpha}}_\lambda(\boldsymbol{w}) = \underset{\boldsymbol{\beta},\boldsymbol{\alpha}}{\operatorname{argmax}} \sum_{i=1}^{N} w_i \left[y_i \left(\boldsymbol{x}_i^T \boldsymbol{\beta} + \alpha_{g[i]} \right) - \ln \left(1 + \exp(\boldsymbol{x}_i^T \boldsymbol{\beta} + \alpha_{g[i]}) \right) \right] - \frac{1}{2\lambda} \sum_{g=1}^{G} \alpha_g^2. \quad (12)$$

A critical question is how to tune λ; it is typically done by minimizing some function $g(\lambda)$ that almost invariably depends on $\hat{\boldsymbol{\beta}}_\lambda(\boldsymbol{w})$ and $\hat{\boldsymbol{\alpha}}_\lambda(\boldsymbol{w})$. The most common approaches, implemented in our software and described in detail in Appendix D, are generalized cross-validation (GCV) or restricted maximum likelihood (REML). Wood (2017) provides an extensive overview of these methods and the general theory behind this approach to calibrating λ.

At a high level of generality, however, one obtains $\hat{\lambda}(\boldsymbol{w}) = \operatorname{argmin}_\lambda g(\lambda; \boldsymbol{w})$. By careful use of the implicit function theorem and tedious algebra, Appendix D shows that $\partial \hat{\lambda}(\boldsymbol{w}) / \partial w_i$ can be efficiently computed. However, $\hat{\lambda} = \hat{\lambda}(\mathbf{1})$ is usually not of direct interest and thus the key quantity is the derivative of the estimator $\hat{\boldsymbol{\beta}}_{\hat{\lambda}(\boldsymbol{w})}(\boldsymbol{w})$ and its covariance with respect to w_i, evaluated at $\boldsymbol{w} = \mathbf{1}$. It is critical to note that this derivative must take into account (i) the direct impact that changing w_i has on $\hat{\boldsymbol{\beta}}_\lambda$ and (ii) the indirect impact that changing w_i has on the optimal amount of smoothing $\hat{\lambda}$ that itself affects $\hat{\boldsymbol{\beta}}_{\hat{\lambda}}$. Equation 13 decomposes for the point estimates using the chain rule

$$\left[\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\lambda}(\boldsymbol{w})}(\boldsymbol{w})}{\partial w_i} \right]_{\boldsymbol{w}=\mathbf{1}} = \left[\frac{\partial \hat{\boldsymbol{\beta}}_\lambda(\boldsymbol{w})}{\partial w_i} \right]_{\lambda=\hat{\lambda}(\mathbf{1}),\boldsymbol{w}=\mathbf{1}} + \left[\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\lambda}}(\boldsymbol{w})}{\partial \hat{\lambda}} \right]_{\lambda=\hat{\lambda}(\mathbf{1}),\boldsymbol{w}=\mathbf{1}} \left[\frac{\partial \hat{\lambda}(\boldsymbol{w})}{\partial w_i} \right]_{\boldsymbol{w}=\mathbf{1}}. \quad (13)$$

The impact on the variance can be computed, as in Proposition B.1, although there is an additional term that comes from the direct dependence of $\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_{\hat{\lambda}})$ on $\hat{\lambda}$ that is missing in the standard regression setting; Appendix D provides details. It also contains a proposition (Proposition D.2) that demonstrates that the earlier results on computational scalability for the generalized linear model, i.e. never evaluating or forming a $p^2 \times N$ matrix, continues to hold.

4 Robustness of Post-Regression Quantities of Interest

There are many post-regression quantities of interest that one might report to provide substantively interpretable summaries of the regression model (King, Tomz and Wittenberg, 2000). We focus on the most popular ones here, although our methodology can be extended more broadly. Assume, from above, that we have estimated a set of regression coefficients $\hat{\boldsymbol{\beta}}$ and their associated covariance matrix $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$. We consider three popular types quantities that researchers report.

4.1 Single-Parameter Statistics

It is common to report coefficients and confidence intervals on single variables of interest, e.g.,

$$\text{CI}_j = \hat{\boldsymbol{\beta}}_j \pm 1.96\sqrt{\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_j)}; \quad t_j = \frac{\hat{\boldsymbol{\beta}}_j}{\sqrt{\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_j)}}. \tag{14}$$

In terms of our goal of quantifying robustness of these quantities, one straightforward approach is to see how deleting the approximately most influential set would pull t_i toward zero or the lower bound of CI_i towards zero if $\hat{\boldsymbol{\beta}}_i$ is positive (Broderick, Giordano and Meager, 2023). This would allow us to see whether deleting a small number of observations might undermine the reported statistical significance of an effect.

The approximate influence can be computed by applying the chain rule for differentiation by noting that the impact of deleting an observation i flows through its impact on *both* $\hat{\boldsymbol{\beta}}$ and $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$. In the two cases in Equation 14, the derivatives, and thus the building block for the approximate most influential set, are

$$\frac{\partial \text{CI}_j}{\partial \boldsymbol{w}} = \frac{\partial \hat{\boldsymbol{\beta}}_j}{\partial \boldsymbol{w}} \pm \frac{1.96}{2}\left[\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_j)\right]^{-1/2}\frac{\partial \text{Var}(\hat{\boldsymbol{\beta}}_j)}{\partial \boldsymbol{w}} \tag{15a}$$

$$\frac{\partial t_j}{\partial \boldsymbol{w}} = \frac{\partial \hat{\boldsymbol{\beta}}_j}{d\boldsymbol{w}} \cdot \frac{1}{\sqrt{\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})}} - \frac{\hat{\boldsymbol{\beta}}_j}{2}\left[\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_j)\right]^{-3/2}\frac{\partial \text{Var}(\hat{\boldsymbol{\beta}}_j)}{\partial \boldsymbol{w}}. \tag{15b}$$

Straightforward extensions to other single-parameter quantities (e.g., p-values) follow by use of the chain rule.

4.2 Multiple-Parameter Tests

However, it is often common to go beyond looking a single parameter and look at statistics that depend on multiple components of $\boldsymbol{\beta}$. We consider two here in detail; the first is a Wald statistic. This is used in testing whether (non-linear) combinations of parameters are zero. In the simple linear combination setting, we consider $\boldsymbol{A}\hat{\boldsymbol{\beta}}$ where $\boldsymbol{A}$ is some matrix of restrictions that is assumed to be full rank.[5] The test statistic is shown below,

$$\widehat{W} = \left[\boldsymbol{A}\hat{\boldsymbol{\beta}}\right]^T \left[\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{A}^T\right]^{-1} \boldsymbol{A}\hat{\boldsymbol{\beta}}; \quad \widehat{\text{Var}}\left(\boldsymbol{A}\hat{\boldsymbol{\beta}}\right) = \boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{A}^T, \tag{16}$$

and is a function of both the parameter estimates and the estimated covariance. In our setting of looking at the approximate robustness, one might wish explore whether deleting a small number of observations substantially *decreases* the Wald statistic. Deleting this AMIS is thus aimed at making the Wald test not statistically significant or, more generally finding a set such that, when deleted, $\boldsymbol{A}\hat{\boldsymbol{\beta}}$ is closer to zero. By again applying the chain rule, we can compute the derivative of $\widehat{W}$ with respect to a single observation w_i is:

$$\frac{\partial \widehat{W}}{\partial w_i} = 2\boldsymbol{A}^T \left[\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{A}^T\right]^{-1} \boldsymbol{A}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{w}_i} - \left[\boldsymbol{A}\hat{\boldsymbol{\beta}}\right]^T \left[\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{A}^T\right]^{-1} \boldsymbol{A}\left[\frac{\partial \widehat{\text{Var}}(\hat{\boldsymbol{\beta}})}{\partial w_i}\right] \boldsymbol{A}^T \left[\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{A}^T\right]^{-1} \left[\boldsymbol{A}\hat{\boldsymbol{\beta}}\right] \tag{17}$$

The derivation is more complex and is discussed in Appendix E; it can be efficiently computed for all $\boldsymbol{w}$, i.e. $\partial\widehat{W}/\partial\boldsymbol{w}$, using vectorization and Kronecker products. If non-linear restrictions are desired, this can also be done using a slight generalization of the above result.

Another popular test involving multiple parameters is a likelihood ratio test; this can be seen as fitting a restricted model, e.g., where some component of $\boldsymbol{\beta}$ is set to zero, and then comparing the difference in the estimated likelihoods. Appendix E provides details.

4.3 Predicted Outcomes and Marginal Effects

The above two sub-sections consider only functions of $\hat{\boldsymbol{\beta}}$. However, the most popular quantities of interest, especially for models with non-Gaussian outcomes, involve creating predictions of the out-

[5] The non-full rank case can be dealt with using a generalized inverse of the middle term; see Appendix E.

come on the original scale (e.g., predicted probabilities) or closely associated quantities (e.g., "first differences" or "average marginal effects"); see King, Tomz and Wittenberg (2000) for discussion.

Formally, consider making N_0 predictions post-estimation, e.g., to create a curve that plots predicted probabilities or marginal effects. Assume there is some function $h(\cdot)$ with derivative $h'(\cdot)$ that produces these functions for input $\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$, e.g. $h(\boldsymbol{x}_0^T\boldsymbol{\beta}) = \exp(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}})/[1+\exp(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}})]$ to produce a predicted probability after a logistic regression. We can define the vector of predictions $\hat{\boldsymbol{h}}_0$ that comes from applying $h(\cdot)$ element-wise to $\boldsymbol{X}_0\hat{\boldsymbol{\beta}}$. As is standard, we approximate the variance of $\hat{\boldsymbol{h}}_0$ using the (multivariate) delta method.

$$\hat{\boldsymbol{h}} = h(\boldsymbol{X}_0\hat{\boldsymbol{\beta}}); \quad \widehat{\mathrm{Var}}(\hat{\boldsymbol{h}}) \approx \boldsymbol{J}_h\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{J}_h^T; \quad \boldsymbol{J}_h = \boldsymbol{X}_0^T \mathrm{diag}\left(h'(\boldsymbol{X}_0\hat{\boldsymbol{\beta}})\right) \tag{18}$$

While this would be sufficient to plot a predicted probability curve, for example, we often wish to go further and take differences between elements of $\hat{\boldsymbol{h}}_0$ or other linear combinations $\boldsymbol{A}\hat{\boldsymbol{h}}$. For example, if one wished to calculate a first difference between two predicted probabilities, one would compute $\hat{\boldsymbol{h}}_1 - \hat{\boldsymbol{h}}_0$. The standard way that the derivative of a curve is produced is by a finite difference method between two close values (e.g., Leeper 2016) which can also be expressed by a particular choice of $\boldsymbol{X}_0$ and $\boldsymbol{A}$. Thus, we also consider $\boldsymbol{A}\hat{\boldsymbol{h}}$, defined below with the accompanying variance estimator

$$\boldsymbol{A}\hat{\boldsymbol{h}} = \boldsymbol{A}h(\boldsymbol{X}_0\hat{\boldsymbol{\beta}}); \quad \widehat{\mathrm{Var}}(\boldsymbol{A}\hat{\boldsymbol{h}}) \approx \boldsymbol{A}\boldsymbol{J}_h\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{J}_j^T\boldsymbol{A}^T \tag{19}$$

Given $\hat{\boldsymbol{h}}$ and $\boldsymbol{A}\hat{\boldsymbol{h}}$, one can conduct similar tests to the single-parameter and multi-parameter tests above (e.g., computing confidence intervals and Wald statistics). Appendix E derives the analogous derivatives of $\hat{\boldsymbol{h}}$ and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{h}})$ with respect to the weights of each observation. Proposition E.3 shows that the derivative of a Wald statistic involving $\boldsymbol{A}\hat{\boldsymbol{h}}$ can be computed in $O(N_0p^2 + Np + N_0^3)$. Thus, as most quantities of interest involve only modest N_0, computing the derivatives of these quantities remains computationally efficient.

In terms of testing robustness of post-estimation quantities of interest, this additional notation allows testing robustness of quantities such as the confidence interval on a first difference or average marginal effect. One might also examine whether an entire *curve* is sensitive to the deletion of a few observations. To do this, we propose the following procedure: For a predicted probability curve $\hat{\boldsymbol{h}}$,

compute its derivative at each point using the finite difference method commonly implemented in existing software (e.g., Leeper 2016; Chang and Goplerud 2024); call this $\hat{\boldsymbol{h}}'$. If the curve was flat, i.e. the covariate of interest had no effect, then the derivaive would be zero at all points. Using this insight, create a Wald statistic—with appropriate rank, see Appendix E—that would test the null hypothesis $\hat{\boldsymbol{h}}' = \mathbf{0}$. Finally, find the AMIS by looking at the derivative of this Wald statistic with respect to $\boldsymbol{w}$ and then examine the deletion of this set on the refitted model and accompanying curve.

5 Empirical Applications

To illustrate our method, we examine the robustness of the analyses by Giger and Klüver (2016) and Woldense and Kroeger (2024). These papers include a number of post-estimation quantities of interest commonly reported in similar analyses in political science. We examine the robustness of some of the key claims in each paper, showing how to address evidence of non-robustness using model extensions such as allowing for non-linear functional forms or modelling heterogeneous effects.

5.1 Giger and Klüver 2016

Giger and Klüver (2016) examine whether the affiliation of legislators with different types of interests groups impacts their likelihood of voting in line with their constituents' preferences. They expect that MPs with more associations with interest groups that focus on subgroups in the population (e.g., "sectional groups" such as farmer's associations) should be more likely to vote in ways that do not align with their constituents' preferences. By contrast, MPs with more associations with groups that represent broad appeals (e.g., "cause groups" such as environmental movements) should be less likely to vote in a way that does not align with their constituents' preferences.

To test these, they use Swiss data that merges public referenda outcomes with parliamentary roll-call votes on identical policy proposals where their primary model is a logistic regression that predicts whether an MP "defects" (i.e., votes in a way that disagrees with their constituents' vote) as a function of the number of sectional and cause groups to which an MP is affiliated, as well as a number of controls and random effects for the MP's party and canton.

We first begin by showing the robustness of each of the two coefficients of primary interest

(sectional groups and cause groups) reported in their original model. Table 1 shows the impact of deleting various AMIS. For each cell, the AMIS for that specific coefficient and α is determined using the approximate method; the presented coefficients are the re-estimation of the original model after removing the AMIS.

Table 1: Approximate Maximum Influence Sets (AMIS) for Giger and Klüver (2016)

	Sectional Groups	Cause Groups
Original	[0.00203, 0.023]	[-0.0517, -0.0188]
$\alpha = 0.0001N$	$[0.000711, 0.0218]$ (✓)	$[-0.0505, -0.0175]$ (✓)
$\alpha = 0.001N$	$[-0.00561, 0.016]$ (×)	$[-0.0446, -0.0114]$ (✓)
$\alpha = 0.005N$	$[-0.0282, -0.00507]$ (×)	$[-0.0254, 0.0086]$ (×)
$\alpha = 0.01N$	$[-0.0534, -0.0286]$ (×)	$[-0.00509, 0.0297]$ (×)
$\alpha = 0.015N$	$[-0.0794, -0.0527]$ (×)	$[0.015, 0.0507]$ (×)
$\alpha = 0.02N$	$[-0.107, -0.0783]$ (×)	$[0.0345, 0.071]$ (×)

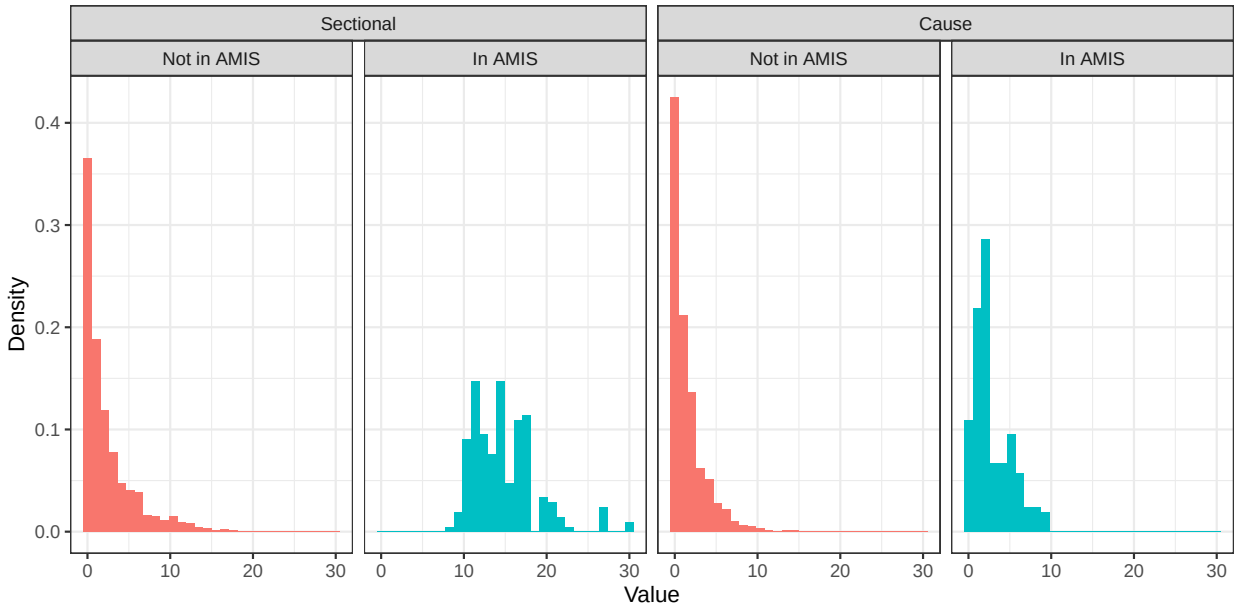
Note: This table shows the refitted model and corresponding 95% confidence interval after deleting the AMIS for the coefficient labelled in each column. Different α-levels as proportions are shown where $N = 20,260$. A ✓ indicates that the refit confidence interval does not contain zero and has the same sign as the original estimate; a × indicates a non-robusut result.

It shows that both of the main effects are relatively non-robust to small deletions of the data. The coefficient on sectional groups can be made insignificant after deleting $\alpha = 0.001N$ observations, i.e. 21 observations. The coefficient on cause groups can be made insignificant after deleting the AMIS found by $\alpha = 0.005N$. Indeed, it is possible to create significant coefficients in the *opposite* direction by deleting sets of no more than $0.015N$.

Given this result, we explore possible causes and adjustments to the original model. Per the discussion in Section 2 about steps to take given a finding of non-robustness, we first try to characterize the AMIS using simple bivariate summaries. We start by looking at the key independent variables of interest where Figure 2 shows the distribution of the sectional and cause group variables for $\alpha = 0.01N$ by whether the observation is in the AMIS. Both distributions are highly skewed; for sectional groups, the AMIS at $\alpha = 0.01N$ contains almost all of the very large values. For cause groups, the AMIS contains a slightly higher distribution of those values but the difference is less

pronounced.

Figure 2: Distribution of Section and Cause Group Variable



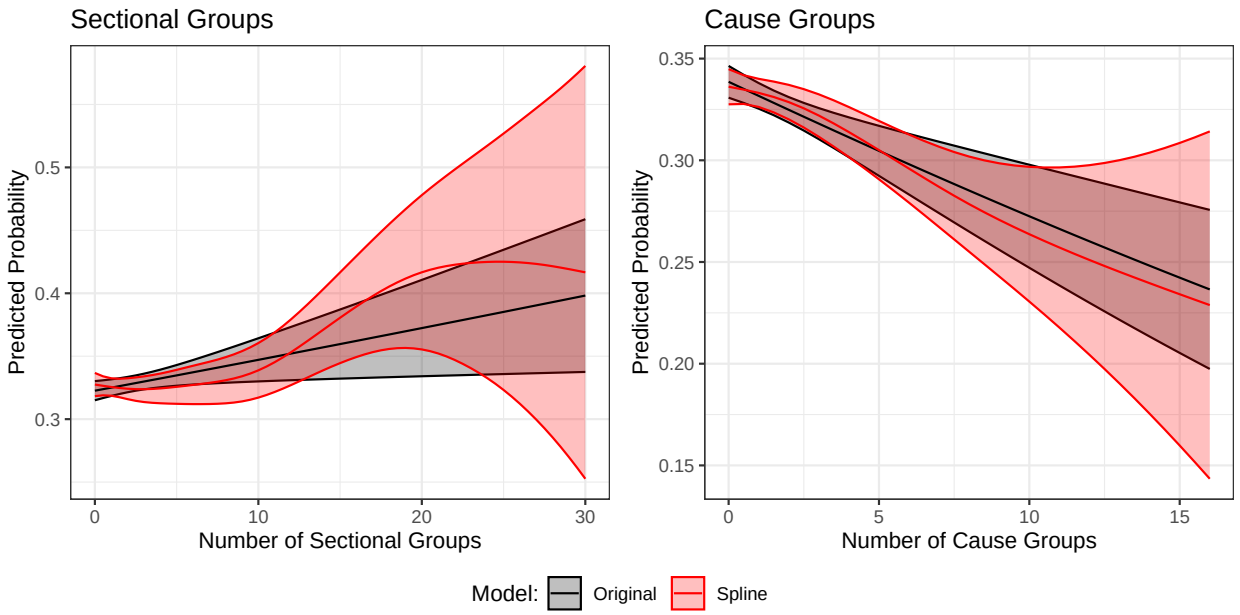
More formally, we fit a decision tree with a maximum depth of three to examine who is in the AMIS. The terminal nodes of the tree, and rules for characterizing membership, are shown below—with the percent of observations in the AMIS that are in each node ("% AMIS"), the share of the entire dataset in the node ("% N"), and the proportion of the node in the AMIS ("Pr").

```
-----Tree for Sectional Groups-----
 % AMIS %N   Pr
 53.20  0.54 98.18: sect_no >= 14 & residual >= 0.45
 45.81 98.35  0.47: sect_no <  14
  0.99  1.11  0.89: sect_no >= 14 & residual <  0.45
----Tree for Cause Groups-----
 % AMIS %N    Pr
 38.92  0.46 84.95: cause_no >=   7 & residual <  -0.39
 31.03  4.52  6.89: cause_no >=   6 & residual >= -0.39
 24.63  0.53 46.30: cause_no is 6 to 7 & residual <  -0.39
  5.42 94.49  0.06: cause_no <  6
```

Both tree confirm the visual analysis: 53% of the AMIS is located in a node with more than 14 sectional group memberships and a positive residual—although this node is only 0.54% of the entire dataset. For cause groups, nearly 40% of the AMIS is located in a small node with many cause groups and a negative residual.

A reasonable conjecture is that the linearity assumption of each variable is likely the source of the problem—given that observations with large values and thus high leverage are in the AMIS. Following the discussion in Section 2, we rely on generalized additive models with penalized splines on both economic and cause groups to try to address this problem. After having estimated this model, one can construct a predicted probability curve using the observed value approach (Hanmer and Kalkan 2013), mirroring Figures 3 in Giger and Klüver (2016). Figure 3 shows these curves for the original model and the model with a spline.

Figure 3: Predicted Probability Curves with Spline Model



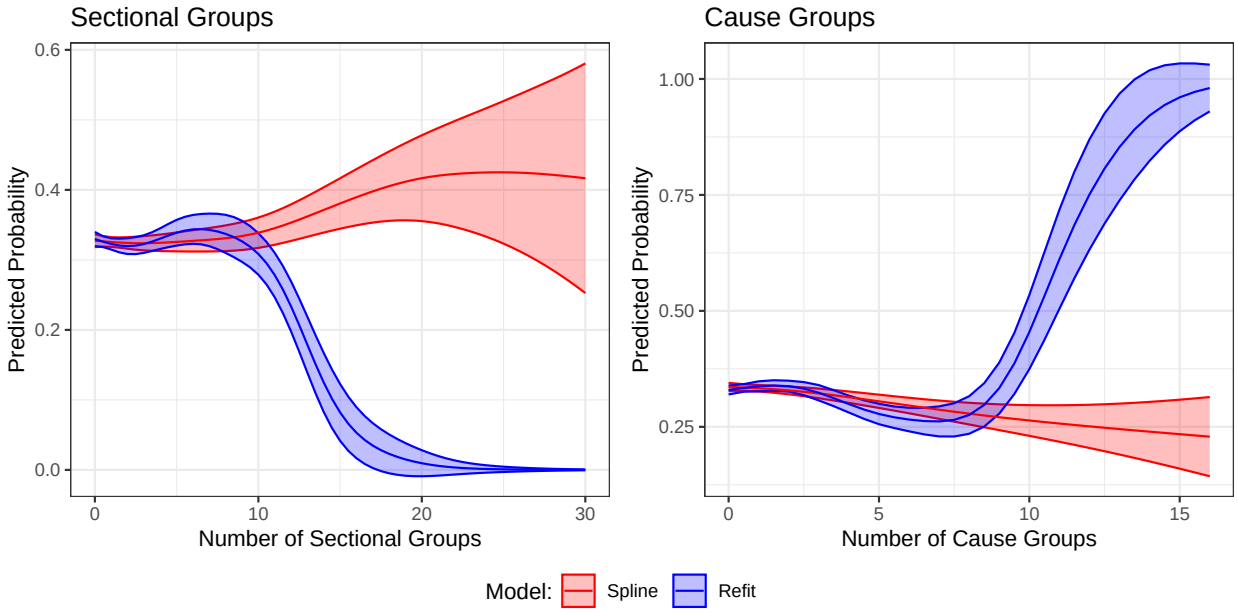
Focusing first on sectional groups, in the range where most of the data is observed (e.g., 95% of observed values of sectional groups are less than or equal to ten), the curve is effectively flat. For larger values, there is some evidence of a positive effect but the standard errors increase considerably. For cause groups, most of the observed values lie between zero and five (95%); this still shows a negative effect—consistent with the original paper.

Thus, using a more flexible functional form suggests that while the results on cause groups

remains robust, there is limited evidence for the effect of sectional groups on MP voting behavior for the majority of MPs.

Finally, we examine the robustness of the spline model. We calculate the Wald statistic for the derivative of the predicted probability curve, using a generalized Wald test with a rank equal to the effective degrees of freedom of the corresponding spline (see Section 4 and Appendix E), and remove observations corresponding to an AMIS of size $\alpha = 0.01N$. Figure 4 plots the original and refit curves.

Figure 4: Refitting Spline Model



For the range of cause and sectional groups where most of the data is observed (i.e,. 0-10 for cause groups and 0-5 for sectional groups), the curves are quite similar—albeit more flat for cause groups. At very large values, the behavior of the spline model and the model excluding the AMIS differ quite considerably; this is expected as removing limited data in those regions should cause large changes in the "tails" of the spline. Thus, for the region where most of the data lies, the predicted probability curves produced by the spline models are considerably more robust than the one with a linear functional form on sectional and cause groups.

5.2 Woldense and Kroeger 2024

Woldense and Kroeger (2024) seek to examine the nature of elite alliances in African electoral

autocracies in the aftermath of the Cold War. By relying on cabinets as proxies for power-sharing arrangements between rulers and elites, the authors find evidence to suggest that governing elite coalitions experienced significant reshuffling at the end of the Cold War (e.g., from 1988-1992) compared to periods before (e.g., 1966-1988) and after (e.g., 1993-2010). They suggest that this unusually high turnover within the governing coalition constituted a survival strategy for rulers amidst geopolitical uncertainty.

Methodologically, their main results build on logistic regressions to predict whether a minister exits a cabinet in a given year (e.g., 1) or remains into the next year (0). A variety of controls are included linearly with two—the number of years that a minister has served and the number of years that the leader has served—are included using cubic polynomials. The key variable of interest is whether the year is in the "Cold War End Window" (CWE), i.e. 1988-1992. Their expectation is that this variable has a positive effect, as they expect higher exit probabilities in the CWE period, controlling for other covariates.

We focus on replicating their main results: the average marginal effect of their key treatment variable, CWE, holding all other covariates at their median (if continuous) or zero (if binary), and a predicted probability curve of the probability of a minister exiting the government as a function of minister tenure, again holding all other covariates at their median or zero.

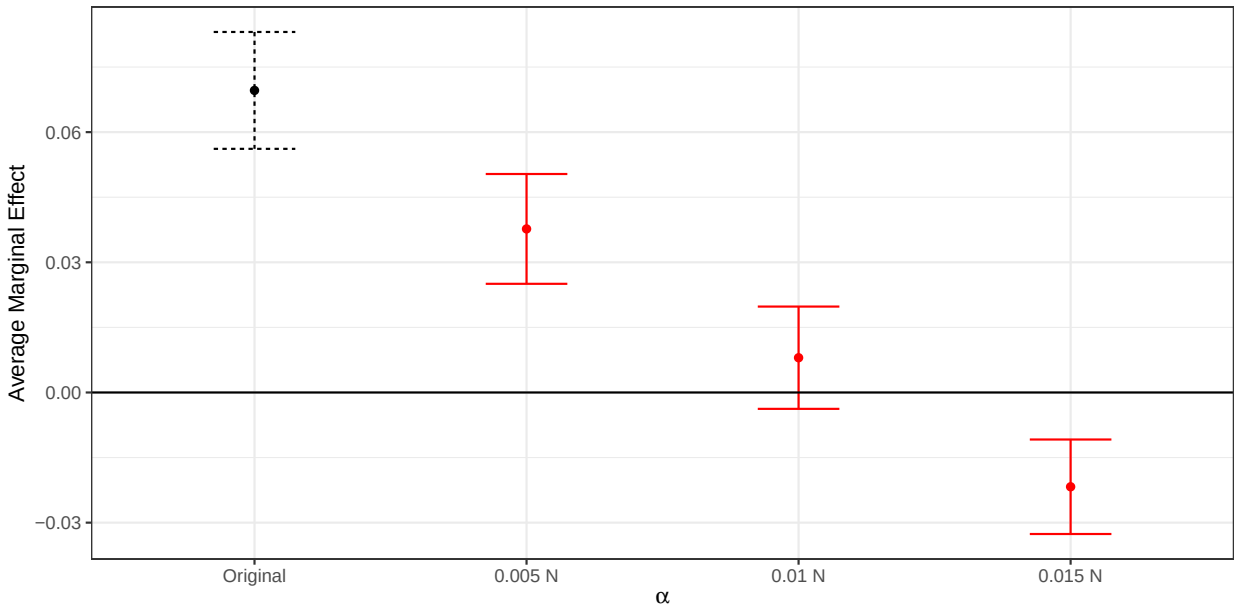
Robustness of the Average Marginal Effect: Figure 5 shows the refitted models after identifying the AMIS based on the Wald statistic (i.e., the squared t-statistic) of the estimated marginal effect.

It shows that deleting a set of $\alpha = 0.01N$ is sufficient to make the reported average marginal effect insignificant. To explore this further, we examined the observations that were in the AMIS for $\alpha = 0.01N$. Visual inspections revealed that (i) all observations in the AMIS were in the Cold War Ends period (i.e., in the "treatment" group) and (ii) had large positive residuals.

A more formal test using a decision tree, limited to have a maximum depth of four, led to the following rules for predicting which observations were in the AMIS:

```
% AMIS  % N  Pr
69.72  0.78 89.06: CWE_5 is 1 & resid >=   0.73 & min_tenure <  9     & lea_tenure >= 4
 9.48  0.12 79.49: CWE_5 is 1 & resid is 0.68 to 0.73     & cabinet_size >= 37
```

Figure 5: Robustness of Average Marginal Effect of CWE




```
8.87  0.80 11.11: CWE_5 is 1 & resid is 0.68 to 0.73     & cabinet_size <  37
6.42  0.09 67.74: CWE_5 is 1 & resid >=   0.81 & min_tenure >= 9
4.59  0.15 31.25: CWE_5 is 1 & resid >=   0.73 & min_tenure <  9     & lea_tenure <  4
0.61  0.14  4.35: CWE_5 is 1 & resid is 0.73 to 0.81 & min_tenure >= 9
0.31 10.71  0.03: CWE_5 is 1 & resid <  0.68
0.00 87.20  0.00: CWE_5 is 0
```

This provides a less clear characterization than in Section 5.1, but a few points are worth noting: First, all observations in the AMIS have CWE equal to one. Second, nearly 70% of the AMIS is located in a leaf where there is a large positive residual, minister tenure is low but leader tenure is rather higher.

Especially given the finding that all units in the AMIS are in the "treated" group, we proceed directly to the third technique for addressing non-robustness discussed in Section 2—estimating heterogeneous effects. Some heterogeneous effects analyses were also explored in other sections of Woldense and Kroeger (2024). We note that observations in the AMIS tend to have large cabinets. Upon examining the log oil rents variable, we note that it has a distribution where 66% have a value of zero and the remaining 33% have values centered around 21. To address this, we add a dummy variable for "zero log oil rents" and then create a standardized variable that (i) gives all

observations with zero rents a value of zero and (ii) subtracts the mean of the non-zero values from the observations with a non-zero value.

Then, to model the heterogeneity, we use a special type of generalized additive model, i.e. a semi-parametric model known as "factor-by-curve" models (Ruppert, Wand and Carroll, 2003) where a spline is used for each continuous variable interacted with each level of CWE, as well as including interactions between all binary variables and CWE. This is the flexible analogue to interacting all covariates with the treatment variable and thus implements a version of the "T-Learner" (Künzel et al., 2019) in a single model.

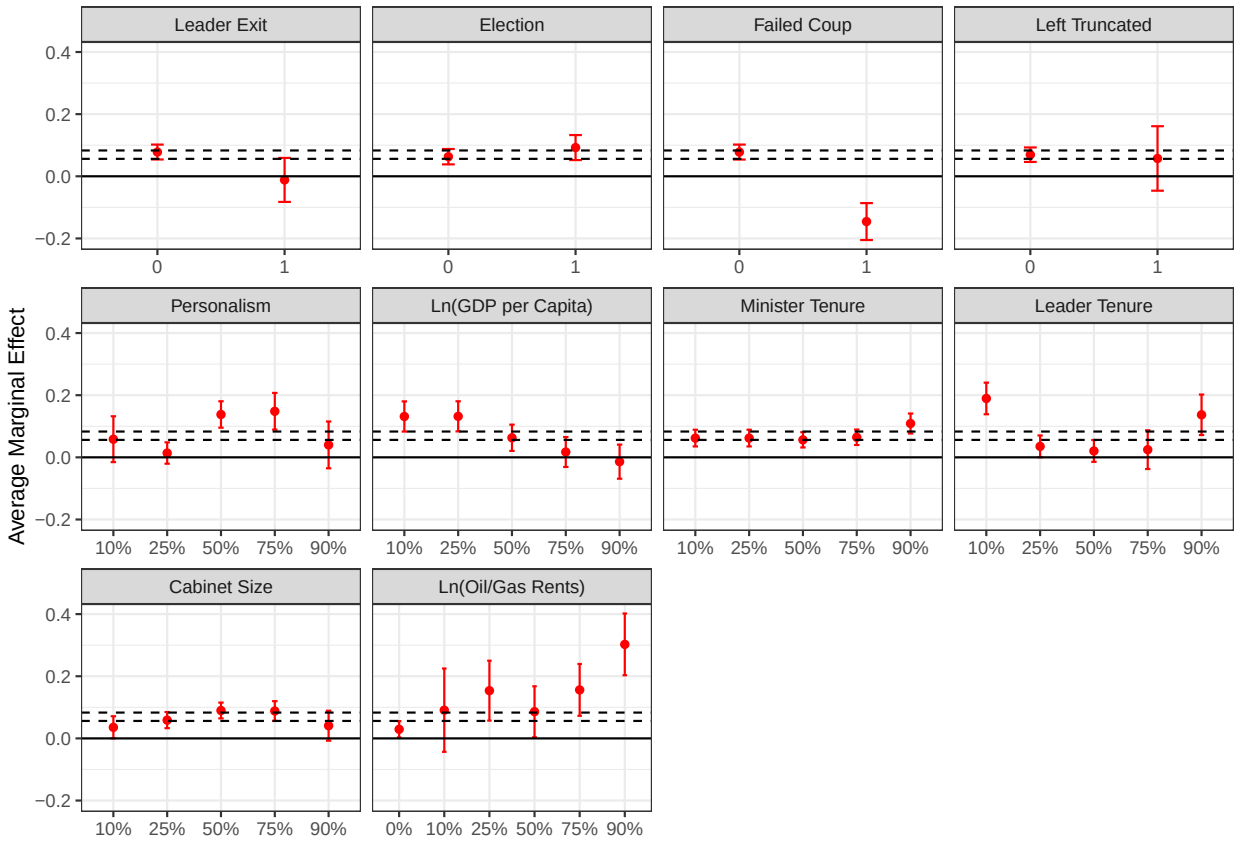
To show the heterogeneous effects, Figure 6 computes the average marginal effect for CWE for each variable in the original regression at a variety of levels; for binary variables, it computes the average marginal effect—across all observations—for both zero and one. For continuous variables, average marginal effects are shown at the 10%, 25%, 50%, 75%, and 90% quantiles of the variable. For log oil rents, 0% shows the effect when this variable is zero and the quantiles are of the non-zero values.

It reveals considerable heterogeneity versus the original results (average marginal effect of 0.07), and thus suggests the finding of non-robustness here may be attributable to un-modelled heterogeneity. Most dramatically, it seems that the impact of failed coups is considerably different in the CWE period; Woldense and Kroeger (2024, p. 186) summarize existing literature by noting that "scholars have shown that failed coup attempts often result in minister dismissals and reshuffles, particularly in personalist autocracies". However, Figure 6 shows that, in countries that have had a failed coup, being in the CWE period results in considerably *less* risk of a minister being dismissed. Put another way, failed coups increase risk of minister exit considerably outside the CWE period, as expected, but in the CWE period, there is a negative and statistically significant average marginal effect.

Other heterogeneous effects are also worth noting; for example, for leaders with short tenure (e.g., in the bottom 10%), there is increased probability of minister exit in the CWE period than outside of it, similarly in countries with very high log oil and gas rents.

We can again check the robustness of this new model by looking at the impact of removing an AMIS of $0.01N$ on the average marginal effect reported in Woldense and Kroeger (2024). We find that this appears to be non-robust (a significant sign flip occurs), but if we focus on the average

Figure 6: Heterogeneous Effects of "Cold War Ends"



marginal effect holding all covariates at *observed* values, this is robust in the refit model.

Use of Polynomial Terms: Building on prior literature, Woldense and Kroeger (2024) use third-order polynomials for "minister tenure" and create a predicted probability plot as a function of this variable in their manuscript. There is a question as to the robustness of this specification; using the Wald statistic on the predicted probability curve that ranges from 1 to 15—following Woldense and Kroeger (2024), but noting that the minister tenure has a farther range (to 29), Figure 7 shows the impact of refitting the model after removing the AMIS of particular α sizes, with the original curve shown in dashed lines.

We see that the curve becomes flatter as larger AMIS are removed and, for $\alpha = 0.01N$, it suggests that there is no effect of minister tenure. To examine whether the robustness can be improved by adjusting the model specification, Figure 8 estimates the curve using a model with splines for minister tenure and leader tenure.

Figure 7: Robustness of Predicted Probability Curve

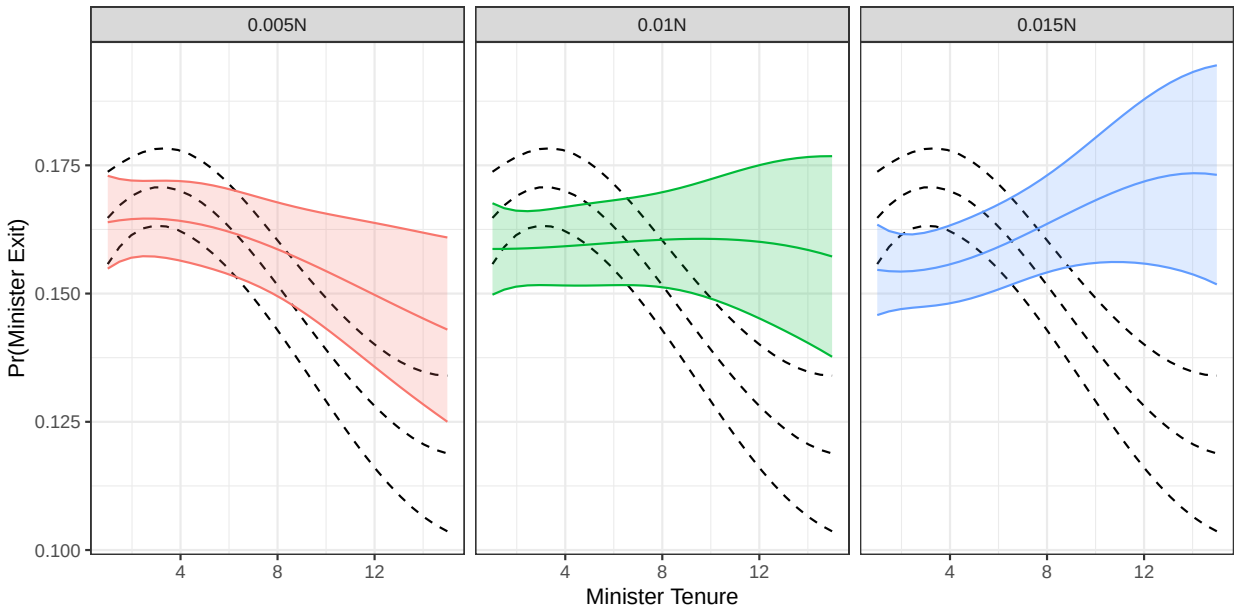
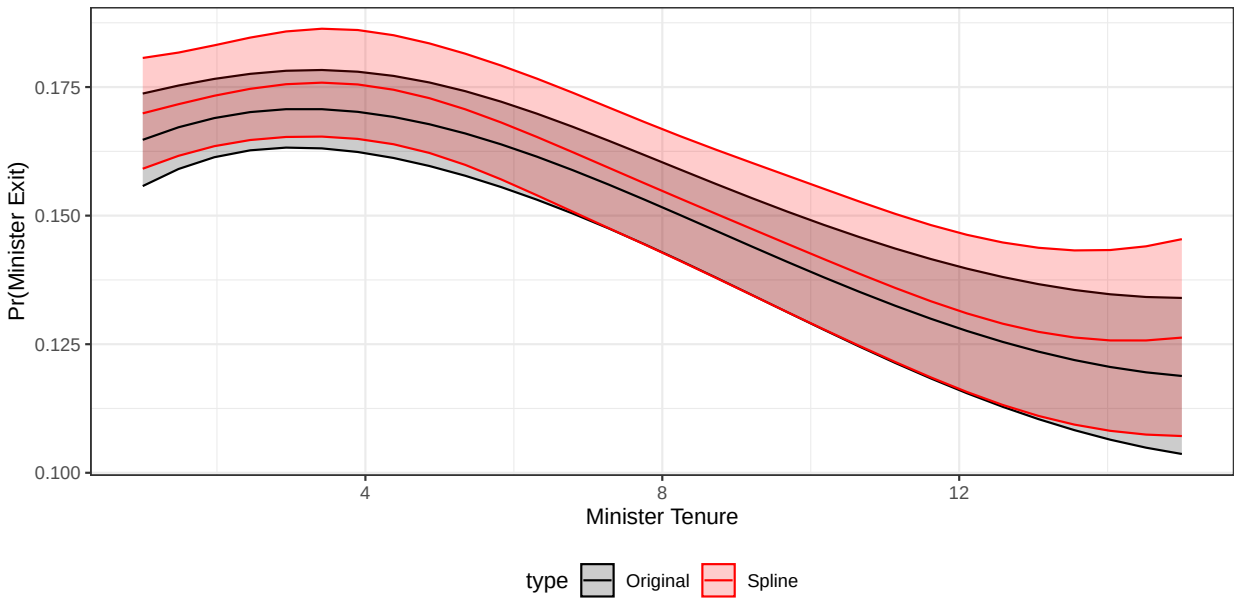


Figure 8: Robustness of Predicted Probability Curve (Woldense and Kroeger 2024)



Broadly speaking, it looks rather similar. Thus, the choice of cubic polynomial seems reasonably robust over the range considered in the original paper.

6 Conclusion

In this paper, we developed a methodology for efficiently identifying a set of observations that, if deleted, are likely to have a large impact on key quantities of interest. Our workflow is built around identifying this set for a small fraction of the data, removing this data, re-estimating the model and quantity of interest, and examining whether it changes. If it does, the result can be seen as "non-robust" as it is sensitive to small changes in the underlying data.

In a stylized example and empirical replications, we suggest that this non-robustness is often the product of model mis-specification or unmodelled heterogeneity. By addressing these using standard techniques (e.g., splines, interactions, etc.), the researcher can improve their initial model. In some cases, these improvements will be sufficient to make their original results robust to deleting small amounts of data, although non-robustness may still appear in even correctly specified models if the underlying true effect is comparatively small.

It would likely be useful in future work to develop a more automated method of figuring out which part of the model may be driving the non-robustness, although we have sketched out some tools for doing so (e.g., examining the identity of the observations in the approximately most influential set using either human observation or decision trees).

Further work might explore other ways to use this approximation to deleting a small set of observations; for example, one could use this method to extremely efficiently evaluate leave-one-out cross-validation as it only requires fitting the model once, obtaining the approximation of the leave-one-out coefficients, and generating a prediction on the test set.

References

Broderick, Tamara, Ryan Giordano and Rachael Meager. 2023. "An automatic finite-sample robustness metric: when can dropping a little data make a big difference?" *arXiv preprint arXiv:2011.14999* .

Cameron, A Colin, Jonah B Gelbach and Douglas L Miller. 2011. "Robust inference with multiway clustering." *Journal of Business & Economic Statistics* 29(2):238–249.

Chang, Qing and Max Goplerud. 2024. "Generalized Kernel Regularized Least Squares." *Political Analysis* 32(2):157–171.

Chatterjee, Samprit and Ali S Hadi. 1986. "Influential observations, high leverage points, and outliers in linear regression." *Statistical science* pp. 379–393.

Currie, Iain D, Maria Durban and Paul HC Eilers. 2006. "Generalized linear array models with applications to multidimensional smoothing." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 68(2):259–280.

Devaux, Martin and Naoki Egami. 2022. "Quantifying Robustness to External Validity Bias.". **URL:** *https://naokiegami.com/paper/external_robust.pdf*

Egami, Naoki and Erin Hartman. 2023. "Elements of external validity: Framework, design, and analysis." *American Political Science Review* 117(3):1070–1088.

Eubank, Nicholas and Adriane Fresh. 2022. "Enfranchisement and incarceration after the 1965 Voting Rights Act." *American Political Science Review* 116(3):791–806.

Findley, Michael G, Kyosuke Kikuta and Michael Denly. 2021. "External validity." *Annual Review of Political Science* 24(1):365–393.

Giger, Nathalie and Heike Klüver. 2016. "Voting against your constituents? How lobbying affects representation." *American Journal of Political Science* 60(1):190–205.

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.

Hazlett, Chad and Leonard Wainstein. 2022. "Understanding, Choosing, and Unifying Multilevel and Fixed Effect Approaches." *Political Analysis* 30(1):46–65.

King, Gary and Margaret E Roberts. 2015. "How robust standard errors expose methodological problems they do not fix, and what to do about it." *Political Analysis* 23(2):159–179.

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.

Künzel, Sören R, Jasjeet S Sekhon, Peter J Bickel and Bin Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects using Machine Learning." *Proceedings of the National Academy of Sciences* 116(10):4156–4165.

Kuschnig, Nikolas, Gregor Zens and Jesús Crespo Cuaresma. 2021. "Hidden in plain sight: Influential sets in linear models.".

Leeper, Thomas J. 2016. "Interpreting Regression Results using Average Marginal Effects with R's margins.".
URL: *https://s3.us-east-2.amazonaws.com/tjl-sharing/assets/AverageMarginalEffects.pdf*

Magnus, Jan R and Heinz Neudecker. 2019. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons.

Martinez, Luis R. 2022. "How much should we trust the dictator's GDP growth estimates?" *Journal of Political Economy* 130(10):2731–2769.

Ruppert, David, Matt P Wand and Raymond J Carroll. 2003. *Semiparametric Regression*. Cambridge University Press.

Williams, D. A. 1987. "Generalized linear model diagnostics using the deviance and single case deletions." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 36(2):181–191.

Woldense, Josef and Alex Kroeger. 2024. "Elite Change without Regime Change: Authoritarian Persistence in Africa and the End of the Cold War." *American Political Science Review* 118(1):178–194.

Wood, Simon N. 2011. "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(1):3–36.

Wood, Simon N. 2013. "On p-values for smooth components of an extended generalized additive model." *Biometrika* 100(1):221–228.

Wood, Simon N. 2017. *Generalized Additive Models*. Chapman and Hall/CRC.

Appendix for "Assessing Robustness of Post-Estimation Quantities of Interest"

A Notation and Matrix Operations

In the appendices, we repeatedly rely on certain matrix operations and corresponding notation that we define below:

- Hadamard or Element-wise Product: $\boldsymbol{a} \odot \boldsymbol{b}$ where $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^p$
- Kronekcer Product: $\boldsymbol{A} \otimes \boldsymbol{B}$ where $\boldsymbol{A} \in \mathbb{R}^{N_A \times p_A}$ and $\boldsymbol{B} \in \mathbb{R}^{N_B \times p_B}$
- Face-Splitting Product or Row-Wise Kronecker Product (see Currie, Durban and Eilers 2006): If $\boldsymbol{A} \in \mathbb{R}^{N \times p_A}$ and $\boldsymbol{B} \in \mathbb{R}^{N \times p_B}$, then,

$$Q(\boldsymbol{A}, \boldsymbol{B}) = \begin{bmatrix} \boldsymbol{a}_1^T \otimes \boldsymbol{b}_1^T \\ \dots \\ \boldsymbol{a}_N^T \otimes \boldsymbol{b}_N^T \end{bmatrix}$$

- Diagonal Matrix: If $\boldsymbol{d}$ is a vector in $\mathbb{R}^p$, then $\text{diag}(\boldsymbol{d})$ is a $p \times p$ diagonal matrix with $\boldsymbol{d}$ on the diagonal.
- Vectorization: If $\boldsymbol{A} \in \mathbb{R}^{N \times p}$, then $\text{vec}(\boldsymbol{A}) \in \mathbb{R}^{Np}$, i.e. a vector that stacks the columns of $\boldsymbol{A}$ vertically.
- Commutation Matrix: $\boldsymbol{K}_{N,p}$ is a permutation matrix such that for $\boldsymbol{A} \in \mathbb{R}^{N \times p}$, $\boldsymbol{K}_{n,p}\text{vec}(\boldsymbol{A}) = \text{vec}(\boldsymbol{A}^T)$.

B Derivation of Approximate Influence Scores

We first consider computing the approximate influence scores for a generalized linear model with likelihood shown below following notation from Wood (2017, ch. 3). We extend this notation to include a vector of observation-specific weights w_i, stacked into a vector as $\boldsymbol{w}$, where $\boldsymbol{w} = \boldsymbol{1}$ reflects the original dataset.

Assume that our data y_i comes from an exponential family distribution where g is the link function such that $g(\mu_i) = (\boldsymbol{x}_i^T \boldsymbol{\beta})$ where $\mu_i = E[y_i]$ and noting that $\mu = b'(\theta_i)$ where $b'(\cdot)$ denotes the derivative of $b(\cdot)$.

$$f_{\theta_i}(y_i) = \exp\left[\omega_i(y_i\theta_i)/\phi + c(y_i, \phi)\right]; \tag{1}$$

We focus on a simplified case where $\omega_i = 1$ for all i. The maximum likelihood estimates of $\boldsymbol{\beta}$ can be found maximizing the log-likelihood,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{i=1}^{N} w_i \left[y_i\theta_i - b_i(\theta_i)\right]/\phi + c_i(\phi, y_i). \tag{2}$$

An equivalent first-order condition is to find $\hat{\boldsymbol{\beta}}$ such that the derivative of the log-likelihood, the score, is equal to zero (Wood, 2017, p. 106):

$$s(\hat{\boldsymbol{\beta}}; \boldsymbol{w}) = \mathbf{0}; \quad s(\boldsymbol{\beta}; \boldsymbol{w}) = \frac{1}{\phi} \sum_{i=1}^{N} w_i \frac{y_i - \mu_i}{g'(\mu_i)V(\mu_i)} \boldsymbol{x}_i \tag{3}$$

One additional important quantity is the Hessian, i.e. the second derivative of the log-likelihood with respect to $\boldsymbol{\beta}$ (Wood, 2017, p. 106):

$$H(\boldsymbol{\beta}; \boldsymbol{w}) = -\frac{1}{\phi} \sum_{i=1}^{N} w_i \frac{\alpha(\mu_i)}{g'(\mu_i)^2 V(\mu_i)} \boldsymbol{x}_i \boldsymbol{x}_i^T; \quad \alpha(\mu_i) = 1 + (y_i - \mu_i)\left[V'(\mu_i)/V(\mu_i) + g''(\mu_i)/g'(\mu_i)\right] \tag{4}$$

From this, it is possible to use the implicit function theorem to find the derivative of $\hat{\boldsymbol{\beta}}$ and its covariance with respect to $\boldsymbol{w}$. Proposition B.1 states the result:

Proposition B.1. *Consider the generalized linear model discussed above where $\phi = 1$. Define w_i^{glm} as the weights used in the score, i.e. $w_i^{\text{glm}}(\boldsymbol{\beta}) = (y_i - \mu_i)/[g'(\mu_i)V(\mu_i)]$. Define $\boldsymbol{w}^{\text{glm}}(\boldsymbol{\beta})$ as the stacked vector of these weights. Define $w^{\text{IRLS}}(\boldsymbol{\beta})$ and the corresponding stacked $\boldsymbol{w}^{\text{IRLS}}(\boldsymbol{\beta})$ as the negative derivative of w_i^{glm} with respect to $\boldsymbol{\beta}$, i.e. $w_i^{\text{IRLS}}(\boldsymbol{\beta}) = \alpha(\mu_i)/[g'(\mu_i)^2 V(\mu_i)]$. Define $\boldsymbol{w}^{\text{dIRLS}}(\boldsymbol{\beta})$ as the stacked derivatives of w^{IRLS} with respect to $\boldsymbol{\beta}$; Wood (2011) provides the explicit form. Define*

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}; \boldsymbol{w}) = \left[\boldsymbol{X}^T \mathrm{diag}\left(\boldsymbol{w} \odot \boldsymbol{w}^{\mathrm{IRLS}}(\hat{\boldsymbol{\beta}})\right) \boldsymbol{X}\right]^{-1}.$$

Then, the derivatives of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ and corresponding estimator of the covariance can with respect to $\boldsymbol{w}$ are,

$$\begin{aligned}
\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{w}} &= -\left[H\left(\hat{\boldsymbol{\beta}}; \boldsymbol{w}\right)\right]^{-1} \boldsymbol{X}^T \mathrm{diag}\left(\boldsymbol{w}^{\mathrm{glm}}(\hat{\boldsymbol{\beta}})\right) \\
\frac{\partial \mathrm{vec}\left(\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}; \boldsymbol{w})\right)}{\partial \boldsymbol{w}} &= -Q\left(\boldsymbol{X}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}), \boldsymbol{X}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\right)^T \left[\mathrm{diag}\left(\boldsymbol{w}^{\mathrm{IRLS}}(\hat{\boldsymbol{\beta}})\right) + \mathrm{diag}\left(\boldsymbol{w} \odot \boldsymbol{w}^{\mathrm{dIRLS}}(\hat{\boldsymbol{\beta}})\right) \boldsymbol{X} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{w}}\right]
\end{aligned}$$

Proof. The derivative of $\hat{\boldsymbol{\beta}}$ with respect to $\boldsymbol{w}$ is a specific case of Equation 10 in Broderick, Giordano and Meager (2023), or a direct application of the implicit function theorem to Equation 3. The derivative of the estimated covariance can be established by noting some results on matrix differentials and vectorization; specifically, $d\boldsymbol{A}^{-1} = -\boldsymbol{A}^{-1} d\boldsymbol{A} \boldsymbol{A}^{-1}$ (Magnus and Neudecker, 2019, p. 168), $\mathrm{vec}(\boldsymbol{X}^T \mathrm{diag}(\boldsymbol{d}) \boldsymbol{X}) = Q(\boldsymbol{X}, \boldsymbol{X})^T \boldsymbol{d}$ (Currie, Durban and Eilers, 2006), and $\boldsymbol{a} \odot \boldsymbol{b} = \mathrm{diag}(\boldsymbol{a})\boldsymbol{b}$. By vectorizing $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ and carefully computing the chain rule over the differentials, the result is established. □

If the expected information is used to compute the estimate of the variance, i.e. $\{1/[g'(\mu_i)^2 V(\mu_i)]\}_{i=1}^N$, then the above formula substitutes this and its derivative for $\boldsymbol{w}^{\mathrm{IRLS}}$ and $\boldsymbol{w}^{\mathrm{dIRLS}}$. The default, at the time of writing in `R`'s `glm` and `mgcv` is to use the expected weights. Note, however, that $H(\boldsymbol{\beta}; \boldsymbol{w})$ must be evaluated in terms of the (Newton) weights when finding $\partial \hat{\boldsymbol{\beta}} / \partial \boldsymbol{w}$.

Improving Computational Cost While Proposition B.1 shows how to the derivative of the estimated covariance with respect to $\boldsymbol{w}$, it is highly impractical to use. Note that computing this matrix is a $p^2 \times N$ matrix—possibly quite large—and computing it has a complexity of $O(Np^3)$ or $O(N^2 p)$ depending on the order of matrix multiplications. This is likely prohibitively expensive in large problems.

Fortunately, most relevant quantities involve derivatives of quadratic forms. For example, if the derivative of a quadratic form, say, $\boldsymbol{a}^T \widehat{\mathrm{Vec}}(\hat{\boldsymbol{\beta}}; \boldsymbol{w}) \boldsymbol{b}$ were desired, this could be naively computed by

$$[\boldsymbol{b} \otimes \boldsymbol{a}]^T \frac{\partial \text{vec}\left(\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}; \boldsymbol{w})\right)}{\partial \boldsymbol{w}} \tag{5}$$

However, in addition to the unfavorable costs to compute the matrix of derivatives, evaluating this product has a complexity of $O(Np^2)$. Proposition B.2 shows that the quadratic form of interest can be evaluated without ever forming a $p^2 \times N$ matrix and with a complexity that scales in $O(Np)$.

Proposition B.2. *Define* $\tilde{\boldsymbol{a}} = \widehat{\text{Var}}(\hat{\boldsymbol{\beta}}; \boldsymbol{w})\boldsymbol{a}$ *and* $\tilde{\boldsymbol{b}} = \widehat{\text{Var}}(\hat{\boldsymbol{\beta}}; \boldsymbol{w})\boldsymbol{b}$*. Then, the following equivalence holds and the right-hand side can be evaluated in a computational complexity that is* $O(Np)$*.*

$$[\boldsymbol{b} \otimes \boldsymbol{a}]^T \frac{\partial \text{vec}\left(\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}; \boldsymbol{w})\right)}{\partial \boldsymbol{w}} = Q\left(\boldsymbol{X}\tilde{\boldsymbol{a}}, \boldsymbol{X}\tilde{\boldsymbol{b}}\right)^T \left[\boldsymbol{w}^{\text{IRLS}}(\hat{\boldsymbol{\beta}}) + \text{diag}\left(\boldsymbol{w} \odot \boldsymbol{w}^{\text{dIRLS}}(\hat{\boldsymbol{\beta}})\right) \boldsymbol{X} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{w}}\right] \tag{6}$$

Proof. The proof of the equivalence is established by noting that for conformable matrices, and noting the definitions in Appendix A, $Q(\boldsymbol{X}, \boldsymbol{Z})\,[\boldsymbol{A} \otimes \boldsymbol{B}] = Q(\boldsymbol{XA}, \boldsymbol{ZB})$. The proof of the complexity can be established as follows: $\boldsymbol{X}\tilde{\boldsymbol{a}}$ are $\boldsymbol{X}\tilde{\boldsymbol{b}}$ are each $O(Np)$ to compute. Computing $Q(\boldsymbol{X}\tilde{\boldsymbol{a}}, \boldsymbol{X}\tilde{b})$ is $O(N)$, i.e. the Hadamard product of the two vectors. The multiplication against $\text{diag}(\boldsymbol{w}^{\text{IRLS}}(\hat{\boldsymbol{\beta}}))$ is $O(N)$. Careful multiplication for the second term reveals that $Q(\boldsymbol{X}\tilde{\boldsymbol{a}}, \boldsymbol{X}\tilde{b})^T \text{diag}(\boldsymbol{w} \odot \boldsymbol{w}^{\text{dIRLS}}(\hat{\boldsymbol{\beta}}))\boldsymbol{X}$ is $O(N + Np)$ and results in a $1 \times p$ matrix. The final multiplication against $\partial\hat{\boldsymbol{\beta}}/\partial\boldsymbol{w}$ is thus $O(Np)$. Thus, collecting terms gives a result that is dominated by $O(Np)$. □

This is useful in establishing that as long as one is only interested in quadratic forms of the variance, it is never necessary to form the $p \times N^2$ matrix and such quadratic forms can be computed in $O(Np)$. A result of this is that the derivative of the diagonal of $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$ with respect to $\boldsymbol{w}$ can be found at $O(Np^2)$ and does not require storing or forming an $p^2 \times N$ matrix. Thus, the derivative of the variance of $\hat{\boldsymbol{\beta}}$ should almost *never* be formed explicitly as almost all quantities of interest (e.g., the Wald tests discuss below) can be expressed as quadratic forms. If $\boldsymbol{Xv}$ can be evaluated at a cost lower than $O(Np)$, i.e. because $\boldsymbol{X}$ is mostly sparse, additional improvements arise.

Non-Fixed ϕ If ϕ is not fixed, as in the case of linear regression, it must also be estimated. Note that $\hat{\boldsymbol{\beta}}$ does not depend on $\hat{\phi}$ and it only affects $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$. Different estimators of $\hat{\phi}$ exist; in the Gaussian setting, the defaults in `R` and `mgcv` coincide as

$$\hat{\phi} = \frac{\sum_{i=1}^{N} w_i (y_i - \hat{\mu}_i)^2 / V(\hat{\mu}_i)}{N - p}; \quad g(\hat{\mu}_i) = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}} \tag{7}$$

To consider the impact on the derivative of the estimated covariance matrix, we will use—only when ϕ is not fixed at 1—an expanded notation—$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}; \phi, \boldsymbol{w})$ to denote the dependence on ϕ. If $\phi = 1$, this coincides with the quantity analyzed in Proposition B.1. Then, if $\hat{\phi}$ is estimated, the derivative can be expressed as follows:

$$\frac{\partial \text{vec}\left(\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}; \hat{\phi}, \boldsymbol{w})\right)}{\partial \boldsymbol{w}} = \text{vec}\left(\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}; 1, \boldsymbol{w})\right) \frac{\partial \hat{\phi}}{\partial \boldsymbol{w}} + \hat{\phi} \cdot \frac{\partial \text{vec}\left(\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}; 1, \boldsymbol{w})\right)}{\partial \boldsymbol{w}}. \tag{8}$$

This follows by noting that $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}; \hat{\phi}, \boldsymbol{w}) = \hat{\phi}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}; 1, \boldsymbol{w})$ and thus estimating ϕ uses the (scaled) derivative from Proposition B.1 plus an additional term to reflect the fact that $\hat{\phi}$ depends on $\boldsymbol{w}$. The derivative of $\hat{\phi}$ with respect to $\boldsymbol{w}$ is shown below, noting that $N = \sum_i w_i$ where $g'(\hat{\boldsymbol{\mu}})$ applies $g'(\mu)$ elementwise to $\hat{\mu}_i$. Wood (2011) provides the relevant derivative of the $\hat{\phi}$ with respect to $\hat{\boldsymbol{\beta}}$.

$$\frac{\partial \hat{\phi}}{d\boldsymbol{w}} = \frac{1}{N-p}\left[\left(\frac{\boldsymbol{y} - \hat{\boldsymbol{\mu}}}{V(\hat{\boldsymbol{\mu}})}\right)^T + \boldsymbol{w}^T \text{diag}\left(-\frac{1}{g'(\hat{\boldsymbol{\mu}})}\left[\frac{2(\boldsymbol{y} - \hat{\boldsymbol{\mu}})}{V(\hat{\boldsymbol{\mu}})} + \frac{(\boldsymbol{y} - \hat{\boldsymbol{\mu}})^2}{V(\hat{\boldsymbol{\mu}})}\frac{V'(\hat{\boldsymbol{\mu}})}{V(\hat{\boldsymbol{\mu}})}\right]\right)\boldsymbol{X}\frac{\partial \hat{\boldsymbol{\beta}}}{d\boldsymbol{w}}\right] - \frac{\hat{\phi}}{N-p}\mathbf{1}^T \tag{9}$$

This presents no additional complication regarding the computational scalability in terms of derivatives of quadratic forms of the variance as one simply computes $\partial\hat{\phi}/\partial\boldsymbol{w}$, and then applies Proposition B.2 to the second term in Equation 8 and computes $[\boldsymbol{b} \otimes \boldsymbol{a}]^T \text{vec}\left(\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}; \hat{\phi}, \boldsymbol{w})\right)$ at a cost of $O(p^2)$.

Note that Equation 9 differs from the software implementation in Broderick, Giordano and Meager (2023) as they do not include the final term $-1/(N-p)\hat{\phi}\mathbf{1}$ that arises because of $N = \sum_i w_i$. For large N, this term vanishes so the difference is rather small in practice.

C Clustered Standard Errors

Consider the common type of clustered standard error that can be expressed as follows in a generalized linear model or generalized additive model (Hazlett and Wainstein 2022; Chang and Goplerud

2024), where $\boldsymbol{G} \in \{0,1\}^{N\times N}$ denotes a (sparse) adjacency matrix where $\boldsymbol{G}_{ij} = 1$ if i and j are in the same group but its *diagonal* is zero,

$$\widehat{\mathrm{Var}_{\boldsymbol{G}}}(\hat{\boldsymbol{\beta}}) = \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) \left[\sum_i \sum_j \boldsymbol{G}_{ij}\boldsymbol{x}_i\boldsymbol{x}_j^T w_i w_j w_i^{\mathrm{IRLS}} w_j^{\mathrm{IRLS}} + \sum_i \boldsymbol{x}_i\boldsymbol{x}_i^T \left(w_i^{\mathrm{IRLS}}\right)^2 w_i\right] \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) \tag{10}$$

$$= \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) \left[\boldsymbol{X}^T \mathrm{diag}\left(\boldsymbol{w}^{\mathrm{IRLS}} \odot \boldsymbol{w}\right) \boldsymbol{G}\, \mathrm{diag}\left(\boldsymbol{w}^{\mathrm{IRLS}} \odot \boldsymbol{w}\right) \boldsymbol{X} + \boldsymbol{X}^T \mathrm{diag}\left[\left(\boldsymbol{w}^{\mathrm{IRLS}}\right)^2 \odot \boldsymbol{w}\right] \boldsymbol{X}\right] \widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) \tag{11}$$

Proposition C.1 provides the derivative of this quantity with respect to $\boldsymbol{w}$ as well as a result on the computational complexity.

Proposition C.1. *Given a matrix $\boldsymbol{G}$ and a cluster robust covariance matrix $\widehat{\mathrm{Var}_{\boldsymbol{G}}}(\hat{\boldsymbol{\beta}})$, define*

$$\boldsymbol{M} = \boldsymbol{X}^T \mathrm{diag}\left(\boldsymbol{w}^{\mathrm{IRLS}} \odot \boldsymbol{w}\right) \boldsymbol{G}\, \mathrm{diag}\left(\boldsymbol{w}^{\mathrm{IRLS}} \odot \boldsymbol{w}\right) \boldsymbol{X} + \boldsymbol{X}^T \mathrm{diag}\left[\left(\boldsymbol{w}^{\mathrm{IRLS}}\right)^2 \odot \boldsymbol{w}\right] \boldsymbol{X}.$$

Then, the derivative with respect to $\boldsymbol{w}$ is

$$\frac{\partial \mathrm{vec}\left(\widehat{\mathrm{Var}_{\boldsymbol{G}}}(\hat{\boldsymbol{\beta}})\right)}{\partial \boldsymbol{w}} = \left[\boldsymbol{I} \otimes \boldsymbol{M}\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}})} + \boldsymbol{M}\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}})} \otimes \boldsymbol{I}\right] \frac{\partial \widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}})}}{\partial \boldsymbol{w}} + \left[\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}})} \otimes \widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}})}\right] \frac{\partial \mathrm{vec}(\boldsymbol{M})}{\partial \boldsymbol{w}}$$

$$\frac{\partial \mathrm{vec}(\boldsymbol{M})}{\partial \boldsymbol{w}} = \begin{array}{l} [\boldsymbol{I} + \boldsymbol{K}]\, Q\left(\boldsymbol{G}\mathrm{diag}\left(\boldsymbol{w}^{\mathrm{IRLS}} \odot \boldsymbol{w}\right) \boldsymbol{X}, \boldsymbol{X}\right)^T \left[\mathrm{diag}\left(\boldsymbol{w}^{\mathrm{IRLS}}\right) + \mathrm{diag}\left(\boldsymbol{w} \odot \boldsymbol{w}^{\mathrm{dIRLS}}\right) \boldsymbol{X} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{w}}\right] + \\ Q(\boldsymbol{X}, \boldsymbol{X})^T \left[\mathrm{diag}\left(\boldsymbol{w}^{\mathrm{IRLS}}\right) + 2\ \mathrm{diag}\left(\boldsymbol{w} \odot \boldsymbol{w}^{\mathrm{IRLS}} \odot \boldsymbol{w}^{\mathrm{dIRLS}}\right) \boldsymbol{X} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{w}}\right] \end{array}$$

The derivative of quadratic forms against $\widehat{\mathrm{Var}_{\boldsymbol{G}}}(\hat{\boldsymbol{\beta}})$ can be evaluated at a cost that is $\mathcal{O}([N + \mathrm{nnz}(\boldsymbol{G})]p)$ where $\mathrm{nnz}(\boldsymbol{G})$ is the number of non-zero elements in $\boldsymbol{G}$.

Proof. The derivative follows from careful application of the chain rule and earlier results. The result on computational complexity of the derivative of the quadratic form follows by noting that $\boldsymbol{G}\mathrm{diag}\left(\boldsymbol{w}^{\mathrm{IRLS}} \odot \boldsymbol{w}\right) \boldsymbol{X}$ can be evaluated at a cost dominated by $\mathcal{O}(Np + \mathrm{nnz}(\boldsymbol{G})p)$. Once this has been computed, the cost is $\mathcal{O}(Np)$ per Proposition B.2. Thus, the cost is dominated by

$\mathcal{O}\left([N + \text{nnz}(\boldsymbol{G})]\, p\right)$. □

This simplifies considerably in the heteroskedastic case, i.e. where $\boldsymbol{G} = \boldsymbol{0}$. In that case, only the second term in $\partial \text{vec}(\boldsymbol{M})/\partial \boldsymbol{w}$ is required. In the clustered setting, if there are C clusters of size C_g, then the number of non-zero elements in $\mathcal{G}$ is $\sum_g C_g^2 - N$. We can consider how this behaves in the case of equally sized clusters: In the case of equally size clusters, $\text{nnz}(\boldsymbol{G}) = N^2/C - N = N\,[C_g - 1]$. If we assume that C_g is fixed as N grows asymptotically (e.g., in the case of panel data that adds new individuals), this can still be computed efficiently for large N. In the case of very large clusters, $\text{nnz}(\boldsymbol{G}) \approx N^2$ and even this "efficient" evaluation of the derivative for a quadratic form can be dominated by $\mathcal{O}(N^2 p)$.

Multiway Clustering: Cameron, Gelbach and Miller (2011) propose multiway clustering; this has been applied in political science for dyadic data. It can be expressed by writing $\boldsymbol{G}$ as an additive combination of sparse matrices; thus, the results above can be immediately applied.

D Generalized Additive Models

We consider a generalized additive model following notation in Wood (2017). Specifically, we now maximize a penalized log-likelihood with a J-length vector of non-negative parameters $\boldsymbol{\lambda}$ that correspond to the amount of regularization or smoothing. To simplify notation, we collect all parameters into $\boldsymbol{\beta}$ and appropriately adjust $\boldsymbol{x}_i$. The maximum likelihood estimate as a function of $\boldsymbol{w}$ and $\boldsymbol{\lambda}$ is shown below:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}(\boldsymbol{w}) = \underset{\boldsymbol{\beta}}{\text{argmax}} \sum_{i=1}^{N} w_i \left[y_i \theta_i - b_i(\theta_i)\right]/\phi + c_i(\phi, y_i) - \frac{1}{2\phi} \sum_{j=1}^{J} \lambda_j \boldsymbol{\beta}^T \boldsymbol{S}_j \boldsymbol{\beta} \tag{12}$$

This formulation can express random effects, splines, and many other common penalized terms; see Wood (2017) for a review. The amount of regularization is calibrated by optimizing some function g that depends on $\boldsymbol{\lambda}$ and, usually, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$.

Define $\ell(\boldsymbol{\beta})$ as the log-likelihood, $\boldsymbol{S}_{\boldsymbol{\lambda}} = \sum_{j=1}^{J} \lambda_j \boldsymbol{S}_j$, $|\boldsymbol{A}|_+$ as the generalized determinant of $\boldsymbol{A}$, and M as the dimension of the nullspace of $\boldsymbol{S}_{\boldsymbol{\lambda}}$. Further define τ as the effective degrees of freedom of the model, typically using

$$\tau = \left[\boldsymbol{X}^T \text{diag}(\boldsymbol{w}^{\text{IRLS}}(\hat{\boldsymbol{\beta}}))\boldsymbol{X} + \boldsymbol{S}_{\boldsymbol{\lambda}}\right]^{-1} \left[\boldsymbol{X}^T \text{diag}(\boldsymbol{w}^{\text{IRLS}}(\hat{\boldsymbol{\beta}}))\boldsymbol{X}\right], \tag{13}$$

although note that `mgcv` uses the expected weights (i.e. Fisher weights; Wood 2017, p. 283). Then, the two most popular types of criterion for tuning $\boldsymbol{\lambda}$ are shown below:

- Generalized Cross-Validation (GCV): Define $D(\boldsymbol{\beta})$ as the deviance, i.e.

$$D(\boldsymbol{\beta}) = 2\phi \left[\sum_{i=1}^{N} w_i \left[\ell_i(\tilde{\boldsymbol{\beta}}_i) - \ell_i(\boldsymbol{\beta})\right]\right] = \sum_{i=1}^{N} w_i \left[y_i \left(\tilde{\theta}_i - \theta_i\right) - b(\tilde{\theta}_i) + b(\theta_i)\right],$$

  where $\ell_i(\boldsymbol{\beta})$ represents the log-likelihood for observation i, and $\tilde{\boldsymbol{\beta}}_i$ and $\tilde{\theta}_i$ represents the maximum likelihood estimates of saturated model, i.e. the model that best fits the observed data. The GCV criterion is shown below (Wood, 2017, p. 261)

  - If ϕ is known,

$$g(\boldsymbol{\lambda}) = D(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) + 2\tau\phi$$

  - If ϕ is estimated,

$$g(\boldsymbol{\lambda}) = \frac{N \cdot D(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})}{(N-\tau)^2}$$

- Restricted Maximum Likelihood (REML): The REML criterion is shown below (Wood, 2017, p. 263):[6]

$$g(\boldsymbol{\lambda}) = \ell(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) + \sum_{j=1}^{J} \left[-\frac{\lambda_j \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}^T \boldsymbol{S}_j \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}}{2\phi}\right] + \frac{1}{2}\ln\left|\frac{\boldsymbol{S}_{\boldsymbol{\lambda}}}{\phi}\right|_+ - \frac{1}{2}\ln\left|\frac{\boldsymbol{X}^T \text{diag}\left(\boldsymbol{w}^{\text{IRLS}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})\right)\boldsymbol{X} + \boldsymbol{S}_{\boldsymbol{\lambda}}}{\phi}\right| + \frac{M}{2}\log(2\pi)$$

For either criterion, the optimal $\boldsymbol{\lambda}$ is defined by maximizing $g(\boldsymbol{\lambda})$; in practice, the relevant software `mgcv` optimizes over the log of $\boldsymbol{\lambda}$ to make the scale unrestricted, i.e.

$$\hat{\boldsymbol{\rho}} = \underset{\boldsymbol{\rho}}{\text{argmax}}\, g\left(\exp(\boldsymbol{\rho})\right); \quad \hat{\boldsymbol{\lambda}} = \exp(\hat{\boldsymbol{\rho}}). \tag{14}$$

Proposition D.1 shows how the derivative of $\hat{\boldsymbol{\rho}}$ can be found with respect to the observation weights $\boldsymbol{w}$.

[6] Note that this follows the corrected version in the errata of that book.

Proposition D.1. *Assume $\phi = 1$. Then, the derivative of $\hat{\boldsymbol{\rho}}$ with respect to $\boldsymbol{w}$ for a given $g(\boldsymbol{\lambda})$ can be found as follows*

$$\frac{\partial \hat{\boldsymbol{\rho}}}{\partial \boldsymbol{w}} = \left[\frac{\partial^2 g(\exp(\boldsymbol{\rho}))}{\partial \boldsymbol{\rho}\boldsymbol{\rho}^T}\right]^{-1}_{\boldsymbol{\rho}=\hat{\boldsymbol{\rho}}} \left[\frac{\partial g(\exp(\boldsymbol{\rho}))}{\partial \boldsymbol{\rho}\boldsymbol{w}^T}\right]_{\boldsymbol{\rho}=\hat{\boldsymbol{\rho}}}; \quad \frac{\partial \hat{\boldsymbol{\lambda}}}{\partial \boldsymbol{w}} = \operatorname{diag}(\hat{\boldsymbol{\rho}}) \frac{\partial \hat{\boldsymbol{\rho}}}{\partial \boldsymbol{w}}.$$

Proof. The proof follows by the implicit function theorem for $\hat{\boldsymbol{\rho}}$ and the chain rule for $\hat{\boldsymbol{\lambda}}$. □

However, $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\lambda}}$ are rarely of direct interest. Rather, one uses $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}}$ for inference with the following estimate of the covariance:

$$\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}}) = \phi \left[\boldsymbol{X}^T \operatorname{diag}\left(\boldsymbol{w} \odot \boldsymbol{w}^{\text{IRLS}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}})\right) \boldsymbol{X} + \boldsymbol{S}_{\hat{\boldsymbol{\lambda}}}\right]^{-1} \tag{15}$$

We start by considering the case of $\phi = 1$, Proposition D.2 shows how the of the derivatives of $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}}$ and $\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}})$ can be computed.

Proposition D.2. *In a generalized additive model defined above with smoothing parameters $\boldsymbol{\lambda}$ and $\phi = 1$ where an estimate $\hat{\boldsymbol{\lambda}}$ is found by minimizing some function $g(\boldsymbol{\lambda})$, the derivative of $\hat{\boldsymbol{\beta}}$ and the estimate of its variance given $\hat{\boldsymbol{\lambda}}$ with respect to $\boldsymbol{w}$ are shown below*

$$\frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}}}{\partial \boldsymbol{w}} = \left[\frac{\partial \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}}{\partial \boldsymbol{w}}\right]_{\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}} - \sum_{j=1}^{J} \hat{\lambda}_j \widehat{\operatorname{Var}}\left(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}}\right) \boldsymbol{S}_j \hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}} \frac{\partial \hat{\boldsymbol{\rho}}_j}{\partial \boldsymbol{w}} \tag{16}$$

$$\begin{aligned}\frac{\partial \operatorname{vec}\left(\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}})\right)}{\partial \boldsymbol{w}} = & - Q\left(\boldsymbol{X}\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}}), \boldsymbol{X}\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}})\right)^T \left[\operatorname{diag}\left(\boldsymbol{w}^{\text{IRLS}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}})\right) + \operatorname{diag}\left(\boldsymbol{w} \odot \boldsymbol{w}^{\text{dIRLS}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}})\right) \boldsymbol{X} \frac{\partial \hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}}}{\partial \boldsymbol{w}}\right] + \\ & - \sum_{j=1}^{J} \left[\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}}) \otimes \widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}})\right] \operatorname{vec}(\boldsymbol{S}_j) \frac{\partial \hat{\rho}_j}{\partial \boldsymbol{w}}.\end{aligned} \tag{17}$$

Further, assuming that $\partial\hat{\lambda}/\partial\boldsymbol{w}$ has already been obtained, the computational complexity of evaluating the derivative of a quadratic form of the $\boldsymbol{a}^T\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}})\boldsymbol{b}$ with respect to $\boldsymbol{w}$ is $\mathcal{O}(...)$.

Proof. The derivatives follow by the chain rule and $d\boldsymbol{A}^{-1} = -\boldsymbol{A}^{-1}d\boldsymbol{A}\boldsymbol{A}^{-1}$.

On computational complexity of derivatives of quadratic forms, the first term in the derivative of $\partial\operatorname{vec}(\widehat{\operatorname{Var}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})/\partial\boldsymbol{w}$ can be tacked identically as in Proposition B.2 and is thus $\mathcal{O}(Np)$. The second term can be simplified using the mixed product property of Kronecker products to

$\left[\boldsymbol{a}^T\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}}) \otimes \boldsymbol{b}^T\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}})\right] \text{vec}(\boldsymbol{S}_j)\partial\hat{\lambda}_j/\partial\boldsymbol{w}$ and thus can be evaluated in $O(p^2J)$. The cost of evaluating the derivative of the quadratic form is thus dominated by $\mathcal{O}(Np + p^2J)$. □

Thus, this proposition established that the generalized additive model setting is quite similar to the standard generalized linear model. However, due to the estimation of $\hat{\boldsymbol{\lambda}}$, there are two channels through which changing $\boldsymbol{w}$ affects the estimates and the estimated covariance matrix. First, *fixing* $\boldsymbol{\lambda}$, there is a direct impact on $\hat{\boldsymbol{\beta}}$ that exactly mirrors the results in Proposition B.1; however, there is now a second channel through which adjusting $\boldsymbol{w}$ affects $\hat{\boldsymbol{\lambda}}$ and therefore also affects $\hat{\boldsymbol{\beta}}$.

If $\hat{\phi}$ is estimated, Proposition D.2 can be adapted analogously to how it is done in the unregularized case. The estimation of $\hat{\phi}$ in the Gaussian case is the estimator in Equation 7, above, but dividing by $N - \tau$ instead of $N - p$.

Clustered Standard Errors : If clustered standard errors are desired alongside the generalized additive model (see Hazlett and Wainstein 2022; Chang and Goplerud 2024), the results from Appendix D can be used here with the same meat ($\boldsymbol{M}$) and the variance matrix $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\lambda}}})$.

Non-Fixed ϕ : If ϕ is not fixed, it must be estimated. This could be done by optimizing $g(\exp(\rho), \phi)$ over ρ and ϕ or by using a plug-in estimator of ϕ (see Equation 7), using $\hat{\phi}(\rho)$ to denote this as a function of ρ, and then optimizing over $g(\exp(\rho), \hat{\phi}(\rho))$. By construction, note that $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}(\boldsymbol{w})$ does not depend on ϕ.

E Derivation of Post-Regression Quantities of Interest

This appendix provides the derivatives of the post-regression quantities of interest discussed in Section 4.

E.1 Wald Statistic

To begin, we consider the Wald statistic $\widehat{W}$, we first consider the standard case where $\boldsymbol{A} \in \mathbb{R}^{M_0 \times p}$ and $\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{A}^T$ is full rank and then provide a generalization to the non-full rank setting.

$$\widehat{W} = \left[\boldsymbol{A}\hat{\boldsymbol{\beta}}\right]^T \left[\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{A}^T\right]^{-1} \left[\boldsymbol{A}\hat{\boldsymbol{\beta}}\right] \tag{18}$$

Proposition E.1 provides the derivative with respect to $\boldsymbol{w}$ in the full-rank setting.

Proposition E.1. *Define* $\widehat{\boldsymbol{V}}_{\boldsymbol{A}} = \left[\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{A}^T\right]^{-1}$. *Then, the derivative of the Wald statistic* $\widehat{W}$ *is*

$$\frac{\partial \widehat{W}}{\partial \boldsymbol{w}} = -\left[\boldsymbol{A}^T\widehat{\boldsymbol{V}}_{\boldsymbol{A}}\boldsymbol{A}\hat{\boldsymbol{\beta}} \otimes \boldsymbol{A}^T\widehat{\boldsymbol{V}}_{\boldsymbol{A}}\boldsymbol{A}\hat{\boldsymbol{\beta}}\right]^T \frac{\partial \widehat{\text{Var}}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{w}} + 2\left[\boldsymbol{A}\hat{\boldsymbol{\beta}}\right]^T \widehat{\boldsymbol{V}}_{\boldsymbol{A}}\boldsymbol{A}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{w}}.$$

This derivative can be evaluated at $\mathcal{O}([M_0]^3 + M_0p^2 + Np)$.

Proof. The proof of the derivative follows from the standard applications of chain rules and vectorizations. The computational complexity of this can be established by noting that $\widehat{\boldsymbol{V}}_{\boldsymbol{A}}$ can be evaluated at $\mathcal{O}(M_0p^2 + [M_0]^3)$. The first term can be evaluated at $\mathcal{O}(M_0p^2 + Np)$ using Proposition B.2. The second term can be evaluated at $\mathcal{O}(M_0p + M_0^2 + Np)$. Collecting terms yields a complexity of $\mathcal{O}([M_0]^3 + M_0p^2 + Np)$. □

However, it is common to have settings where $\widehat{\boldsymbol{V}}_{\boldsymbol{A}}$ is not full rank. We follow Wood (2013) in proposing a Wald statistic in that setting. To do so, Lemma E.1 defines a rank r inverse of a real symmetric matrix $\boldsymbol{A}$ and note its differential. We use $\boldsymbol{A}^-$ to indicate the standard generalized inverse.

Lemma E.1. *For real and symmetric matrix* $\boldsymbol{A}$ *with eigendecomposition* $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$, *for positive integer* r, *define* $\boldsymbol{A}^{r-}$ *as the rank* r *generalized inverse of* $\boldsymbol{A}$, *i.e.*

$$\boldsymbol{A}^{r-} = \sum_{k=1}^{r} \frac{1}{\lambda_k}\boldsymbol{q}_k\boldsymbol{q}_k^T$$

Then, following the conditions of (Magnus and Neudecker, 2019, p. 180) on the differentials of an eigendecomposition (e.g., all $\{\lambda_1, \cdots, \lambda_r\}$ *are simple eigenvalues), then the differential of* $d\text{vec}\left(\boldsymbol{A}^{r-}\right)$ *is shown below:*

$$d\text{vec}(\boldsymbol{A}^{r-}) = \left(\sum_{k=1}^{r} -\frac{1}{\lambda_k^2}\left[\boldsymbol{q}_i \otimes \boldsymbol{q}_i\right]\left[\boldsymbol{q}_i \otimes \boldsymbol{q}_i\right]^T + \frac{1}{\lambda_k}\left[\boldsymbol{q}_i^T \otimes \left[\boldsymbol{q}_k \otimes \boldsymbol{I} + \boldsymbol{I} \otimes \boldsymbol{q}_k\right]\left[\lambda_k\boldsymbol{I} - \boldsymbol{A}\right]^-\right]\right)\text{vec}(d\boldsymbol{A})$$

Proof. This follows by careful application of the differentials of an eigendecomposition, $d\lambda_k = \boldsymbol{q}_k^T d\boldsymbol{A}\boldsymbol{q}_k$ and $d\boldsymbol{q}_k = \left[\lambda_k\boldsymbol{I} - \boldsymbol{A}\right]^- d\boldsymbol{A}\boldsymbol{q}_k$ (Magnus and Neudecker, 2019, p. 180), and the mixed product properties of Kronecker products. □

Further note that the evaluation of $[\lambda_k \boldsymbol{I} - \boldsymbol{A}]^-$ can be computed trivially for any λ_k if the eigendecomposition of $\boldsymbol{A}$ has been obtained noting that if $\boldsymbol{B} = c\boldsymbol{I} + \boldsymbol{A}$, then the eigenvalues of $\boldsymbol{B}$ are the eigenvalues of $\boldsymbol{A}$ plus c and the eigenvectors are unchanged.

We also note a result from Magnus and Neudecker (2019) that simplifies for the generalized, i.e. $\boldsymbol{A}^- = \boldsymbol{A}^{\text{rank}(\boldsymbol{A})-}$

$$d\boldsymbol{A}^- = -\boldsymbol{A}^- d\boldsymbol{A}\boldsymbol{A}^- + \boldsymbol{A}^- \left[\boldsymbol{A}^-\right]^T d\boldsymbol{A}^T \left[\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^-\right] + \left(\boldsymbol{I} - \boldsymbol{A}^-\boldsymbol{A}\right) d\boldsymbol{A}^T \left[\boldsymbol{A}^-\right]^T \boldsymbol{A}^- \tag{19}$$

Now, Proposition E.2 defines the Wald statistic and its differential in the non-full rank setting.

Proposition E.2. *For a given r, assume that $r \leq \text{rank}(\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{A}^T)$. Define*

$$\widehat{W}_r = \left[\boldsymbol{A}\hat{\boldsymbol{\beta}}\right]^T \left[\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{A}^T\right]^{-r} \left[\boldsymbol{A}\hat{\boldsymbol{\beta}}\right],$$

and $\widehat{\boldsymbol{V}}_{\boldsymbol{A}}^{-r} = \left[\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{A}^T\right]^{-r}$. Then, the derivative is shown below:

$$\frac{\partial \widehat{W}_r}{\partial \boldsymbol{w}} = \left[\boldsymbol{A}\hat{\boldsymbol{\beta}} \otimes \boldsymbol{A}\hat{\boldsymbol{\beta}}\right]^T \frac{\partial \text{vec}\left(\widehat{\boldsymbol{V}}_{\boldsymbol{A}}^{-r}\right)}{\partial \boldsymbol{w}} + 2\left[\boldsymbol{A}\hat{\boldsymbol{\beta}}\right]^T \widehat{\boldsymbol{V}}_{\boldsymbol{A}}^{-r} \boldsymbol{A} \frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{w}}$$

Proof. This follows from previous results on the chain rule. The derivative of $\widehat{\boldsymbol{V}}_{\boldsymbol{A}}^{-r}$ can be computed using Lemma E.1. □

Efficient Evaluation As before, note that evaluating the derivative of $\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{A}^T$ with respect to $\boldsymbol{w}$ is undesirable for the reasons discussed in Appendix B. Note, however, that similar careful use of the differentials of quadratic forms can simplify computation dramatically. For example, note that for fixed $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$d\left[\boldsymbol{a}^T\boldsymbol{A}^{r-}\boldsymbol{b}\right] = \left(\sum_{k=1}^{r} -\frac{1}{\lambda_k^2}\left[(\boldsymbol{a}^T\boldsymbol{q}_k)\boldsymbol{q}_k \otimes (\boldsymbol{b}^T\boldsymbol{q}_k)\boldsymbol{q}_k\right]^T + \frac{1}{\lambda_k}\boldsymbol{q}_k^T \otimes \left[(\boldsymbol{a}^T\boldsymbol{q}_k)\boldsymbol{b} + (\boldsymbol{b}^T\boldsymbol{q}_k)\boldsymbol{a}\right]^T \left[\lambda_k\boldsymbol{I} - \boldsymbol{A}\right]^+\right) \text{vec}(d\boldsymbol{A}) \tag{20}$$

Noting that in the Wald setting, the differential on the right hand side becomes $(\boldsymbol{A}\otimes\boldsymbol{A})\text{vec}(d\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}))$. The term multiplying $\text{vec}(d\boldsymbol{A})$ is the sum of Kronecker products of $\boldsymbol{a}^T \otimes \boldsymbol{b}^T$ and thus the results from Proposition B.2 apply, so the computational complexity in the rank r case mirrors Proposition E.1.

Non-Integer r Wood (2013) proposes using the effective degrees of freedom τ of the relevant smooth as r; this requires a method for non-integer r generalized inverses. Wood proposes the following generalization where $k = \lfloor r \rfloor + 1$, $\nu = r - k + 1$ and $\rho = \sqrt{\nu(1-\nu)/2}$

$$\widehat{\boldsymbol{V}}_{\boldsymbol{A}}^{r-} = \boldsymbol{Q}\,\text{bdiag}\left(\text{diag}\left(\{1/\lambda_\ell\}\right)_{\ell=1}^{k-2}, \boldsymbol{B}\right)\boldsymbol{Q}^T; \quad \boldsymbol{B} = \tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{B}}\tilde{\boldsymbol{\Lambda}}; \quad \tilde{\boldsymbol{\Lambda}} = \begin{bmatrix} \lambda_{k-1}^{-1/2} & 0 \\ 0 & \lambda_k^{-1/2} \end{bmatrix}; \quad \tilde{\boldsymbol{B}} = \begin{bmatrix} 1 & \rho \\ \rho & \nu \end{bmatrix}$$

This can be re-expressed as

$$\widehat{\boldsymbol{V}}_{\boldsymbol{A}}^{r-} = \sum_{\ell=1}^{k-2} \frac{1}{\lambda_\ell}\boldsymbol{q}_\ell\boldsymbol{q}_\ell^T + \sum_{\ell=k-1}^{k} \boldsymbol{q}_\ell\boldsymbol{B}\boldsymbol{q}_\ell^T$$

Its partial derivative can be computed with respect to $\boldsymbol{w}$.

E.2 Derivation of Predicted Probabilities and Marginal Effects

We first consider the differentials of predicted probabilities and consider two types, although discussion works for any twice differentiable function $h(\cdot)$. The first is the predicted probability evaluated at a single observation, i.e. $h(\boldsymbol{x}_0^T\hat{\boldsymbol{\beta}})$. This would be used for, say, predicted probabilities setting all other covariates to their median and varying one covariate.

The second is the "average" predicted probability, i.e.

$$\hat{h}_0 = \frac{1}{N}\sum_{i=1}^{N} h(\boldsymbol{x}_{0,i}^T\hat{\boldsymbol{\beta}}) = \boldsymbol{c}_0^T h(\boldsymbol{X}_{0,i}\hat{\boldsymbol{\beta}}). \tag{21}$$

We assume, for generality, that we seek to evaluate $\{\hat{h}\}_{i=1}^{N_0}$ where $\hat{h}$ is defined as above for $\boldsymbol{c}_i$ and $\boldsymbol{X}_{0,i}$. $\boldsymbol{c}_i = 1$ and $\boldsymbol{X}_{0,i} = \boldsymbol{x}_0^T$ recovers the simple case shown in the main text. We define $\hat{\boldsymbol{h}}$ as the stacked (scalar) $\hat{h}_i$ into a vector. Defining $\tilde{\boldsymbol{c}}_i = \boldsymbol{X}_{0,i}\text{diag}(h'(\boldsymbol{X}_{0,i}\hat{\boldsymbol{\beta}})\boldsymbol{c}_i$, then the usual application of the delta rule gives us the variance of $\hat{h}_0$ and $\hat{\boldsymbol{h}}$, shown below.

$$\widehat{\text{Var}}(\hat{h}_0) = \tilde{\boldsymbol{c}}_0^T\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{c}}_0; \quad \widehat{\text{Var}}(\hat{\boldsymbol{h}}) = \tilde{\boldsymbol{C}}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{C}}^T; \quad \tilde{\boldsymbol{C}}^T = [\tilde{\boldsymbol{c}}_1, \cdots, \tilde{\boldsymbol{c}}_{N_0}] \tag{22}$$

Proposition E.3 finds the derivative of $\hat{\boldsymbol{h}}$ with respect to $\boldsymbol{w}$ and that it can be computed in favorable computational complexity.

Proposition E.3. *Define* $\tilde{\boldsymbol{C}}_i^J = \boldsymbol{X}_{0,i}^T \text{diag}\left(h''(\boldsymbol{X}_{0,i}\hat{\boldsymbol{\beta}})\right)\boldsymbol{X}_{0,i}$. *The derivative of* $\hat{\boldsymbol{h}}$ *and its variance with respect to* $\boldsymbol{w}$ *is shown below, where* $\boldsymbol{K}_{N_0,N_0}$ *is a commutation matrix.*

$$\frac{\partial \hat{\boldsymbol{h}}}{\partial \boldsymbol{w}} = \tilde{\boldsymbol{C}}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{w}}$$

$$\frac{\partial \widehat{\text{Var}}(\hat{\boldsymbol{h}})}{\partial \boldsymbol{w}} = \left[\tilde{\boldsymbol{C}} \otimes \tilde{\boldsymbol{C}}\right]\frac{\partial \widehat{\text{Var}}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{w}} + \left[\boldsymbol{I} + \boldsymbol{K}_{N_0,N_0}\right]\left[\boldsymbol{I} \otimes \tilde{\boldsymbol{C}}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\right]\begin{bmatrix}\tilde{\boldsymbol{C}}_1^J \\ \dots \\ \tilde{\boldsymbol{C}}_{N_0}^J\end{bmatrix}\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{w}}$$

Further, for $\boldsymbol{a}$ *and* $\boldsymbol{b}$, *the partial derivative of* $\boldsymbol{a}^T\widehat{\text{Var}}(\hat{\boldsymbol{h}})\boldsymbol{b}$ *with respect to* $\boldsymbol{w}$ *can be evaluated in* $\mathcal{O}(N_0p^2 + Np)$. *For some matrix* $\boldsymbol{A} \in \mathbb{R}^{M_0 \times N_0}$, *the derivatives of the Wald statistic corresponding to* $\boldsymbol{A}\hat{\boldsymbol{h}}$ *can be evaluated in* $O(N_0p^2 + Np + M_0^3)$.

Proof. Proof of this proposition begins by noting the following differentials, found using standard rules (see Magnus and Neudecker 2019):

$$d\hat{h}_i = \boldsymbol{c}_i^T \text{diag}\left(h'(\boldsymbol{X}_{0,i})\right)\hat{\boldsymbol{\beta}}))\boldsymbol{X}_{0,i}d\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{c}}_i^T d\hat{\boldsymbol{\beta}}$$
$$d\tilde{\boldsymbol{c}}_i = \boldsymbol{X}_{0,i}^T \text{diag}\left[h''(\boldsymbol{X}_{0,i}\hat{\boldsymbol{\beta}})\right]\boldsymbol{X}_{0,i}d\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{C}}_i^J d\hat{\boldsymbol{\beta}}$$

The derivative of $\hat{\boldsymbol{h}}$ is straightforward; the chain rule applied to $\widehat{\text{Var}}(\hat{\boldsymbol{h}})$ gives the derivative with respect to $\boldsymbol{w}$. Consider a quadratic form $\boldsymbol{a}^T\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{b}$; the derivative simplifies to

$$\frac{\partial \boldsymbol{a}^T\widehat{\text{Var}}(\hat{\boldsymbol{h}})\boldsymbol{b}}{\partial \boldsymbol{w}} = \left[\boldsymbol{b}^T\tilde{\boldsymbol{C}} \otimes \boldsymbol{a}^T\tilde{\boldsymbol{C}}\right]\frac{\partial \widehat{\text{Var}}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{w}} + \left[\boldsymbol{a}^T\begin{bmatrix}\boldsymbol{b}^T\tilde{\boldsymbol{C}}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{C}}_1^J \\ \dots \\ \boldsymbol{b}^T\tilde{\boldsymbol{C}}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{C}}_{N_0}^J\end{bmatrix} + \boldsymbol{b}^T\begin{bmatrix}\boldsymbol{a}^T\tilde{\boldsymbol{C}}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{C}}_1^J \\ \dots \\ \boldsymbol{a}^T\tilde{\boldsymbol{C}}\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})\tilde{\boldsymbol{C}}_{N_0}^J\end{bmatrix}\right]\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \boldsymbol{w}}$$

The first term can be evaluated at the cost of evaluating the derivative of a quadratic form of $\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})$ plus $\mathcal{O}(2N_0p)$. From Proposition B.2, this is $\mathcal{O}(Np)$. The second term can be evaluated at a cost of $\mathcal{O}(N_0p^2 + Np)$. Thus, the overall complexity is $\mathcal{O}(N_0p^2 + Np)$.

On the Wald statistic, for $\boldsymbol{A}$, this involves computing in the full-rank case

$$\widehat{W} = \left[\boldsymbol{A}\hat{\boldsymbol{h}}\right]^T \left[\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{h}})\boldsymbol{A}^T\right]^{-1} \left[\boldsymbol{A}\hat{\boldsymbol{h}}\right].$$

Using results from Proposition E.1, define $\check{\boldsymbol{a}} = \boldsymbol{A}^T\widehat{\boldsymbol{V}}_{\boldsymbol{A}}\boldsymbol{A}\hat{\boldsymbol{h}}$, this becomes using the mixed product property of Kronecker products and vectorization,

$$\frac{\partial\widehat{W}}{\partial\boldsymbol{w}} = -\frac{\partial\check{\boldsymbol{a}}^T\widehat{\text{Var}}(\hat{\boldsymbol{h}})\check{\boldsymbol{a}}}{\partial\boldsymbol{w}} + 2\left[\boldsymbol{A}\hat{\boldsymbol{h}}\right]^T\left[\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{h}})\boldsymbol{A}^T\right]^{-1}\boldsymbol{A}\frac{\partial\hat{\boldsymbol{h}}}{\partial\boldsymbol{w}}.$$

Given the above result on the cost of the derivative of the quadratic form, the computational cost is thus $\mathcal{O}(N_0p^2 + Np + M_0^3)$. □

Non-Full Rank Case: For a generalized Wald statistic with rank r, the above results can be immediately extended where the eigendecomposition of $\boldsymbol{A}\widehat{\text{Var}}(\hat{\boldsymbol{h}})\boldsymbol{A}^T$ is computed. The computational cost mirrors that in the standard generalized Wald by using the results in Proposition E.2 and Proposition E.3.

F Additional Tests for Empirical Analyses

This appendix contains additional tests for the empirical analyses.

F.1 Cook's Distance

In the linear model, Cook's distance, C_i, is defined as the impact that deleting an observation i has on the fitted values of the model or, equivalently, on the vector of estimated coefficients (Chatterjee and Hadi, 1986). There are many ways to write Cook's distance; we display two below. Using the notation $\hat{\boldsymbol{\beta}}_{-i}$ to mean estimating $\hat{\boldsymbol{\beta}}$ excluding observation i, i.e. $w_i = 0; \boldsymbol{w}_{-i} = 1$ and where r_i is the residual $y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ and h_{ii} is the leverage (i.e., $\boldsymbol{x}_i^T\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}\boldsymbol{x}_i$), C_i can be expressed as

$$C_i = \frac{\sum_{j=1}^{J}\left(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{-i}\right)^2}{\hat{\sigma}^2\cdot p} = \left[\frac{r_i}{(1-h_{ii})}\right]^2\cdot\frac{h_{ii}}{\hat{\sigma}^2\cdot p}. \tag{23}$$

There is a very close connection between Cook's distance and the influence scores we consider in this paper; using results from Chatterjee and Hadi (1986, p. 383), we can write C_i in terms of

the "exact" influence, i.e. $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{-i}$, as well as our influence scores $\partial\hat{\boldsymbol{\beta}}/\partial w_i$.[7]

$$C_i = \frac{1}{\hat{\sigma}^2 p}\left[\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_i\right]^T \boldsymbol{X}^T\boldsymbol{X}\left[\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_i\right] = \frac{1}{\hat{\sigma}^2 p}\frac{1}{(1-h_{ii})^2}\left[\frac{\partial\hat{\boldsymbol{\beta}}}{\partial w_i}\right]^T \boldsymbol{X}^T\boldsymbol{X}\left[\frac{\partial\hat{\boldsymbol{\beta}}}{\partial w_i}\right]. \tag{24}$$

Thus, Cook's distance is closely related to the sum of the squared influence scores used in the main analysis. This illustrates the *global* nature of Cook's distance and the source of its divergence from our measure of identifying influential observations for *specific* quantities insofar as a large influence on one coefficient $\hat{\boldsymbol{\beta}}_j$ does not necessarily imply a large influence on all coefficients.

For generalized linear models, if r_i is the Pearson residual for observation i, i.e. $(y_i - \hat{\mu}_i)/\sqrt{V(\hat{\mu}_i)}$ (see Appendix B) and the estimated dispersion is $\hat{\phi}$, then, following Williams (1987), an approximate C_i is defined as

$$C_i = \left[\frac{r_i}{(1-h_{ii})}\right]^2 \cdot \frac{h_{ii}}{\hat{\phi}\cdot p}. \tag{25}$$

In the case of a generalized additive model, Wood (2017) replaces p with the effective degrees of freedom τ (see Appendix D).

Now, in our empirical examples, we can compare Cook's distance against the $\partial\phi(\hat{\boldsymbol{\beta}}, \widehat{\text{Var}}(\hat{\boldsymbol{\beta}}))/\partial\boldsymbol{w}$ used to create the AMIS. We find that, unsurprisingly, there is a rather weak relationship. Thus, the measures we consider in this paper are tapping into something both conceptually and empirical distinct from Cook's distance.

Figures 9 and 10 compares Cook's distance against the (absolute value of the) influence scores for $\hat{\boldsymbol{\beta}}$, the corresponding confidence interval bound $\text{CI}(\hat{\boldsymbol{\beta}})$ in the two empirical examples in Section 5, and, for Woldense and Kroeger (2024), the average marginal effect. It shows that, in all cases, points with high influence—i.e. those in the AMIS—are not the points with the highest Cook's distance.

[7]Note that in a linear regression $\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} - i$ has an exact closed form of $\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}\boldsymbol{x}_i r_i/(1-h_{ii})$ which is the $\partial\hat{\boldsymbol{\beta}}/\partial w_i$ scaled by $1/(1-h_{ii})$.

Figure 9: Cook's Distance for Giger and Klüver (2016)

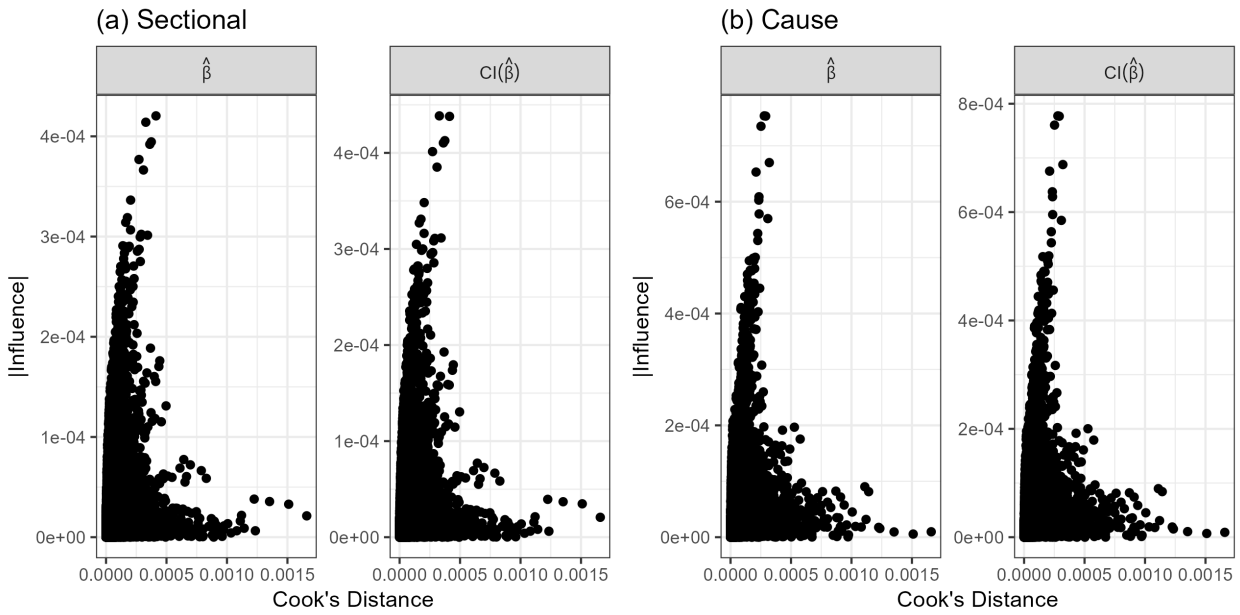


Figure 10: Cook's Distance for Woldense and Kroeger (2024)