Discussion: Misspecification in Econometrics

Hiroaki Kaido
Boston University

Abstract

This discussion reviews recent advances in econometric analysis of model misspecification, drawing on the chapters by Armstrong and by Andrews, Chen, and Tecchio. It highlights how misspecification affects bias, variance, and the interpretation of estimands, and discusses the role of expanded models that capture plausible departures from baseline assumptions. The discussion contrasts local and global misspecification frameworks, examines their implications for robust inference, and underscores the role of user-chosen components—such as weighting matrices and tuning parameters—in shaping empirical conclusions. By connecting these ideas to examples from causal inference, social interaction models, and partially identified settings, the discussion outlines directions for developing inference procedures that remain credible in the presence of model misspecification.

Key words: Model misspecification, Robustness, Credible Inference

1 Introduction

The econometrics literature has long examined how to build and test empirical models. The standard toolkit available to applied researchers includes specification tests for a wide range of parametric and moment-based models. Econometric theory has also developed formal frameworks for analyzing local misspecification, robustness, and bias-aware inference. The implications of misspecification and specification testing for inference, such as their effects on standard errors and post-regularization procedures, have been extensively studied. Building on these developments, the preceding chapters by Armstrong and by Andrews, Chen, and Tecchio (ACT) address key current issues related to misspecification and offer practical recommendations for both econometricians and applied researchers.

These chapters share several recurring themes. First, they underscore the challenges and pitfalls associated with commonly used estimands, such as those based on linear models, when these models are misspecified. Second, they highlight the gap between the textbook treatment of misspecification and empirical practice. Third, both contributions offer practical guidance on how to handle user-chosen components in estimation.

A point that merits attention is the gap between the textbook treatment of misspecification and empirical practice. On the one hand, textbooks emphasize formal model checks based on statistical tests—such as the Hausman test for models that admit both efficient and robust estimators (Hausman, 1978), the information matrix test for parametric models (White, 1982), and the J-test for moment-based models (Hansen, 1982). On the other hand, as documented by ACT, such tests are rarely reported in recent empirical work, raising the question of why this gap exists. Several factors may explain it. First, conventional specification tests can indicate that a model is misspecified but offer little guidance on how to fix the problem; moreover, a non-significant result does not imply that the model is correctly specified. Second, practitioners often rely on informal methods of assessing model fit, and such practices have become increasingly prevalent. Third, concerns about sequential testing, regularization, and their implications for valid inference may discourage researchers from conducting formal tests, leading to a preference for more flexible modeling approaches. Given this state of the empirical practice, the two chapters provide a timely review of existing tools and introduce new proposals aimed at making inference more credible in the presence of potential misspecification. The discussion below highlights several key insights from these chapters and outlines directions for further research.

1.1 Impacts of Misspecification on Bias and Variance

The chapter by Armstrong offers a framework for assessing how model misspecification affects inference. The central idea is to broaden a baseline model along the key dimensions where misspecification is most likely to occur. This approach requires the researcher to be explicit about (i) how the model is expanded and (ii) which class of estimators or inference methods is considered. Implementing such an analysis, in turn, calls for examining the statistical properties of the proposed procedures within the expanded model.

A key object is the *influence function* of the estimand $T(\theta_0)$ of interest. For an asymptotically linear estimator $\hat{T}$ of $T(\theta_0)$ such that $\sqrt{n}(\hat{T}-T(\theta_0)) = n^{-1/2}\sum_{i=1}^{n} IF(W_i)+o_p(1)$, $IF(\cdot)$ is called the influence function. It characterizes the sensitivity (and hence potential bias) of the estimator to perturbations of the data-generating process (Hampel, 1974; Ichimura and Newey, 2022).[1] The set of admissible perturbation directions, and thus the departures for which $E_P[IF(W)]$ may be nonzero, is determined by how the baseline model is expanded. In many applications, this expansion governs both worst-case bias and variance. Consequently, the statistical risk of $\hat{T}$ depends on the expansion amount M as well as on any user-chosen components of the estimation procedure. For example, consider a model characterized by moment conditions

$$E[g(X_i, \theta)] = 0 \tag{1.1}$$

for some $\theta \in \Theta$. One way to expand this baseline model is to allow for *local misspecification*, where some violation $c/\sqrt{n}$ of the moment restrictions with the same order of sampling variation is allowed. This leads to a bias-variance trade-off of the following form:

$$\overline{bias}^2 = \sup_{c \in \mathcal{C}(M)} |k_W' c/\sqrt{n}|^2, \quad \text{and} \quad V(k_w) = k_W' \Sigma k_W/n, \tag{1.2}$$

where $\mathcal{C}(M) = \{c : \|c\| \leqslant M\}$, and k_W is the GMM estimator's influence function as given in Armstrong's chapter, and M controls the size of the expanded model. The weighting matrix can then be chosen to balance these quantities (Armstrong and Kolesár, 2021b). In this expanded model, the amount of expansion enters the worst-case bias but not the variance.

[1] If the data-generating distribution changes from P to $P_\epsilon = (1-\epsilon)P + \epsilon\delta_w$, where δ_w represents a point mass (or an approximating sequence of it) at w, then the pathwise derivative of the functional $T(\theta(P))$ in the direction of w is $\frac{d}{d\epsilon}T(\theta(P_\epsilon))\big|_{\epsilon=0} = IF(w)$.

This framework is useful when a small degree of misspecification around a trusted model is suspected.

As discussed in ACT's chapter (Section 4), an alternative is to work under global misspecification, where the moment condition fails, i.e., $0 \notin \{E[g(X_i, \theta)] : \theta \in \Theta\}$. In this case, the GMM estimator converges to the minimizer of the population criterion, and smooth functionals of this pseudo-true value remain estimable. However, the bias does not vanish with sample size, and the asymptotic variance generally differs from that under correct specification. ACT review and recommend misspecification-robust variance estimators, either analytic formulas or the bootstrap without re-centering, that are consistent under global misspecification. The same guidance extends to semiparametric moment-condition models. In particular, Ai and Chen (2007) derive the asymptotic variance for semiparametric minimum-distance estimators under misspecification and show that, with additional structure (e.g., linearly separable nuisance components), misspecification need not affect the first-order variance.

Certain combinations of expanded models and estimators can yield tractable characterizations of statistical risk that provide insights into the trade-off between robustness and efficiency. Examples of such parameter spaces include (i) Huber's gross error model, (ii) convex parameter spaces for parametric and nonparametric models, and (iii) the set of distributions characterized by convex capacities. Key results on robust inference and optimality in such classes were established in the statistics literature (Huber, 1964; Huber and Strassen, 1973; Donoho, 1994), which have been applied to regression and moment-based models (Armstrong and Kolesár, 2018, 2021a,b) and structural incomplete models (Kaido and Zhang, 2019). Exploring such structures further and extending the analysis to a broader class of applications is a natural next step. For instance, a growing literature on models of social interaction investigates how different modeling assumptions affect identification and inference (Aronow and Samii, 2017; Athey et al., 2018; Leung, 2022; Sävje, 2024). To illustrate and connect with the examples discussed in the chapter, consider a simple setting in which the outcome is generated by

$$Y_i = \alpha + \tau D_i + r(S_i) + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} (0, \sigma^2),$$

where $r(\cdot)$ is an unknown function capturing spillover effects, and $S_i = S_i(D) \in \mathcal{S} = [0,1]$ denotes the exposure mapping (Manski, 2013; Aronow and Samii, 2017)—for example, the fraction of treated peers within a cluster such as a school—which depends on the treatment vector D. A simple approach to estimating τ is to regress Y_i on D_i while approximating $r(\cdot)$ by a step function defined over K exposure bins $B = \{B_1, \ldots, B_K\}$ forming a partition of $\mathcal{S}$. In this specification, misspecification of either $r(\cdot)$ or $S_i(\cdot)$ may induce bias by conflating direct and indirect effects, and may also inflate variance: the coarseness of the exposure partition affects the sampling variability of the estimator. Different bias–variance tradeoffs can thus emerge depending on the choice of the expanded parameter space for $(r(\cdot), \{S_i(\cdot)\}_{i=1}^n)$, the sampling design, and the asymptotic approximation adopted.

Returning to the point on how to seek robustness, there are often multiple ways to proceed when expanding a baseline model. Armstrong's chapter categorizes them into several broad classes: (i) data-driven lower bounds for M; (ii) model augmentation through auxiliary assumptions; (iii) placebo tests; and (iv) model expansions defined through statistical distance measures. This menu will likely continue to grow in the coming years.

These approaches differ in the dimensions along which robustness to misspecification is sought, but they share the common feature of explicitly defining an expanded model. A central question is how to choose an appropriate parameterization, particularly the degree of expansion, as this choice directly affects inference. Recent studies have explored ways to guide the choice of sensitivity parameters—such as those discussed in Category (ii) in the context of omitted variables, by Manski and Pepper (2018); Rambachan and Roth (2023) in event-study settings, and similar developments have emerged for calibrating the degree of unobserved confounding, using either domain knowledge or data-driven heuristics (Hsu and Small, 2013; Kallus and Zhou, 2021). While formalizing this process remains challenging, connecting the selection of sensitivity parameters to an underlying decision problem may provide a promising path forward. Such a connection has been explored in the context of treatment choice by Kallus and Zhou (2021).

1.2 Misspecification, Estimands, and Credible Inference

The chapter by ACT highlights four key themes: (i) the distinction between econometric misspecification, a failure of the underlying model assumptions, and statistical misspecification, which concerns the discrepancy between the observable data distribution and the class of permitted distributions; (ii) the definition and interpretation of estimands under misspecification; (iii) inference procedures under global misspecification, emphasizing the role of robust standard errors; and (iv) a novel use of the J-statistic under local misspecification. Taken together, these elements provide a coherent framework for diagnosing and interpreting model misspecification, clarifying what the researcher is estimating when some of the assumptions fail, and guiding inference in such settings.

The chapter discusses these points effectively, using examples from linear causal and moment-based models. Here, I expand the discussion of a nonparametric potential-outcome framework in Section 2.1 of ACT.

Example 1 (Potential Outcome Model): Let $(Y, D, Z) \in \mathcal{Y} \times \mathcal{D} \times \mathcal{Z}$ denote the observed outcome, treatment, and instrument. Let $Y^* = (Y(d,z))_{d\in\mathcal{D},\, z\in\mathcal{Z}}$ and $D^* = (D(z))_{z\in\mathcal{Z}}$ be potential outcomes and treatments. Assume Z is independent of (Y^*, D^*), possibly conditional on covariates. In ACT's notation, Q is the joint law of (Y^*, D^*) and P is the law of (Y, D, Z). Many causal estimands are functionals of Q, e.g. average and policy-relevant treatment effects

$$\theta(Q) = E_Q[Y(d) - Y(d')] \quad \text{and} \quad \theta(Q) = E_Q\big[Y(D(z)) - Y(D(z'))\big],$$

which makes target parameters explicit and interpretable.

Key assumptions in this class include exclusion and shape restrictions on (Y^*, D^*). For example, Imbens and Angrist (1994) impose (i) *exclusion*: $Y(d,z) = Y(d,z')$ for all $(d,z,z') \in \mathcal{D} \times \mathcal{Z}^2$, and (ii) *selection monotonicity*: for any $z, z' \in \mathcal{Z}$, either $D(z) \geqslant D(z')$ or $D(z) \leqslant D(z')$. These conditions restrict the support of (Y^*, D^*) and hence the admissible set $\mathcal{Q}$ of joint laws, helping define and interpret causal parameters and discipline unobserved heterogeneity. Other support restrictions include monotonicity of outcomes or vector-valued treatments and mediation assumptions (Heckman and Pinto, 2018; Mogstad et al., 2021; Kwon and Roth, 2024). As pointed out by ACT, the joint distribution Q on a postulated support can be viewed as a *structure* that remains fixed when we undertake a policy change.

Depending on θ and the maintained assumptions, $\theta(Q)$ may be point- or set-identified. The partial-identification literature has developed methods for characterizing sharp identification regions and inference for $\theta(Q)$ (Manski, 2003; Molinari, 2020). Although the modeling restrictions concern latent objects, they yield nontrivial testable implications. Examples include tests of IV validity (Balke and Pearl, 1997; Kitagawa, 2015; Kédagni and Mourifié, 2020), "judge IV" designs (Frandsen et al., 2023), encouragement designs (Bai et al., 2025) and full mediation (Kwon and Roth, 2024), among others.

In the context of moment-based models, the chapter makes a creative proposal for using the J-statistic as a diagnostic tool in credible empirical research. The GMM weighting matrix W is a user-specified component. While standard econometric texts describe the form of W that guarantees asymptotic efficiency and discuss practical implementations of the efficient weighting matrix, a substantial literature has noted that such estimators may exhibit poor finite-sample performance. The choice of the weighting matrix, therefore, remains a subtle and consequential aspect of empirical implementation.

The chapter by ACT introduces an important additional perspective. Because W is researcher-chosen, it can be strategically selected to produce more favorable empirical results. The J-statistic provides a natural diagnostic for assessing the credibility of inference in this setting. Specifically, within the local misspecification framework discussed in the previous section, they establish the following equivalence (in the asymptotic sense):

$$\mathrm{VIF}(\hat{\theta}_W) \leqslant 1+\tau^2 \qquad \Leftrightarrow \qquad \hat{\theta}_W \in [\hat{\theta}_{\text{eff}} \pm \tau\hat{\sigma}_{\text{eff}}\sqrt{J}],$$

where $\hat{\theta}_{\text{eff}}$ is an asymptotically efficient GMM estimator, $\hat{\sigma}_{\text{eff}}$ an estimator of the variance of $\hat{\theta}_{\text{eff}}$, and the variance inflation factor $\mathrm{VIF}(\hat{\theta}_W) = \hat{\sigma}^2_{\hat{\theta}_W}/\hat{\sigma}^2_{\hat{\theta}_{\text{eff}}}$ measures how much the asymptotic variance of $\hat{\theta}_W$ gets inflated relative to the efficient one. This relation implies that a smaller J leaves less scope for weight-hacking, while a larger J permits greater sensitivity to W. Reporting J therefore quantifies the potential for weight choice to matter. A natural question that follows is how practitioners should discipline themselves against weight-hacking when J is large. One pragmatic approach may be to restrict attention to a more structured subclass of weighting matrices—for example, diagonal or block-diagonal forms—which limit flexibility. Further work on principled restrictions and reporting practices would be valuable. Finally, the insight behind the J-statistics may extend to other settings, such as regular semiparametric models (Chen and Santos, 2018). The key structure behind the above result is that an estimator can be expressed as $\hat{\theta}_W = \hat{\theta}_{\text{eff}} + v_W' Z$ where the second term captures the W-specific departure from efficiency, and the J-statistic quantifies the size of Z. Thus, in any problem where a researcher-chosen component induces a deviation from an efficient estimator, the same reasoning can bound how much that choice can move the estimate. A full treatment would require model-specific regularity conditions; nonetheless, the approach taken by ACT can potentially inform a wide range of problems.

More broadly, researchers often encounter settings where (i) the model may be misspecified yet difficult to detect in finite samples; (ii) inference relies on user–chosen components; and (iii) diagnostics that summarize the extent of potential misspecification are available. Even when the structure discussed above for the J-statistic does not apply verbatim, it is useful to develop methods that translate such diagnostics into credible inference.

As a concrete example, consider structural discrete-choice models for individual decisions or firm competition, which allow flexible treatment of partially understood elements such as choice-set formation or equilibrium selection (see, e.g., Tamer, 2003; Barseghyan et al., 2021). In these models, misspecification of primitives, such as functional forms, solution concepts, and exogeneity assumptions, remains a central concern. Under misspecification, the set of parameter values satisfying the model-implied restrictions may be empty, analogous to what occurs in globally misspecified moment equation models. This observation motivates an approach that conducts inference on a non-empty set of pseudo-true parameter values, which remains well-defined even when the model is misspecified. In this spirit, Kaido and Molinari (2025) and Andrews and Kwon (2024) study inference for pseudo-true parameters under misspecification: the former defines pseudo-true values via minimization of the Kullback–Leibler divergence, following White (1982), while the latter defines them in terms of minimal violations of moment inequality restrictions.

When the degree of misspecification is mild, the identified set may remain non-empty, but inference requires additional care. Andrews and Kwon (2024) show that existing specification tests often have limited power to detect mild misspecification, and that inference procedures which ignore misspecification can be substantially distorted. Nevertheless, the magnitude of violations of certain moment restrictions, such as those underlying specification tests, may still be informative, in a manner analogous to the J-statistic. Designing procedures that incorporate such quantities into estimation and reporting to improve transparency is a promising direction.

References

Ai, C. and X. Chen (2007): "Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables," *Journal of Econometrics*, 141, 5–43.

Andrews, D. W. K. and S. Kwon (2024): "Misspecified Moment Inequality Models: Inference and Diagnostics," *Review of Economic Studies*, 91, 45–76.

Armstrong, T. B. and M. Kolesár (2018): "Optimal Inference in a Class of Regression Models," *Econometrica*, 86, 655–683.

——— (2021a): "Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness," *Econometrica*, 89, 1141–1177.

——— (2021b): "Sensitivity analysis using approximate moment condition models," *Quantitative Economics*, 12, 77–108.

Aronow, P. M. and C. Samii (2017): "Estimating Average Causal Effects Under General Interference, with Application to a Social Network Experiment," *The Annals of Applied Statistics*, 11, 1912–1947.

Athey, S., D. Eckles, and G. W. Imbens (2018): "Exact P-Values for Network Interference," *Journal of the American Statistical Association*, 113, 230–240.

Bai, Y., S. Huang, and M. Tabord-Meehan (2025): "Sharp Testable Implications of Encouragement Designs," ArXiv preprint: 2411.09808.

Balke, A. and J. Pearl (1997): "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

Barseghyan, L., M. Coughlin, F. Molinari, and J. C. Teitelbaum (2021): "Heterogeneous choice sets and preferences," *Econometrica*, 89, 2015–2048.

Chen, X. and A. Santos (2018): "Overidentification in Regular Models," *Econometrica*, 86, 1771–1817.

Donoho, D. L. (1994): "Statistical Estimation and Optimal Recovery," *The Annals of Statistics*, 22, 238–270.

Frandsen, B., L. Lefgren, and E. Leslie (2023): "Judging judge fixed effects," *American Economic Review*, 113, 253–277.

Hampel, F. R. (1974): "The Influence Curve and Its Role in Robust Estimation," *Journal of the American Statistical Association*, 69, 383–393.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hausman, J. A. (1978): "Specification Tests in Econometrics," *Econometrica*, 46, 1251–1271.

Heckman, J. J. and R. Pinto (2018): "Unordered monotonicity," *Econometrica*, 86, 1–35.

Hsu, J. Y. and D. S. Small (2013): "Calibrating Sensitivity Analyses to Observed Covariates in Observational Studies," *Biometrics*, 69, 803–811.

Huber, P. and V. Strassen (1973): "Minimax tests and the Neyman-Pearson lemma for capacities," *Annals of Statistics*, 1, 251–263.

Huber, P. J. (1964): "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35, 73–101.

Ichimura, H. and W. K. Newey (2022): "The Influence Function of Semiparametric Estimators," *Quantitative Economics*, 13, 29–61.

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467.

Kaido, H. and F. Molinari (2025): "Information Based Inference in Models with Set-Valued Predictions and Misspecification," ArXiv:2401.11046.

Kaido, H. and Y. Zhang (2019): "Robust likelihood ratio tests for incomplete economic models," ArXiv preprint arXiv:1910.04610.

Kallus, N. and A. Zhou (2021): "Minimax-Optimal Policy Learning Under Unobserved Confounding," *Management Science*, 67, 2870–2890.

Kédagni, D. and I. Mourifié (2020): "Generalized instrumental inequalities: testing the instrumental variable independence assumption," *Biometrika*, 107, 661–675.

Kitagawa, T. (2015): "A Test For Instrument Validity," *Econometrica*, 83, 2043–2063.

Kwon, S. and J. Roth (2024): "Testing Mechanisms," ArXiv: https://arxiv.org/abs/2404.11739.

Leung, M. P. (2022): "Causal Inference Under Approximate Neighborhood Interference," *Econometrica*, 90, 267–293.

Manski, C. F. (2003): *Partial Identification of Probability Distributions*, Springer Series in Statistics, New York, NY: Springer.

——— (2013): "Identification of Treatment Response with Social Interactions," *Econometrics Journal*, 16, S1–S23.

Manski, C. F. and J. V. Pepper (2018): "How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions," *The Review of Economics and Statistics*, 100, 232–244.

Mogstad, M., A. Torgovitsky, and C. R. Walters (2021): "The Causal Interpretation of Two-Stage Least Squares with Multiple Instrumental Variables," *American Economic Review*, 111, 3663–3698.

Molinari, F. (2020): "Chapter 5 - Microeconometrics with partial identification," in *Handbook of Econometrics, Volume 7A*, ed. by S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin, Elsevier, vol. 7 of *Handbook of Econometrics*, 355–486.

Rambachan, A. and J. Roth (2023): "A More Credible Approach to Parallel Trends," *The Review of Economic Studies*, 90, 2555–2599.

Sävje, F. (2024): "Causal Inference with Misspecified Exposure Mappings: Separating Definitions and Assumptions," *Biometrika*, 111, 1–15.

Tamer, E. (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *The Review of Economic Studies*, 70, 147–165.

White, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1–25.