Imputation of Counterfactual Outcomes when the Errors are Predictable: Rejoinder

Sílvia Gonçalves*
Department of Economics, McGill University

and Serena Ng

Department of Economics, Columbia University and NBER

August 15, 2024

We thank Bruno Ferman (BF), Marcelo Medeiros (MM), Yuya Sasaki (YS), and Kaspar Wüthrich (KW) for their constructive comments on our paper. Below are some further thoughts.

1 On the Interpretation of pup

The discussants offered alternative interpretations of PUP. As we noted, PUP-like alternatives are already available in the literature. Assuming that the covariate space is small relative to the sample size, our derivations in Section 4 provide a formal justification for exploiting residuals of other series as predictors. This is similar to the implementation of FARMTREAT considered in Fan et al. (2022) developed to accommodate heterogeneity and high-dimension covariates in estimation. But while FARMTREAT is motivated as a variable selection procedure in a high dimension setting, we motivated it from the viewpoint of optimal prediction when the number of predictors is fixed. We thank MM for providing new simulations to reinforce the usefulness of PUP in other settings. Also mentioned is ARCO of Carvalho et al. (2018) which delivers a time averaged treatment effect over the post-intervention sample. We have primarily focused on the building block of the average, which is the treatment effect for unit i at a given $T_0 + h$.

KW shows that if selection into treatment is based on past shocks with $D_i = 1(e_{i,T_0} < 0)$, the PUP term $\rho e_{i,T_0}$ will correct for selection bias in a standard prediction if e_{it} is correctly specified to be an AR(1). We thank KW for this simple and convincing interpretation.

It is well known that a Difference-in-Difference regression identifies the average treatment effect δ under a parallel trend assumption. If the dynamics are correctly specified, then a Lagged Dependent

*The authors gratefully acknowledge support from the NSERC grant (Canada) and the National Science Foundation (Ng SES 018369)

Variable (LDV) model can also identify δ. YS first notes that the PUP regression $Y_{i,T_0+h} = \alpha_i + \lambda_{T_0+h} + \delta D_{i,T_0+h} + \rho_h Y_{i,T_0+h} + e_{iy,T_0+h}$+error nests DiD and LDV as special cases, and then points out that PUP can estimate δ without imposing the identifying assumptions required of DiD or LDV. In this regard, PUP is doubly robust. This is an insightful observation, though it would be prudent to qualify that robustness is limited to the two nested models, and it is possible that the true model is neither DiD or LDV.

Hsiao et al. (2012) remarked that a vector-autoregressive model (VAR) can be used for what we refer to as $\mathcal{M}(\cdot)$. This motivates MM to note that PUP is closely related to a VAR. This is true to the extent that both take dynamics into account. By the same argument that a LDV can identify δ, a VAR (which is a multivariate LDV) will also identify δ if the dynamics are correctly specified. But when the VAR is not correctly specified such as when $\mathcal{M}^*$ is a VARMA, there is still a role for PUP.

BF draws attention to the fact that the error term is specific to the model being analyzed, and thus to the assumptions underlying the choice of the conditional mean model $\mathcal{M}(\beta; \mathcal{H})$ used to estimate m_{it}. We cannot agree with this more because by construction, $e_{it} = Y_{it}(0) - m_{it}$ is defined from m_{it}. A 'model' depends not only on the form of $\mathcal{M}(\cdot)$, but also the conditioning information $\mathcal{H}$ which is necessarily context dependent. In a time series setting, a natural definition of $\mathcal{H}$ is the history of the dependent variable Y and its covariates X as of T_0, which we may denote by $\mathcal{H}_{T_0}$. We prefer to define $\mathcal{H}_{T_0}$ in terms of e since it is spanned by Y and X, but is more parsimonious if there are many X's. However, in the treatment effect setting considered in Section 4, we have data not just for the treated up to T_0, but also the post-treatment outcomes of the control group. It is inappropriate in this case to index $\mathcal{H}$ by T_0. Failing to find a satisfactory notation, we opted to simply write $\mathcal{H}$. We note, however, that $\mathcal{H}$ should only include observables or functions of them used in estimation and imputation, for if excluded variables were helpful, they should have been used in the first place.

2 On Misspecified Models

The discussants raised several aspects of misspecification that are of interest. Suppose for simplicity that the treated unit is $i = 1$ and recall that we view $\mathcal{M}(\cdot; \mathcal{H})$ as the 'pseudo-true' mean conditional on information $\mathcal{H}$ but that it may not coincide with the true $\mathcal{M}^*$. If m_{1t} generated by $\mathcal{M}$ coincides with m^*_{1t} generated by $\mathcal{M}^*$, the error $e_{1t} = Y_{1t}(0) - m_{1t}$ should not be predictable. KW and YS consider settings when $\mathcal{M}(\cdot; \mathcal{H})$ is specified such that $E[e_{1t}|\mathcal{H}] \neq 0$ even though $E[e_{1t}] = 0$, and in both cases, PUP improves prediction. In a sense, PUP improves upon $\mathcal{M}(\cdot;)$ by adding an estimate of $E[e_{1t}|\mathcal{H}]$ to m_{1t}. PUP has a control function flavor, but re-estimation is not involved.

Another type of misspecification concerns the model for e_{1t} used to generate the PUP correction for a given choice of $\mathcal{M}(\cdot; \mathcal{H})$. Now the premise of PUP is not to find the correct model for e_{1t}, but rather to have a 'good enough' model that would mop up its predictability as much as possible. After all, quoting the statistician George Box, 'all models are wrong, but some models are more useful'. Whether the AR(1) model is useful in this context depends on the predictability that remains in $\hat{e}^+_{1,T_0+1} = \hat{e}_{1,T_0+1} - \hat{\rho}_1 \hat{e}_{1,T_0}$. This can be checked by studying the correlogram of $\hat{e}^+_{1t}$.

With some data snooping, prediction error can be further reduced by adjusting the model for $\hat{e}_{1t}$.

An anonymous referee asked about the merits of a correction based on a model richer than the simple AR(1) that we focused in the paper. Unfortunately, even if we take minimizing prediction mean-squared error as our goal, we were unable to obtain precise results for an AR(2) prediction. The main reason is that the variance of PUP becomes a complicated expression of the autocorrelation coefficients of e_{1t}. A good reference is Kunitomo and Yamamoto (1985), who derive the prediction mean square error for misspecified AR(p) models when the true DGP is AR(m) with $m \geq p$. As their Theorem 3 and Corollary 5 make clear, this variance is very involved even if we abstract from estimation uncertainty.

A more delicate issue is that while prediction mean-squared error will be smaller than that of the standard prediction even when the model for e_{1t} is not correctly specified as Lemma 1 shows, precise statements cannot be made for coverage. To see why, assume that $e_{1,T_0+1} \overset{d}{\sim} N(0, \sigma^2_{e,1})$. The standard prediction (i.e. without the PUP correction) has conditional coverage probability

$$\Phi\Big(\text{bias}_1 + \text{bias}_2 z_{1-\alpha/2}\Big) - \Phi\Big(\text{bias}_1 - \text{bias}_2 z_{1-\alpha/2}\Big) \neq 1-\alpha$$

as $T_0 \to \infty$. Distorted inference can arise because of a location (bias_1), a shift (bias_2) in the quantiles, or because the normality assumption is incorrect. We focused in the paper on bias_1 and bias_2 and showed that in the AR(1) case, $\text{bias}_1 = -\frac{\phi_1}{\sigma_{v,1}} e_{1,T_0}$ and $\text{bias}_2 = \frac{\sigma_{e,1}}{\sigma_{v,1}}$. In contrast, a PUP prediction interval

$$\left[\hat{\delta}_{1T_0+1} - \hat{\rho}_1 \hat{e}_{1,T_0} - \hat{\sigma}_{\delta_1} z_{1-\alpha/2}, \;\; \hat{\delta}_{1T_0+1} - \hat{\rho}_1 \hat{e}_{1,T_0} + \hat{\sigma}_{\delta_1} z_{1-\alpha/2}\right] \tag{1}$$

will have the correct coverage asymptotically if e_{1t} is a Gaussian AR(1) model as assumed. Note that $\hat{\sigma}_{\delta_1}$ is a consistent estimator of the standard error of the conditional distribution of $e_{1,T_0+1} - E[e_{1,T_0+1}|\mathcal{H}]$ given $\mathcal{H}$, which is equal to $\sigma_{v,1}$ under the AR(1) model. Furthermore, $\frac{e_{1,T_0+1} - \rho_1 e_{1,T_0}}{\sigma_{v,1}} \sim N(0,1)$ by the assumption of normality and the AR(1) specification for e_{1t}, and $z_{1-\alpha/2}$ is approximately 2. However, if e_{1t} is not an AR(1), bias_1 will not be zero and bias_2 will not be 1. Even if e_{1,T_0} is truly Gaussian, the implications of the two biases for inference are unclear.

BF's concern is that PUP inference will still be distorted when the model for e_{1t} is misspecified. To guard against misspecification, BF suggests an inference procedure that can yield misspecification robust intervals for δ_1. His procedure relies on an ingenious use of bounds on the misspecification bias for the conditional mean of e_{1,T_0+1} given $\mathcal{H}$, and avoids a distributional assumption on e_{1t} by using quantiles of the empirical distribution of the estimated residuals $\hat{e}_{1t}$. When applied to the standard estimator $\hat{\delta}_1$ for which $\hat{e}_{1t} = \hat{e}^+_{1t}$, this procedure can eliminate the two types of bias (bias_1 and bias_2) discussed above, as we now explain. Let $\hat{F}$ be the empirical cdf of $\hat{e}_{1t}$ at level α and $Q_{\hat{F}}(\alpha)$ be the corresponding quantile function. Note that under Gaussianity of e_{1t}, this quantile is asymptotically equivalent to $Q_{\hat{F}}(\alpha) = \hat{\sigma}_{e,1} z_\alpha$. A standard $100(1-\alpha)\%$ level prediction interval for δ_{1,T_0+1} is

$$\left[\hat{\delta}_{1,T_0+1} - Q_{\hat{F}}(1-\alpha/2), \quad \hat{\delta}_{1,T_0+1} - Q_{\hat{F}}(\alpha/2)\right].$$

The conditional coverage probability of this interval, given $\mathcal{H}$, is equal to

$$P\left(Q_{\hat{F}}(\alpha/2) \leq e_{1,T_0+1} \leq Q_{\hat{F}}(1-\alpha/2)|\mathcal{H}\right),$$

which is different from $1-\alpha$ for the two main reasons discussed above (bias$_1$ and bias$_2$).

The premise of BF is the existence of a worst case bias $\Delta \geq 0$ such that $|E[e_{1,T_0+1}|\mathcal{H}]| \leq \Delta$ almost surely. Letting $Q^*(\alpha)$ denote the α-quantile of $e_{1,t} - E[e_{1,T_0+1}|\mathcal{H}]$, a conditionally valid interval for δ_{1,T_0+1} is given by

$$\left[\hat{\delta}_{1,T_0+1} - \Delta - Q^*(1-\alpha/2), \quad \hat{\delta}_{1,T_0+1} + \Delta - Q^*(\alpha/2)\right].$$

Although valid conditionally, this interval is infeasible because it depends on $Q^*(\alpha/2)$ and $Q^*(1-\alpha/2)$, the quantiles of the conditional distribution of $e_{1,t} - E[e_{1,T_0+1}|\mathcal{H}]$. To obtain a conditionally valid interval that is robust to misspecification of the conditional mean $E[e_{1,T_0+1}|\mathcal{H}]$, BF relies on the fact that $Q_{\hat{F}}(u) - \Delta \leq Q^*(u) \leq Q_{\hat{F}}(u) + \Delta$ for any $u \in (0,1)$. The proposed level $100(1-\alpha)\%$ conditional prediction interval for δ_{1,T_0+1} is thus

$$\left[\hat{\delta}_{1,T_0+1} - 2\Delta - Q_{\hat{F}}(1-\alpha/2), \quad \hat{\delta}_{1,T_0+1} + 2\Delta - Q_{\hat{F}}(\alpha/2)\right].$$

This interval bears some resemblance to our PUP prediction interval given in (1) as both control for bias$_1$. Whereas PUP removes $-\hat{\rho}_1\hat{e}_{1,T_0}$ from both end points of the interval, which is the right centering for the AR(1) model, the BF interval adds and subtracts the bound Δ to adjust for the location bias. The cost of robustness is that when the AR(1) model is correctly specified, using $\pm\Delta$ increases the length of the interval. This effect is amplified by the fact that to correct for bias$_2$, the BF interval multiplies Δ by 2. It remains to be seen whether robustness can justify the possible loss in power due to a wider prediction interval.

PUP can be used in conjunction with any asymptotically unbiased estimator of $\mathcal{M}(\cdot)$, and many such estimators are available. A question of interest is whether we can test if the different predictions are equal. For the sake of discussion, suppose that we have two predictions A and B for Y_{T_0+h}. In the forecasting exercises when Y_{T_0+h} is eventually observed, a simple test is suggested in Diebold and Mariano (1995). For a given loss function $g(\cdot)$ and out-of-sample forecast errors $\{e^A\}$ and $\{e^B\}$, it holds that $E[g(e^A_{T_0+1:T_0+h}) - g(e^B_{T0+1:T_0+h})] = 0$ under the null hypothesis of equal predictability. Since $Y_{T_0+1:T_0+h}$ is observed, the Diebold-Mariano test amounts to evaluating if the sample average of $g(\hat{e}^A_{T_0+1:T_0+h}) - g(\hat{e}^B_{T_0+1:T_0+h})$ is zero. This test is, however, not feasible in the treatment effect setting because $Y_{T_0+h}(0)$ is not observed for any $h > 0$, making this extension challenging. However, if A is the standard prediction and B is PUP we can in principle do an an LM test as suggested in the paper. The size and power of such a test remains to be explored.

Disclosure Statement The authors report that there are no competing interests to declare.

Acknowledgement We also thank Atsushi Inoue and two anonymous referees for comments on the first draft of this paper. Comments from seminar participants at Stanford University, the 2024 SETA conference in Taipei, International Panel Data Conference at Orleans, France, and the 2024 NBER Summer Institute are also appreciated. This work was supported by an NSERC grant and the National Science Foundation (Ng SES 018369).

References

Carvalho, C., Masini, R. and Medeiros, M. 2018, ArCo: An Artificial Counterfactual Approach for High-Diemsnional Panel Time Series Data, *Journal of Econometrics* **207**, 352–380.

Diebold, F. and Mariano, R. S. 1995, Comparing Predictive Accuracy, *Journal of Business and Economic Statistics* **13:3**, 253–264.

Fan, J., Masini, R. and Medeiros, M. 2022, Do We Exploit All Information for Counterfactual Analysis? Benefits of Factor Models and Idiosyncractic Correction, *Journal of the American Statistical Association* **117:538**, 574–590.

Hsiao, C., Ching, H. and Wan, S. 2012, A Panel Data Approach for Program Evaluation: Measuring the Benefits of Political and Eocnomic Integration of Hong Kong with Mainland China, *Journal of Applied Econometrics* **27**, 705–740.

Kunitomo, N. and Yamamoto, T. 1985, Properties of Predictors in Misspecified Autoregressive Time Series Models, *Journal of the American Statistical Assocation* **80:392**, 941–950.