Imputation of Counterfactual Outcomes when the Errors are Predictable

Sílvia Gonçalves* Serena Ng†

June 29, 2024

Abstract

A crucial input into causal inference is the imputed counterfactual outcome. Imputation error can arise because of sampling uncertainty from estimating the prediction model using the untreated observations, or from out-of-sample information not captured by the model. While the literature has focused on sampling uncertainty, it vanishes with the sample size. Often overlooked is the possibility that the out-of-sample error can be informative about the missing counterfactual outcome if it is mutually or serially correlated. Motivated by the best linear unbiased predictor (BLUP) of Goldberger (1962) in a time series setting, we propose an improved predictor of potential outcome when the errors are correlated. The proposed PUP is practical as it is not restricted to linear models, can be used with consistent estimators already developed, and improves mean-squared error for a large class of strong mixing error processes. Ignoring predictability in the errors can distort conditional inference. However, the precise impact will depend on the choice of estimator as well as the realized values of the residuals.

Keywords: treatment effect, synthetic control, missing values, BLUP.

JEL classification: C22, C23

*McGill University, Department of Economics, 855 Sherbooke St. W., Montreal, PQ, H3A,2T7. Email: silvia.goncalves@mcgill.ca

†Columbia University and NBER. Department of Economics, 420 W. 118 St. MC 3308, New York, NY 10027. Email: serena.ng@columbia.edu

We thank Atsushi Inoue, Simon Lee, Jushan Bai, Kaspar Wüthrich, Marcelo Medeiros, Bruno Ferman, Yuya Sasaki, Bruce Hansen, Jeff Wooldridge, Pedro Sant'Anna, Barbara Rossi, and participants at the JBES session at the 2024 ASSA meeting and the Gary Chamberlain Seminar in Econometrics for helpful comments. This work was supported by an NSERC grant and the National Science Foundation (Ng SES 018369).

1 Introduction

Understanding the effects of policies is an important aspect of economic analysis and many questions of interest involve an individual's or a group's response to multiperiod interventions. Given treatment status, a researcher observes the outcome of unit i after intervention at $t = T_0 + 1$ (denoted $Y_{it}(1)$) and wants to compare it to the hypothetical outcome without intervention (denoted $Y_{it}(0)$). Since we do not observe $Y_{it}(0)$ for $t > T_0$, these values need to be imputed. Now imputation is concerned with the prediction of values that will never be sampled, and from results in the prediction literature, we know that in-sample estimation uncertainty should diminish with the sample size and what dominates total prediction error asymptotically is the variation not explained by the model. Yet, in applications, we tend to perform robust inference taking the residuals as given, when an improved prediction is possible by removing predictable variations that might still be in the residuals.

To illustrate, consider Figure 1 which studies the impact of the German reunification in 1990 on Y_{1t} = log GDP. Because the GDP data are non-stationary, we estimate common factors from a 16 country panel of GDP growth $(\Delta Y_{2,1:T}(0), \ldots, \Delta Y_{17,1:T}(0))$, where $\Delta Y_{i,1:T}(0) \equiv (\Delta Y_{i1}(0), \ldots, \Delta Y_{iT}(0))'$. The top-left panel displays actual (log) GDP over the full sample, along with the in-sample fit $\widehat{Y}_{1,1:T_0}(0)$, and the counterfactual values $\widehat{Y}_{1,T_0+1:T}(0)$. The effect of reunification on GDP is stark, but masks the fact that the (in-sample) residuals $\widehat{e}_{1,1:T_0} = \Delta Y_{1,1:T_0}(0) - \widehat{\Delta Y}_{1,1:T_0}(0)$ are persistent. This can be seen from the correlogram in the bottom left panel, or from the plot of the series itself in the top right panel. The series $\widehat{e}_{1,1:T_0}$ is also cross-correlated with other errors, though many of the $\text{corr}(\widehat{e}_{1,1:T_0}, \widehat{e}_{j,1:T_0})$ are not statistically significant as shown in the bottom right panel. The in-sample residuals of the log level model are also serially correlated, as shown in Figure 2. Time series and cross-section correlation of the in-sample residuals is not specific to this example.

This paper considers the implications of non-spherical errors for model-based imputation. Non-spherical errors, which induce predictability, can arise because the model for $Y_{it}(0)$ is mis-specified or because $Y_{it}(0)$ cannot be adequately captured by observed information without further signal extraction. We build on the best-linear-unbiased predictor (hereafter, BLUP) developed in Goldberger (1962) for linear models with non-spherical errors. The key to BLUP is not that it is based on GLS estimation, but that it has a correction term that depends on the covariance structure of the errors. We suggest a practical predictor (PLUP) that is asymptotically equivalent to the infeasible BLUP to a first order. Furthermore, if predictability is due to serial correlation, a simple AR(1) correction will reduce the mean-squared prediction error for a large class of stationary mixing error processes. This is not to say that an AR(1) correction is best as the desired adjustment will necessarily be data dependent, but to point out that simple modifications can reduce the mean-squared error of the standard prediction.

We adapt Goldberger's result for linear prediction to an imputation setting when the counterfactual outcomes are never observed. We derive infeasible BLUP for linear panel data models and make precise its dependence on the covariance structure of the errors. We show that when T_0 is large, a PLUP that controls for time and/or cross-section correlation can be constructed. But the idea of correcting the standard prediction for predictable errors is more general and can be applied to non-linear models when direct modeling of dynamics is not so straightforward. Thus when linearity is not required, we refer to the practical unbiased predictor as PUP.[1]

In addition to inefficient point predictions, ignoring correlation in the errors also has implications for inference. Though a standard prediction will yield asymptotically valid unconditional inference, the prediction interval will be wider because correlated residuals inflate error variance. More concerning is that standard prediction is biased conditional on pre-treatment outcomes of both the treated and the untreated, as well as the post-treatment outcomes of the untreated. This bias may distort inference and the precise impact will depend not just on persistence of the residuals, but also on the realized values of the residuals relevant for imputing $Y_{i,T_0+h}(0)$. In the German unification example, $\widehat{e}_{1,T_0}$ is 1.03 in the growth rate model. This non-zero value yields PUP growth rates that are slightly different from the standard prediction.

PUP is concerned with reducing the out-of-sample error and does not preclude the use of robust standard errors or resampling schemes to account for correlation in the in-sample residuals. In practice, a PUP for unit i will use residuals of unit i before treatment, and possibly of the untreated units after T_0. For serial correlation type error dependence that should die off as h increases, a PUP correction is most effective in imputating $Y_{i,T_0+h}(0)$ at small h. Recent work by Chernozhukov, Wüthrich, and Zhu (2021), Fan, Masini, and Medeiros (2022), and Ferman (2023) can be seen from a PUP perspective.

Our central message is that in-sample uncertainty is asymptotically dominated by variability of the out-of-sample prediction error, and more attention should be paid to improving the point-prediction before turning to inference. The rest of the paper proceeds as follows. Section 2 sets up the econometric framework and provides motivating examples for predictable errors. Section 3 summarizes the properties of BLUP and then presents PLUP. Its mean-squared error is analyzed using asymptotic expansions of the population prediction

[1]We thank Bruce Hansen for this suggestion.

error under mixing conditions. Section 4 switches focus from predicting future outcomes to imputing missing values. Unconditional and conditional coverage of prediction intervals are analyzed in Section 5. With some abuse of language, we sometimes use 'imputation' and 'prediction' interchangeably. Our discussion will focus on time dependence but the arguments also hold for spatial and cross-section dependence.

2 The Econometric Framework

We will use the standard potential outcome framework for analysis. Let $Y_{it}(1)$ be the potential response for unit i at time t if it was exposed to treatment (or policy intervention), and $Y_{it}(0)$ be the potential response of a (control) unit i that was not exposed to intervention at t. We observe $Y_{it} = Y_{it}(0)(1 - D_{it}) + Y_{it}(1)D_{it}$ where treatment status $D_{it} = 1$ if unit i is exposed in period t and is zero otherwise. Without loss of generality, we order the N_1 exposed units before the $N_0 = N - N_1$ unexposed units. We observe

$$Y_{it} = \begin{cases} Y_{it}(0), & i = 1, \dots, N, \quad t = 1, \dots, T_0 \\ Y_{it}(0), & i = N_1 + 1, \dots, N, \quad t = T_0 + 1, \dots, T \\ Y_{it}(1) & i = 1, \dots, N_1, \quad t = T_0 + 1, \dots, T \end{cases}$$

and are interested in the effect on unit $i \in [1, N_1]$ in $h > 0$ periods after treatment begins in $T_0 + 1$. Different average treatment effects can be derived from the individual treatment effect, defined as

$$\delta_{i,T_0+h} = \underbrace{Y_{i,T_0+h}(1)}_{\substack{\text{observed outcome}\\ \text{in period } T_0+h}} - \underbrace{Y_{i,T_0+h}(0)}_{\substack{\text{outcome without treatment}\\ \text{in period } T_0+h}}.$$

The econometrics challenge is that $Y_{it}(0)$ is not observed for $i \leq N_1$ when $t > T_0$.

Following the literature, we assume that $Y_{it}(0)$ has a pseudo-true conditional mean (or mean-unbiased proxy) $m_{it} = \mathcal{M}(\beta; \mathcal{H})$ that is parameterized by a vector β given some information set $\mathcal{H}$, and $e_{it} = Y_{it}(0) - m_{it}$ is such that $E(e_{it}) = 0$. For example, an AR(1) approximation would make $m_{it} = \rho y_{it-1}$ and $\mathcal{H}$ would be y_{is} for $s \leq T_0$. Being a pseudo-true mean, m_{it} may not coincide with the true conditional mean say, m^*_{it}, where $e^*_{it} = Y_{it}(0) - m^*_{it}$. For each $i = 1, \dots, N_1$,

$$\begin{aligned} Y_{it}(0) &= m_{it} + e_{it}, \qquad t = 1, \dots, T \\ Y_{it}(1) &= m_{it} + \delta_{it} + e_{it}, \qquad t > T_0. \end{aligned}$$

Let $\widehat{m}_{it}$ be a consistent estimate of m_{it}. Then

$$Y_{it}(0) = \widehat{m}_{it} + \widehat{e}_{it}.$$

Since $\widehat{Y}_{it}(0) = \widehat{m}_{it}$, the treatment effect on unit i at a given $t = T_0 + h$ is then estimated by

$$\begin{aligned}
\widehat{\delta}_{i,T_0+h} &= Y_{i,T_0+h}(1) - \widehat{Y}_{i,T_0+h}(0) \\
&= Y_{i,T_0+h}(1) - Y_{i,T_0+h}(0) + (Y_{i,T_0+h}(0) - \widehat{m}_{i,T_0+h}) \\
&= \delta_{i,T_0+h} + e_{i,T_0+h} + (m_{i,T_0+h} - \widehat{m}_{i,T_0+h}).
\end{aligned}$$

The pointwise imputation/prediction error is

$$\widehat{\delta}_{i,T_0+h} - \delta_{i,T_0+h} = (m_{i,T_0+h} - \widehat{m}_{i,T_0+h}) + e_{i,T_0+h}.$$

This error has two sources of variation: one from in-sample estimation of m_{it}, and one due to the out-of-sample error e_{i,T_0+h} not captured by the model. The first error will be negligible as T_0 increases provided that $\widehat{m}_{it}$ is consistent for m_{it} in some well defined sense, but the second error does not vanish with the sample size and thus total prediction error variance is minimized asymptotically if m_{it} is chosen such e_{it} does not contain predictable information. However, theory actually allows e_{it} to be serially and/or mutually correlated, and while the assumption of no correlation is convenient, it is not always appropriate. In the next subsection, we provide some examples for dependence in the errors.

2.1 Examples when e_{it} is predictable

We will first clarify what we mean by in-sample and out-of-sample errors. To fix ideas, suppose that unit 1 is being treated and the model is linear so that $m_{1t} = x_t'\beta$. Single equation estimation yields the imputed value $\widehat{\delta}_{1,T_0+1} = x'_{T_0+1}\widehat{\beta}$ and imputation error $\widehat{\delta}_{1,T_0+1} - \delta_{1,T_0+1} = -x'_{T_0+1}(\widehat{\beta} - \beta) + e_{1,T_0+1}$ whose variance is

$$\text{var}(\widehat{\delta}_{1,T_0+1} - \delta_{1,T_0+1}) = \sigma_e^2 + x'_{T_0+1}\text{var}(\widehat{\beta})x_{T_0+1}.$$

Correlation in $x_t e_t$ may necessitate robust standard errors for $\widehat{\beta}$, but provided that $E[x_t e_{1t}] = 0$, $\widehat{\beta}$ is consistent in the sense that $\text{var}(\widehat{\beta}) \to 0$ as $T_0 \to \infty$. Thus, the variance of imputation error is dominated by the out-of-sample error variance $\sigma_e^2 \equiv \text{var}(e_{1,T_0+1})$ asymptotically. This variance is minimized when e_{1,T_0+1} is uncorrelated. Serial correlation can arise because of temporal aggregation, but residual correlation (temporally or mutually) is usually a symptom of misspecification of the model or conditioning information. We give some examples below.

Example misspecification 1: Suppose that $Y_{1,t}(0) = \phi_1 Y_{1,t-1}(0) + \phi_2 Y_{1,t-2}(0) + v_{1t}$ is an AR(2) process with iid innovations v_{1t}, but the researcher assumes an AR(1) model. Then $m_{1t} = \beta Y_{1,t-1}(0)$, the pseudo true parameter is $\beta = \frac{\phi_1}{1-\phi_2} \neq \phi_1$, and $e_{1t} = v_{1t} + (\phi_1 - \beta)Y_{1,t-1}(0) + \phi_2 Y_{1,t-2}(0)$ is serially correlated when $\phi_2 \neq 0$.

Example misspecification 2: Suppose that the potential outcome has an interactive fixed effect structure: $Y_{1t}(0) = \lambda_1' F_t + \epsilon_{1t}$ but a researcher specifies an additive fixed effect model $Y_{1t}(0) = \lambda_1 + F_t + e_{1t}$. Then e_{1t} will be serially correlated if F_t is serially correlated, even if ϵ_{1t} is white noise.

Factor-based imputation assumes $X_{it} = \lambda_i' F_t + e_{it}$ where F_t is a vector of r latent common factors with λ_i as loadings and e_{it} is an idiosyncratic error. An appeal of factor-based imputation is that under some conditions, the space spanned by F can be consistently estimated without modeling the (weak cross-section or time) dependence in the idiosyncratic errors. While Xu (2017) iteratively estimates F and the missing values jointly by principal components (PCA), Bai and Ng (2021) impute the missing values (or complete the matrix) using two full-sample applications of PCA. Athey, Bayati, Doudchenko, Imbens, and Khosravi (2021); Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021) estimate the low rank component using singular value thresholding (SVT).[2] Though all consistent estimators of F imply that $\widehat{F}_{T_0+h}$ can be used as though they were observed predictors, the possibility remains that e_{1,T_0+h} can be predicted by information available.

Example Correlated Idiosyncratic Errors (from Fan, Masini, and Medeiros (2022)) Suppose that $Y_{it}(0) = \lambda_i' F_t + e_{it}$ and the researcher correctly assumes $m_{1t} = \lambda_1' F_t$ but e_{it} is correlated with e_{jt} for j in some index set C. Then $e_{1t} = v_{1t} + \sum_{j \in C} \theta_j e_{jt}$ is predictable by those e_{jt} where $j \in C$.

The method of synthetic control (SC) developed in Abadie and Gardeazabal (2003) assumes that there exist weights β_j^* such that a perfect fit $Y_{1t}(0) = \sum_{j=2}^N \beta_j^* Y_{j,t}(0)$ exists for every $t \leq T_0$. Abadie, Diamond, and Hainmueller (2010) make additional use of K economic predictors $X_t = (X_{1t}, X_{0t})$, where $X_{0t} = (X_{2t}, \ldots, X_{N,t})$ for the unexposed. The Synthetic Difference-in-Difference (SDID) in Arkhangelsky, Athey, Hirshberg, Imbens, and Wager (2021) also reweights the pre-treatment time periods to balance the pre-and post exposure time periods and nests SC and DID as special cases. However, an increasing number of papers suggest that an 'imperfect pretreatment fit' may prevent recovery of β^*.

Example Imperfect Fit 1: (from Ben-Michael, Feller, and Rothstein (2021)) Suppose that $Y_{it}(0) = \phi_1 Y_{i,t-1}(0) + v_{it}$ for all $i = 1, 2, \ldots, N$, and one constructs $m_{1t} = \sum_{j=2}^N \beta_j^* Y_{j,t}$. We can show that the error $e_{1t} = Y_{1t}(0) - m_{1t}$ can be decomposed as $e_{1t} = \phi_1 e_{1,t-1} +$

[2] Regularization is not necessary to consistently estimate the missing values, but could give a lower rank common component than the one in Bai and Ng (2021).

$v_{1t} - \sum_{j=2}^N \beta_j^* v_{jt}$. The error e_{1,T_0+1} contains an imbalance component $\phi_1 e_{1,T_0} = \phi_1(Y_{1,T_0} - \sum_{j=2}^N \beta_j^* Y_{j,T_0})$ (which is zero if there is perfect fit but not otherwise) as well as a noise component $v_{1,T_0+1} - \sum_{j=2}^N \beta_j^* v_{j,T_0+1}$, and both can contribute to serial correlation.

Example Imperfect Fit 2: (from Ferman and Pinto (2021)) Suppose that $Y_{it}(0) = c_i + \Lambda_i' F_t + \epsilon_{it}$ and one estimates $\widehat{\beta} = \text{argmin}_b \|Y_1(0) - X_0 b)\|_2^2$ where $X_0 = (Y_2, \ldots, Y_N)$. With $\beta^* = \text{plim}\widehat{\beta}$, the population imputation error is

$$e_{1t} = Y_{1t}(0) - X_{0t}'\beta^* = \Big(c_1 - \sum_{j=2}^N \beta_j^* c_j\Big) + F_t'\Big(\lambda_1 - \sum_{j=2}^N \beta_j^* \lambda_j\Big) + \Big(\epsilon_{1t} - \sum_{j=2}^N \beta_j^* \epsilon_{jt}\Big).$$

The first two terms vanish only if $\sigma_\epsilon \equiv \text{var}(\epsilon_{it}) = 0$; otherwise, $(c_1, \lambda_1) \neq (\sum_{j=2}^N \beta_j^* c_j, \sum_{j=2}^N \beta_j^* \lambda_j)$. Ferman and Pinto (2021) suggest to remove the bias with a mean adjustment but this may not remove serial or mutual correlation in the errors.

In the above examples, e_{it} absorbs all sorts of deficiencies in m_{it} and thus contains information about $Y_{it}(0)$. Cross-section, spatial, and time dependence in e_{it} are examples of non-spherical errors.

3 Prediction with Non-Spherical Errors

This section uses classical results in linear prediction to motivate how information in the errors can be used to improve prediction. We will consider optimal linear prediction in a time series setting so that the i subscript can be dropped.

3.1 Goldberger's BLUP

This subsection summarizes results for best linear unbiased prediction, BLUP. The concept seems to be first considered in Henderson (1950) in the animal breeding literature to predict the quality of offsprings. It is still widely used in estimation of random effects in linear mixed models for cross-section data.[3] Goldberger (1962) formalizes the idea in a setting where the linear model for predicting a scalar variable y_t is given by

$$y_t = X_t'\beta + e_t \tag{1}$$

where X_t is a $K \times 1$ vector of completely observed predictors assumed to be fixed in repeated samples, β is a vector of time invariant parameters, e_t is a zero mean stationary process that is possibly serially correlated, and Ω is the $T_0 \times T_0$ covariance matrix of the $T_0 \times 1$ vector e.

[3]Robinson (1991) provides a survey of its many derivations, including a Kalman filter interpretation, see also Spall (1991). Taub (1979) and Baltagi (2008, 2013) use it in variance components analysis of panel data.

Goldberger (1962) is interested in a linear unbiased prediction of y_m at some $m > T_0$ given information up to T_0 when Ω is positive definite but has non-zero off-diagonal entries. Let X be at $T_0 \times K$ matrix of regressors. The assumption of squared loss $E[(y_m - y_{m|T_0})^2]$ implies a linear predictor of the form $y_{m|T_0} = A'y$ with prediction error $y_{m|T_0} - y_m = (A'X - X'_m)\beta + A'e - e_m$. The unbiasedness constraint $E[y_{m|T_0}] = y_m$ requires that $A'X - X'_m = 0$, implying a prediction variance of

$$\begin{aligned} \sigma_m^2 &= E[(y_{m|T_0} - y_m)^2] = E[A'ee'A + e_m^2 - 2A'e_m e] \\ &= A'\Omega A + E\left(e_m^2\right) - 2A'\omega \end{aligned}$$

where $\omega \equiv E[e_m e]$. Let λ be the Lagrange multiplier on the unbiasedness constraint. Minimizing $A'\Omega A - 2A'\omega - 2\lambda'(X'A - X_m)$ with respect to A gives the best linear unbiased prediction (BLUP)

$$y^*_{m|T_0} = x'_m\beta_{GLS} + \omega'\Omega^{-1}e_{GLS},$$

where $\beta_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$ is the infeasible GLS estimator, and $e_{GLS} = y - X\beta_{GLS}$ is a $T_0 \times 1$ vector of errors. Notably, $y^*_{m|T_0}$ depends on assumptions about Ω, and in a time series setting, this depends on the dynamics of e_t. If $e_t = \phi_1 e_{t-1} + v_t$, where $v_t \overset{d}{\sim} (0, \sigma_v^2)$, Ω is $\sigma_e^2 = \sigma_v^2/(1-\phi_1^2)$ times a $T_0 \times T_0$ Toeplitz matrix with ϕ_1^{i-1} on the i-th diagonal. Then $\omega \equiv E[e_m e] = \phi_1^{-T_0+m}\Omega_{T_0}$, where Ω_{T_0} is the last column of Ω. The AR(1) assumption implies a BLUP at $m = T_0 + h$ of

$$y^*_{T_0+h|T_0} = X'_{T_0+h}\beta_{GLS} + \phi_1^h e_{GLS,T_0}$$

with prediction error $e^*_{T_0+h|T_0} = y_{T_0+h} - y^*_{T_0+h|T_0} = e_{T_0+h} - \phi_1^h e_{T_0} + o_p(1) = v_{T_0+h} + o_p(1)$ where the $o_p(1)$ term converges to 0 as $T_0 \to \infty$.

BLUP is infeasible because ϕ_1 is not observed. Feasible BLUP requires iterative Cochrane-Orcutt or Prais-Winsten estimation of ϕ_1, or direct estimation of β and ϕ_1 from a Durbin equation.[4] These feasible estimators are all efficient and consistent. Cochrane and Orcutt (1949) suggest to improve the standard prediction by incorporating lags of regressors and the dependent variable. The difference is that $\widehat{e}_{GLS,t}$ summarizes the dynamic relation between y and X into a single signal and can be appealing when X is of high dimension.

A feasible BLUP differs from the OLS prediction in two ways. First, it uses $\widehat{\beta}_{GLS}$ instead of $\widehat{\beta}_{OLS}$ and thus requires the dynamics of e_t to be specified. Second, BLUP adds to the GLS prediction a term that adjusts for serial correlation in e which in this AR(1) example is $\phi_1 e_{T_0}$

[4] Cochrane-Orcutt performs least squares regression of $y_t - \phi_1 y_{t-1}$ on $X_t - \phi_1 X_{t-1}$ for given ϕ_1 using data from $t = 2, \ldots, T_0$, and then estimates ϕ_1 from an autoregression in $y_t - X'_t\widehat{\beta}$ till convergence. The Prais-Winsten estimator additionally exploits information in $t = 1$. It is also possible to estimate β and ϕ_1 directly from the Durbin equation $y_t = X'_t\beta + y_{t-1}\phi_1 + X'_{t-1}\gamma$+error.

for $h = 1$. Since BLUP is an optimal prediction, it is more efficient than an OLS prediction. To make this point precise, consider again the AR(1) case. At $h = 1$, feasible BLUP

$$\widehat{y}^*_{T_0+1|T_0} = X'_{T_0+1}\widehat{\beta}_{GLS} + \widehat{\phi}_1 \widehat{e}_{GLS,T_0}$$

has prediction error $\widehat{e}^*_{T_0+1|T_0} = y_{T_0+1} - \widehat{y}^*_{T_0+1|T_0}$, or

$$\begin{aligned} \widehat{e}^*_{T_0+1|T_0} &= v_{T_0+1} - (X_{T_0+1} - \phi_1 X_{T_0})' (\widehat{\beta}_{GLS} - \beta) - \left(\widehat{\phi}_{1,GLS} - \phi_1\right) \widehat{e}_{GLS,T_0} \\ &= v_{T_0+1} + o_p(1) \end{aligned} \tag{2}$$

where the $o_p(1)$ term comes from the fact that the jointly estimated $\widehat{\beta}_{GLS}$ and $\widehat{\phi}_{1,GLS}$ are $\sqrt{T_0}$ consistent for β and ϕ_1. Since $\widehat{e}^*_{T_0+1|T_0}$ is asymptotically v_{T_0+1} whose variance is σ^2_v, feasible BLUP achieves the same asymptotic efficiency as infeasible BLUP.

In contrast, the OLS prediction error is

$$\begin{aligned} \widehat{e}_{T_0+1|T_0} &= e_{T_0+1} - X'_{T_0+1}(\widehat{\beta}_{OLS} - \beta) \\ &= e_{T_0+1} + o_p(1) \end{aligned} \tag{3}$$

where the $o_p(1)$ term comes from $\sqrt{T_0}$ consistency of $\widehat{\beta}_{OLS}$ for β. But $\widehat{e}_{T_0+1|T_0}$ is asymptotically e_{T_0+1} whose variance is $\sigma^2_e \geq \sigma^2_v$. Thus, the MSE improvement of BLUP over OLS is due to the additional term $\phi_1 e_{T,GLS}$ in the prediction, not because of GLS versus OLS estimation per se. Building on this idea, we will consider a linear prediction that is also asymptotically unbiased but can improve upon the OLS prediction without a priori knowledge of the precise dynamic structure of e_t.

3.2 From BLUP to PLUP

This subsection suggests a practical variant (PLUP) and studies its mean-squared error (MSE) using asymptotic expansions, first for $h = 1$, and then for $h > 1$ when direct and iterative forecasts are possible.

Our point of departure is that any predictor that controls for serial correlation will have the same first order effect as feasible BLUP. Let $\widehat{\beta}$ denote the least squares estimate of β. Consider modifying the (standard) least-squares prediction $\widehat{y}_{T_0+1|T_0} = X'_{T_0+1}\widehat{\beta}$ as follows:

$$\widehat{y}^+_{T_0+1|T_0} = \widehat{y}_{T_0+1|T_0} + \widehat{\rho}_1 \widehat{e}_{T_0} \tag{4}$$

where $\widehat{e}_{T_0} = y_{T_0} - X'_{T_0}\widehat{\beta}$ is the OLS residual, and

$$\widehat{\rho}_1 = \frac{\sum_{t=1}^{T_0} \widehat{e}_{t-1}\widehat{e}_t}{\sum_{t=1}^{T_0} \widehat{e}^2_{t-1}} \tag{5}$$

is the least squares estimate of the first order autocorrelation coefficient of $\widehat{e}_t$. Note that unlike feasible BLUP which re-estimates β after $\widehat{\rho}_1$ is available, we simply adjust the OLS prediction $\widehat{y}_{T_0+1|T_0}$ for serial correlation with $\widehat{\rho}_1\widehat{e}_{T_0}$. The prediction error $\widehat{e}^+_{T_0+1|T_0} = y_{T_0+1} - \widehat{y}^+_{T_0+1|T_0} = y_{T_0+1} - \widehat{y}_{T_0+1|T_0} - \widehat{\rho}_1\widehat{e}_{T_0}$ is

$$\begin{aligned} \widehat{e}^+_{T_0+1|T_0} &= \widehat{e}_{T_0+1|T_0} - \widehat{\rho}_1\widehat{e}_{T_0} \\ &= e_{T_0+1} - \rho_1 e_{T_0} + o_p(1), \end{aligned}$$

where the last equality follows because $\widehat{\beta}\xrightarrow{p}\beta$ and $\widehat{\rho}_1\xrightarrow{p}\rho_1$. If e_t is indeed an AR(1) model, then $\rho_1 = \phi_1$ and

$$\widehat{e}^+_{T_0+1|T_0} = v_{T_0+1} + o_p(1),$$

which is asymptotically equal to v_{T_0+1}, the prediction error of the infeasible BLUP.

As it turns out, adding the term $\widehat{\rho}_1\widehat{e}_{T_0}$ to any consistent prediction $\widehat{y}_{T_0+1|T_0}$ will yield an efficiency gain even when the true model is not an AR(1). We will refer to the prediction

$$\widehat{y}^+_{T_0+1|T_0} = X'_{T_0+1}\widehat{\beta} + \widehat{\rho}_1\widehat{e}_{T_0}$$

as practical BLUP (or PLUP), practical because it does not require GLS estimation and it is asymptotically as efficient as BLUP. To formalize the properties of the PLUP error $\widehat{e}^+_{T_0+1|T_0} = y_{T_0+1} - \widehat{y}^+_{T_0+1|T_0}$, we assume the following.

Assumption A1

(a) $E\left|e_t\right|^r < \infty$ for some $r > 2$, for all t.

(b) $\{e_t\}$ is a zero mean strictly stationary strong mixing process with mixing coefficients $\alpha(k) = O\left(k^{-\frac{r}{r-2}-\delta}\right)$ for some $\delta > 0$.

We define the strong mixing coefficients as $\alpha(k) = \sup_{A,B}\left|P(A\cap B) - P(A)P(B)\right|$ where A and B vary over events in the sigma fields generated by $\{e_s : s \le 0\}$ and $\{e_s : s \ge k\}$, respectively. Assumption A1 includes linear processes $e_t = \sum_{j=0}^{\infty}\psi_j v_{t-j}$, where $\sum_{j=0}^{\infty}|\psi_j| < \infty$ and v_t is i.i.d.$(0, \sigma_v^2)$ with $E\left|v_t\right|^r < \infty$, which includes stationary invertible ARMA(p,q) processes and nonlinear weakly dependent processes with GARCH and ARCH innovations.

Lemma 1 *($h = 1$): Let X_t be predictors and e_t be the errors in the model defined by (1). Suppose that $\{e_t\}$ satisfies Assumption A1 and that for $j = 0, 1$, $E(X_t e_{t-j})$, $E(X_{t-j}e_t)$ and $E\left(X_t X'_{t-j}\right)$ exist such that $\widehat{\beta}\xrightarrow{p}\beta$. Then as $T_0 \to \infty$,* *(i)* $\widehat{\rho}_1 \xrightarrow{p} \rho_1 \equiv \frac{\gamma_1}{\gamma_0}$, *where* $\gamma_k \equiv E\left(e_t e_{t-k}\right)$ *for all* k;

(ii) Standard prediction error: $\widehat{e}_{T_0+1|T_0} = e_{T_0+1} + o_p(1)$ *where* $e_{T_0+1} \overset{d}{\sim} (0, \gamma_0)$;

(iii) PLUP *error:* $\widehat{e}^{+}_{T_0+1|T_0} = e_{T_0+1} - \rho_1 e_{T_0} + o_p(1)$ *where* $e_{T_0+1} - \rho_1 e_{T_0} \overset{d}{\sim} (0, \gamma_0(1-\rho_1^2))$.

Part (i) shows that $\widehat{\rho}_1$ converges to the first order autocorrelation coefficient of e_t. Parts (ii) and (iii) describe the asymptotic expansion of the prediction errors ignoring the estimation error uncertainty. Part (ii) implies that the standard prediction is asymptotically unconditionally unbiased in spite of not accounting for serial correlation because e_{T_0+1} is mean zero by assumption, and it has asymptotic variance $\gamma_0 \equiv \text{var}\left(e_{T_0+1}\right) = \sigma_e^2$. The PLUP error in (iii) also has an unconditional mean of zero, but its variance is $\gamma_0(1-\rho_1^2)$. Since $|\rho_1| \leq 1$,

$$\gamma_0(1-\rho_1^2) \leq \gamma_0,$$

implying that the PLUP mean-squared dominates the standard prediction. If e_t is truly generated as $e_t = \phi_1 e_{t-1} + v_t$, the PLUP error $\widehat{e}^{+}_{T_0+1|T_0}$ will be asymptotically serially uncorrelated since $\rho_1 = \phi_1$.

However, an AR(1) correction will improve upon the standard prediction even when e_t is not an AR(1), provided that e_t is a mixing process satisfying A1. For instance, if e_t is an AR(2) defined by $e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + v_t$, then $\rho_1 = \frac{\phi_1}{1-\phi_2} \neq \phi_1$. But it will still be the case that $\widehat{\rho}_1 \xrightarrow{p} \rho_1$ as stated in (i). The PLUP error is now $\widehat{e}^{+}_{T_0+1|T_0} = e_{T_0+1} - \rho_1 e_{T_0} + o_p(1) = (\phi_1 - \rho_1)e_{T_0-1} + \phi_2 e_{T_0-2} + o_p(1)$, while the standard prediction error is $\widehat{e}_{T_0+1|T_0} = \phi_1 e_{T_0-1} + \phi_2 e_{T_0-2} + o_p(1)$. Both have a mean of zero, implying that misspecifying the dynamics will not contribute to unconditional bias. Nonetheless, the PLUP variance is $(1-\rho_1^2)\sigma_e^2$, which is smaller than the standard prediction error variance of σ_e^2 since $|\rho_1| \leq 1$. Thus the AR(1) correction unambiguously reduces one-step asymptotic mean squared prediction error. In theory, an AR(p) correction with known parameters should improve prediction when e_t is an AR(p). But in practice, sampling uncertainty may offset some gains. Furthermore, when the assumed AR(p) is not the true model, the mean-squared error is no longer tractable as shown in Kunitomo and Yamamoto (1985) even without sampling error. The AR(1) correction is appealing because it is simple to implement, and precise mean-squared error statements can be made when there is no sampling uncertainty as stated in Lemma 1.

Next, consider cases when $h > 1$. The standard prediction $\widehat{y}_{T_0+h|T_0} = X'_{T_0+h}\widehat{\beta}$ has error $\widehat{e}_{T_0+h|T_0} = y_{T_0+h} - \widehat{y}_{T_0+h|T_0} = e_{T_0+h} + o_p(1)$, where $E[e_{T_0+h}] = 0$ and $\text{var}(e_{T_0+h}) = \gamma_0$. There are two ways to implement PLUP. The first to use the AR(1) model for e_t to iteratively predict e_{T_0+h}. *Iterated* PLUP (or PLUPI), defined as

$$\widehat{y}^{+I}_{T_0+h|T_0} = X'_{T_0+h}\widehat{\beta} + \widehat{\rho}_1^h \widehat{e}_{T_0},$$

has error $\widehat{e}^{+I}_{T_0+h|T_0} = e_{T_0+h} - \rho_1^h e_{T_0} + o_p(1)$. As shown in the Appendix,

$$e_{T_0+h} - \rho_1^h e_{T_0} \overset{d}{\sim} (0, \gamma_0[1 + \rho_1^{2h} - 2\rho_1^h \rho_h]).$$

The second approach is to directly predict e_{T_0+h} using information up to T_0. Let $\widehat{\rho}_h = \widehat{\gamma}_h/\widehat{\gamma}_0$ be the h^{th} order sample autocorrelation coefficient of $\{\widehat{e}_t : t = 1, \ldots, T_0\}$. Direct PLUP (or PLUPD) is defined as

$$\widehat{y}^{+d}_{T_0+h|T_0} = X'_{T_0+h}\widehat{\beta} + \widehat{\rho}_h \widehat{e}_{T_0}.$$

PLUPD has error $\widehat{e}^{+d}_{T_0+h|T_0} = e_{T_0+h} - \rho_h e_{T_0} + o_p(1)$, where

$$e_{T_0+h} - \rho_h e_{T_0} \overset{d}{\sim} (0, \gamma_0 \left(1 - \rho_h^2\right)).$$

Lemma 2 *($h \geq 1$) Under the same assumptions as in Lemma 1, the asymptotic MSE of* PLUPD *is always smaller than or equal to that of* PLUPI *and that of the standard predictor for all $h \geq 1$.*

The proof is given in the Appendix. The standard predictor, PLUPI and PLUPD are all asymptotically unbiased provided that $E[e_t] = 0$. But

$$\text{var}(e_{T_0+h} - \rho_h e_{T_0}) - \text{var}\left(e_{T_0+h} - \rho_1^h e_{T_0}\right) = -\gamma_0 \left(\rho_h - \rho_1^h\right)^2 \leq 0$$

with equality when the AR(1) model is correctly specified. Furthermore,

$$\text{var}(e_{T_0+h} - \rho_h e_{T_0}) - \text{var}(e_{T_0+h}) = -\rho_h^2 \gamma_0 \leq 0$$

with equality when $\rho_h = 0$. Hence, absent sampling uncertainty, the asymptotic MSE of PLUPD using e_{T_0} to improve the standard prediction can be no larger than PLUPI which uses the same information for correction, or the standard predictor which ignores e_{T_0} for any $h \geq 1$. This is not to say that a richer dynamic model would not produce further improvements. What's noteworthy is that even a simple correction will reduce the prediction MSE at any h. We can also expect the PLUPD gains to be largest at $h = 1$ and diminish with h because the long horizon forecast of a covariance stationary process is the unconditional mean. Though precise statements can be made for PLUPD, we can only say that the asymptotic MSE of PLUPI is smaller or equal than that of the standard predictor if $1 + \rho_1^{2h} - 2\rho_1^h \rho_h \leq 1$ (see the Appendix for a proof).

Once a prediction is made, we can construct prediction intervals. We will be studying PLUP based inference under normality in the context of causal inference. As we will see in Section 5, while the standard prediction is unconditionally unbiased, it is conditional biased and conditional inference will, in general, have the wrong coverage.

3.3 Simulations for Linear Predictions

This subsection evaluates the unconditional and conditional prediction bias, MSE, and coverage with and without PLUP correction in a single equation setting where by unconditional inference, we mean that e_{i,T_0} is random in repeated sampling, and by conditional inference, we mean that e_{i,T_0} is treated as fixed with respect to some conditioning information, as would be the case in practice.

In each of the 5000 replications, we first simulate $K = 2$ regressors and $y_t = X_t'\beta + e_t$ where for $T = 1, \ldots, 200$, $e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + v_t$. With $v_t \overset{d}{\sim} N(0, .05)$, the R^2 of the regression is about 2/3. In Case 1, e_t is an AR(1) with $\phi_1 = .8$, and in Case 2, e_t is an AR(2) with $(\phi_1, \phi_2) = (1.3, -.4)$. Table 1 reports four sets of errors in predicting y_{T_0+h}. The column labeled 'best' is the infeasible prediction when β, ϕ_1, ϕ_2, and e_{T_0} are known. The column labeled 'noadj' is also infeasible but unlike 'best', it does not take into account information in e_{T_0}. The column labeled 'ols' is the standard prediction $\widehat{y}_{T_0+h}$ using the least squares estimate $\widehat{\beta}$. The columns PLUPI and PLUPD are iterative and direct PLUP respectively. Both are based on a simple AR(1) correction, ie. even when the true DGP is AR(2). Note that they are identical when $h = 1$.

The top panel of Table 1 reports the unconditional bias and MSE for horizons $h = 1, 2, 5, 10$. The average over all 10 horizons is reported in the row labeled 'avg'. Since $E[e_{T_0}] = 0$, the unconditional prediction bias is close to zero. However, the unconditional prediction MSE is much smaller with PLUP corrections. In the AR(1) case, the MSE for the standard (OLS) prediction is 0.14 at $h = 1$, but the PLUP corrections reduce the MSE to 0.05. In the AR(2) case, the OLS prediction has an MSE of 0.43 while the PLUP corrections reduce the MSE to 0.06. The MSE improvements are smaller when $h > 1$, but still non-trivial.

The middle panel of Table 1 shows conditional prediction errors when (e_{T_0-1}, e_{T_0}) are fixed to $(0.5, 1)$. All predictions are conditionally biased, but the PLUP biases are significantly smaller. When $h = 1$ and the errors are AR(1), the standard prediction has a conditional bias of 0.79 while the PLUP corrections reduce it to 0.01. When the errors are an AR(2) process but an AR(1) correction is implemented, the conditional OLS bias at $h = 1$ is reduced from 1.09 to 0.18. Correspondingly, the MSE is reduced from 1.26 to 0.09. Note that the biases are largest when $h = 1$ because the predictability of a stationary ergodic process decreases with the forecast horizon. The improvements in MSE at $h = 1$ translate into improved average predictions over 10 periods. Without the corrections, the average prediction in the AR(2) case has a bias of 0.61 and an MSE of 0.56. The AR(1) PLUP direct correction reduces bias to 0.17 and MSE to 0.22.

4 Imputation of Counterfactual Outcomes

Imputation concerns prediction of values that are never observed. The problem is widely studied in a static setup, but there are few results for a dynamic setting. Little and Rubin (2019, Ch. 11) consider an AR(1) model where $y_1, y_3, \ldots, y_{T-1}$ are observed but not $y_2, y_4, \ldots, y_T$. The adjustments, shown to require an implicit regression of y_t on y_{t-1} and y_{t+1}, can be seen as smoothed estimates of a suitably defined Kalman filter. Chow and Lin (1971) consider missing values occurring between two releases of low frequency data and show that the best prediction involves a correction term that has a BLUP form. Ng and Scanlan (2024) consider factor-based imputation of weekly missing values of a scalar series occurring throughout the sample.

Causal inference concerns imputation of missing potential outcomes that tend to occur at the end of the sample. The problem is typically studied for an iid setting when it is natural to assume that the errors are uncorrelated[5]. As suggested in Section 2, correlation in the residuals cannot be ruled out. We will consider the imputation problem from the perspective of optimal prediction, with the goal of using the insights of BLUP to improve the imputation of $Y_{i,T_0+h}(0)$. We assume that $e_{it} = Y_{it}(0) - m_{it}$ are strong mixing processes and rule out non-stationary data. In addition, we impose the following high level assumption:

Assumptions A2: For $h \geq 1$, $\widehat{m}_{i,T_0+h} - m_{i,T_0+h} = o_p(1)$ and $T_0^{-1} \sum_{t=1}^{T_0} (\widehat{m}_{it} - m_{it})^2 = o_p(1)$.

Assumption A2 is verified in Chernozhukov, Wüthrich, and Zhu (2021) for estimators including synthetic control, matrix completion, factor-based methods. Given an asymptotically unbiased $\widehat{m}_{i,T_0+h}$ satisfying Assumptions A1 and A2, the estimated treatment effect

$$\widehat{\delta}_{i,T_0+h} = m_{i,T_0+h} + \delta_{i,T_0+h} + e_{i,T_0+h} - \widehat{m}_{i,T_0+h}$$

has error

$$\widehat{\delta}_{i,T_0+h} - \delta_{i,T_0+h} = (m_{i,T_0+h} - \widehat{m}_{i,T_0+h}) + e_{i,T_0+h}. \tag{6}$$

This error has two components: an in-sample estimation uncertainty component that depends on the estimator but vanishes as $T_0 \to \infty$, and an out-of-sample prediction component that depends on the choice of m_{it} and the information $\mathcal{H}$ used in the imputation.

[5]Brodersen, Gallusser, Koehler, Remy, and Scott (2015) consider state space estimation of the counterfactual outcomes in the presence of trends, but serial correlation in idiosyncratic shocks and/or the factors are not allowed. Carvalho, Masini, and Medeiros (2018); Masini and Medeiros (2021, 2022) consider causal inference in a high-dimensional setting when the data are persistent and possibly non-stationary.

In order to extend Goldberger's BLUP from a complete data setting to a potential outcomes setting, define the $n \times 1$ vector of (observed) control outcomes $\mathcal{Y}(0)$ by

$$\underbrace{\mathcal{Y}(0)}_{n\times 1} = \begin{pmatrix} Y_{1,1:T_0} \\ \vdots \\ Y_{N,1:T_0} \\ Y_{N_1+1,T_0+1:T} \\ \vdots \\ Y_{N,T_0+1:T} \end{pmatrix} \equiv \begin{pmatrix} \mathcal{Y}^{pre}(0)_{(NT_0\times 1)} \\ \\ \mathcal{Y}^{post}(0)_{(N-N_1)T_1\times 1} \end{pmatrix}$$

where $n = (NT - N_1T_1)$. Note that $\mathcal{Y}(0)$ includes not only the pre-intervention outcomes on all units $\mathcal{Y}^{pre}(0) \equiv (Y_{1,1:T_0} \dots Y_{N,1:T_0})'$ but also the post-intervention outcomes on the control units $\mathcal{Y}^{post}(0) \equiv (Y_{N_1+1,T_0+1:T} \dots Y_{N,T_0+1:T})'$.

Let $\mathcal{M}$ be the pseudo-conditional mean for $\mathcal{Y}(0)$ and $\mathcal{E}$ be the corresponding errors. The matrices $\mathcal{M}$ and $\mathcal{E}$ contain typical elements m_{it} and e_{it}, respectively. With this notation,

$$\begin{aligned} \mathcal{Y}(0) &= \mathcal{M} + \mathcal{E}, \qquad \mathcal{E} = \begin{pmatrix} \mathcal{E}^{pre} \\ \mathcal{E}^{post} \end{pmatrix} \overset{d}{\sim} (0, \Gamma) \\ \Gamma &= E[\mathcal{E}\mathcal{E}'] = \begin{pmatrix} \Gamma_{pre,pre} & \Gamma_{pre,post} \\ \Gamma_{post,pre} & \Gamma_{post,post} \end{pmatrix}. \end{aligned}$$

The $n \times n$ matrix Γ depends on $\Sigma = E[e_t e_t']$, which is the $N \times N$ covariance matrix of $e_t = (e'_{1:N_1,t}, e'_{N_1+1:N,t})'$. It also depends on $\Omega_i = E[e_i e_i']$, which is the time series covariance of unit i, and its dimension can be $T \times T$ or $T_0 \times T_0$ depending on whether i is treated.

Consider the linear case $\mathcal{M} = \mathcal{X}\beta$, where $\mathcal{X}$ contains observed predictors. Consider obtaining BLUP given information on $\mathcal{Y}(0)$ and $\mathcal{X}$. Re-doing Goldberger's problem gives the following:

Proposition 1 *Assume that treatment assignment is known, $\mathcal{Y}(0) = \mathcal{M} + \mathcal{E}$, where $\mathcal{M} = \mathcal{X}\beta$ where β is constant across i and t. Let Γ be the $n \times n$ variance-covariance of $\mathcal{E}$, where $n = NT - N_1T_1$. The* BLUP *of the counterfactual outcome for unit $i \in [1, N_1]$ is*

$$\mathcal{Y}^{+}_{i,T_0+h}(0) = X'_{i,T_0+h}\beta_{GLS} + \omega'_{ih}\Gamma^{-1}\mathcal{E}_{GLS}$$

where β_{GLS} is the vector of infeasible GLS estimates and $\mathcal{E}_{GLS}$ are the corresponding residuals, $\omega_{ih} = E[\mathcal{E}e_{i,T_0+h}]$ is $n \times 1$ vector of covariances between the unexplained errors in the vector of observed control outcomes and unit i's counterfactual outcome not explained by the model at $T_0 + h$.

Proposition 1 provides the individual level best linear unbiased prediction in a treatment effects setting. Two cases are of special interest.

Case 1: serial correlation only: If $E[e_{\ell t}e_{js}] = 0$ for $\ell \neq j$ and for all t, s, then for $i \in [1, N_1]$,

$$\omega_{ih}'\Gamma^{-1}\mathcal{E}_{GLS} = \theta_i' e_{GLS,i,1:T_0} \tag{7}$$

where $\theta_i = (E[e_{i,1:T_0}e_{i,1:T_0}'])^{-1}E[e_{i,1:T_0}e_{i,1:T_0+h}]$ is the $T_0 \times 1$ vector of coefficients from projecting e_{i,T_0+h} on $(e_{i,1}, \ldots, e_{i,T_0})$.

The result in (7) follows from the fact that when there is no cross-section dependence in the errors, then the correction for unit i only depends on Ω_i, the $T_0 \times T_0$ autocovariance structure of e_i. For instance, if $N_1 = 1$ and $i = 1$ is the treated unit,

$$\omega_{ih} = \begin{pmatrix} E[e_{i,1:T_0}e_{i,T_0+h}] \\ 0_{N_0 T\times 1} \end{pmatrix},$$

where $N_0 = N - N_1$, and the prediction simplifies to $m_{i,T_0+h} + \rho_i^h e_{GLS,i,T_0}$ if e_{it} is assumed to be an AR(1), where ρ_i is the first order autocorrelation coefficient of e_{it}. This coincides with Goldberger's correction reviewed in Section 3.

Case 2: cross-section correlation only: If $E[e_{\ell t}e_{js}] = 0$ for $t \neq s$ and for all ℓ, j, then for $i \in [1, N_1]$,

$$\omega_{ih}'\Gamma^{-1}\mathcal{E}_{GLS} = \sum_{j=1}^{N-N_1} \theta_{i,N_1+j}e_{GLS,N_1+j,T_0+h}. \tag{8}$$

where θ_{i,N_1+j} are the slope coefficients from projecting e_{it} on $e_{N_1+1,t}, \ldots, e_{N,t}$ using $t = 1, \ldots, T_0$.

BLUP corrections with cross-section dependence have been derived in a static variance components setting but not in our set up. The result in (8), which is new, is based on two features that follow from no serial correlation. First, for any treated unit $i \in [1, N_1]$,

$$\omega_{ih} = \begin{pmatrix} E(\mathcal{E}^{pre}e_{i,T_0+h}) \\ E(\mathcal{E}^{post}e_{i,T_0+h}) \end{pmatrix} = \begin{pmatrix} 0_{NT_0\times 1} \\ E(\mathcal{E}^{post}e_{i,T_0+h}) \end{pmatrix}.$$

Second, the covariance matrix of errors Γ has a block diagonal structure

$$\Gamma = \begin{pmatrix} \Sigma \otimes I_{T_0} & 0 \\ 0 & \Sigma_{00} \otimes I_{T_1} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{10} \\ \Sigma_{10}' & \Sigma_{00} \end{pmatrix},$$

where $\Sigma_{11} = E(e_{1:N_1,t}e_{1:N_1,t}')$, $\Sigma_{00} = E(e_{N_1+1:N,t}e_{N_1+1:N,t}')$, and $\Sigma_{10} = E(e_{1:N_1,t}e_{N_1+1:N,t}')$. Let $[\Sigma_{10}]_{i,.}$ be the i-th row of the matrix Σ_{10}. Then, for any $i \in [1, N_1]$ and $h > 0$, the non-zero entries of the $n \times 1$ vector ω_{ih} are given by

$$E(\mathcal{E}^{post}e_{i,T_0+h}) = [\Sigma_{01}]_{i.}' \otimes J_h$$

where J_h is the h-th column of the identity matrix of dimension T_1. As a consequence of the two features, the BLUP correction is

$$\omega'_{ih}\Gamma^{-1}\mathcal{E}_{GLS} = ([\Sigma_{10}]_{i,.}\Sigma_{00}^{-1} \otimes J'_h)\mathcal{E}_{GLS}^{post}.$$

It is a linear combination of $\mathcal{E}_{GLS}^{post}$, the GLS errors of the control units in the post-treatment sample, with weights given by the $N - N_1$ vector $\theta'_i = (\theta_{i,N_1+1}, \ldots, \theta_{i,N}) = \Sigma_{00}^{-1}[\Sigma_{10}]'_{i,.}$, where θ_{i,N_1+j} is the population coefficient associated with $e_{N_1+j,t}$ in the regression of e_{it} on $e_{N_1+1:N,t} = (e_{N_1+1,t}, \ldots, e_{N,t})'$. Given the definition of J_h, we can write the BLUP correction for unit i in period $T_0 + h$ as $\omega'_{ih}\Gamma^{-1}\mathcal{E}_{GLS} = \theta'_i e_{GLS,N_1+1:N,T_0+h}$, as given in the Proposition. The θ_i vector can be sparse such as in factor models when the idiosyncratic errors can only be weakly dependent in the sense that if $E(e_{it}e_{jt}) = \tau_{ij,t}$, $|\tau_{ij,t}| \leq |\bar{\tau}_{ij}|$ for some $\bar{\tau}_{ij}$ for all t, and $\sum_{j=1}^{N} |\bar{\tau}_{ij}| \leq M \leq \infty$ for all i.

4.1 From blup to pup

In Section 3, we take as a starting point that the first order improvement of BLUP comes from controlling for the predictability in e_{it}. While BLUP is developed for linear predictions, linearity is not necessary to obtain an improved predictor. Our *practical unbiased predictor* (henceforth, PUP) can be used with any choice of $\mathcal{M}$ that can be consistently estimated so that the residuals $\widehat{\mathcal{E}} = \mathcal{Y}(0) - \widehat{\mathcal{M}}$ can be used to improve prediction. In practice, implementation still requires parametric assumptions on Ω and Σ. For a serially correlated process e_{it} satisfying mixing conditions, we have the following

Lemma 3 *Under Assumptions A1 and A2, a standard imputation has error* $\widehat{\delta}_{i,T_0+1} - \delta_{i,T_0+1} = e_{i,T_0+1} + o_p(1)$*, whose variance is* $\sigma_{e,i}^2 = \gamma_{0,i}$*. A* PUP *correction has error* $\widehat{\delta}_{i,T_0+1}^{+} - \delta_{i,T_0+1} = e_{i,T_0+1} - \rho_{i,1}e_{i,T_0} + o_p(1)$*, whose variance is* $(\sigma_{e,i,1}^{+})^2 = \gamma_{0,i}(1 - \rho_{i,1}^2)$*, where* $\rho_{i,1}$ *is the first-order autocorrelation of* e_{it}*. Since* $|\rho_{i,1}| \leq 1$*,* $(\sigma_{e,i,1}^{+})^2 \leq \sigma_{e,i}^2$*.*

Lemma 3 follows immediately from Lemmas 1 and 2 but the results are presented in a treatment effect setting where the $o_p(1)$ term vanishes with T_0. For $h = 1$, the MSE of a PUP imputation will always be smaller than that of a standard imputation. We focus on $h = 1$ since the gain in MSE should be largest, and furthermore, it is also the case when the direct and iterative corrections coincide. In theory, a direct PUP using $\rho_{i,h}$ has better properties than an iterative PUP using $\rho_{i,1}^{h}$ when $h > 1$. But in simulations when sampling uncertainty is present, the two behave similarly and both dominate the standard predictor.

It is also possible to entertain both time and cross-section dependence in e_{it}. For example,

$$\widehat{Y}^+_{i,T_0+h}(0) = \widehat{Y}_{i,T_0+h}(0) + \sum_{s=0}^{p_i} \widehat{\rho}_{is}\widehat{e}_{i,T_0-s} + \sum_{j\neq i}^{N} \sum_{s=-p_j}^{p_j} \widehat{\theta}_{ij,s}\widehat{e}_{j,T_0+h-s}. \tag{9}$$

The first correction captures the time series information from $\widehat{e}_i$'s own history, while the second correction captures cross-section dependence using estimates of the current and past idiosyncratic errors of the control units.

Though optimal prediction of counterfactual outcomes has not been studied, PUP like corrections have recently been considered. Chernozhukov, Wüthrich, and Zhu (2021) consider time dependence in e_{1t} and suggest adding to the standard imputed value an AR(p) estimate of the residuals, but the idea is not flushed out. Fan, Masini, and Medeiros (2022) consider a factor model with observables W_{it} and assume contemporaneously correlated idiosyncratic errors $e_{it} = \theta_i' e_{-i,t} + v_{it}$, where $e_{-i,t}$ is a $N-1$ vector that excludes e_{it}. This correlation can be controlled by $\widehat{e}_{jt}$ from estimation of the factor model. To circumvent overfitting the augmented prediction model $Y_{it}(0) = \gamma_i' W_{it} + \lambda_i' F_t + \theta_1' \widehat{e}_{-i,t} + v_{it}$ when N_0 is large, LASSO is used to select which of the $\widehat{e}_{j,t}$ to keep. Our analysis provides a framework for thinking about these PUP like modifications.

4.2 Simulations: $\widehat{\delta}_{it}$ vs $\widehat{\delta}^+_{it}$

In Table 1, we saw in a linear prediction setting that PLUP yields significant reductions in bias and mean-squared error. Since imputation is a form of prediction, we can anticipate improvements from using PUP in imputation settings. To illustrate this, we generate $Y_{it}(0)$ using a factor model:

$$\begin{aligned} Y_{it}(0) &= c_i + \Lambda_i' F_t + e_{it}, \quad t = 1, \ldots, T \\ Y_{1t}(1) &= Y_{1t}(0) + \delta_{1t}, \qquad \delta_{1t} = 0.1 \quad t = T_0 + 1, \ldots, T. \end{aligned}$$

We assume that unit 1 is treated with $\delta_{1t} = 0.1$ and $e_{it} = \phi_i e_{i,t-1} + v_{it}$, $v_{it} \overset{d}{\sim} N(0, .25)$, $\phi_1 = 0.6$, $\phi_i = 0$ for $i > 1$. We set $c_i = 0$ for all i, $r = 2$, $F_{1t} = .8F_{1t-1} + e^F_{1t}$ and $F_{2t} = .5F_{2,t-1} + e^F_{2t}$, $\Lambda_{ik} \overset{d}{\sim} N(0,1)$, $e^F_{1t} \overset{d}{\sim} N(0, .5)$, $e^F_{2t} \overset{d}{\sim} N(0, .3)$. The 'best' prediction is $Y_{1,T_0+h} = \Lambda_1' F_{T_0+h} + \phi_1^h e_{1,T_0}$ and the standard prediction is based on the principal components $\widehat{\Lambda}_i' \widehat{F}_{T_0+h}$ of the demeaned data for the control group, which will be denoted PCA.

Table 2 reports the error in imputing $\widehat{\delta}_{1,T_0+h}$ for $(T, N) = (50, 20)$. Because $Y_{1t}(1)$ is observed, $\widehat{\delta}_{1t} - \delta_{1t} = Y_{1t}(1) - \widehat{Y}_{1t}(0)$ is due entirely to $\widehat{Y}_{1t}(0)$. The top panel considers $e_{1t} = 0.6e_{1t-1} + v_{1t}$. We see that though PCA is unconditionally unbiased, its MSEs are larger compared to the PUP ones. Furthermore, the conditional biases upon fixing $e_{1,T_0} = 1.0$ are smaller with PUP, as is MSE. For mutually correlated errors when $e_{1t} = 0.5e_{2t} + v_{1t}$, the conditional case randomly draws $e_{2,T_0+1:T_1}$ once and keeps the vector fixed in the replications. Again, the two imputations based on PUP reduce MSE unconditionally and conditionally. Unlike in the time series case, the improvements can occur at any horizon h. Though direct PUP has slightly better population properties than iterative PUP , they have rather similar properties in simulations. To simplify notation, we use the term PUP in the next subsection, with the understanding that either direct or iterated correlation can be used.

5 Inference Based on Imputed Counterfactual Outcomes

Let $F(x) = P(e_{it} \leq x)$ be the marginal distribution function of e_{it}, and $F^+(x)$ be the marginal distribution of an adjusted PUP error. Lemma 3 suggests that

$$\begin{aligned}
\widehat{\delta}_{i,T_0+1} - \delta_{i,T_0+1} &\xrightarrow{d} e_{i,T_0+1} \overset{d}{\sim} F \\
\widehat{\delta}^{+}_{i,T_0+1} - \delta_{i,T_0+1} &\xrightarrow{d} e_{i,T_0+1} - \rho_{i,1} e_{i,T_0} \overset{d}{\sim} F^{+}.
\end{aligned}$$

Though both F and F^+ are centered at zero, F is more dispersed than F^+. There is surprisingly little work on inference based on a feasible BLUP even in a single equation setting presumably because further assumptions on F are needed. We will assume that $\{e_{it}\}$ is a Gaussian process with autocovariance at lag j of $\gamma_{j,i}$, and let $z_{1-\alpha}$ be such that $\Phi(z_{1-\alpha}) = 1 - \alpha$, where Φ is the cdf of the standard normal distribution. The Gaussian assumption is only made to illustrate the issues created by omitting predictability. Other distributions can be used in its place so long as e_{it} satisfies our mixing assumption.

We consider intervals of the form $\widehat{\delta} \pm \widehat{\sigma}_{\delta} z_{1-\alpha/2}$ where $\widehat{\delta}$ is either the standard predictor, or a PUP predictor with variance defined in Lemma 3. These intervals, denoted PI_{ih}, and PI^{+}_{ih}, can be based on asymptotic theory or resampling methods as in Li (2020) and Cattaneo, Feng, and Titiunik (2021), among others. The intervals will be used for both unconditional and conditional inference. By unconditional inference, we mean that e_{i,T_0} is random in repeated sampling. By conditional inference, we mean that e_{i,T_0} is treated as fixed with respect to some conditioning information, as would be the case in practice. Phillips (1979) notes that while unconditional inference is useful for evaluating econometric methods, conditional inference has a role in applications.

We begin with unconditional inference. We say that a prediction interval is *unconditionally* asymptotically valid if it contains δ_{i,T_0+h} with probability $1 - \alpha$ as $T_0 \to \infty$. We will focus on pointwise results for unit i where $e_{i,T_0+h} \equiv Y_{i,T_0+h}(0) - m_{i,T_0+h}$. It is easy to show that under the assumptions of Lemma 3 and assuming that $\{e_{it}\}$ is Gaussian, all three intervals are asymptotically valid unconditionally. Consider for instance PI_{ih}, an interval for the standard prediction error for unit i. Since $\widehat{\delta}_{i,T_0+h} - \delta_{i,T_0+h} = e_{i,T_0+h} + o_p(1)$ and $e_{i,T_0+h} \overset{d}{\sim} N(0, \gamma_{0,i})$, where $\gamma_{0,i} \equiv \sigma^2_{e,i}$, we have

$$\begin{aligned}
P\left(\delta_{i,T_0+h} \in \text{PI}_{ih}\right) &= P\left(\widehat{\delta}_{i,T_0+h} - \widehat{\sigma}_{e,i} z_{1-\alpha/2} \leq \delta_{i,T_0+h} \leq \widehat{\delta}_{i,T_0+h} + \widehat{\sigma}_{e,i} z_{1-\alpha/2}\right) \\
&= P\left(-z_{1-\alpha/2} \leq \widehat{\sigma}^{-1}_{e,i}(\widehat{\delta}_{i,T_0+h} - \delta_{i,T_0+h}) \leq z_{1-\alpha/2}\right) \\
&= P\left(-z_{1-\alpha/2} \leq \sigma^{-1}_{e,i} e_{i,T_0+h} \leq z_{1-\alpha/2}\right) + o(1) \\
&= \Phi\left(z_{1-\alpha/2}\right) - \Phi\left(-z_{1-\alpha/2}\right) + o(1) \\
&= 1 - \alpha + o(1).
\end{aligned}$$

In the above, the third equality follows because $\widehat{\sigma}^{-1}_{e,i}(\widehat{\delta}_{i,T_0+h} - \delta_{i,T_0+h}) = \sigma^{-1}_{e,i} e_{i,T_0+h} + o_p(1)$ and the fourth equality uses the Gaussianity assumption on e_{it}. The argument for the PUP prediction intervals is similar and thus all three intervals are (unconditionally) asymptotically valid. In the absence of sampling uncertainty, PI^+ is asymptotically narrower than PI for all h whether or not e_{it} is truly an AR(1) process.

5.1 Conditional Inference

While the three intervals provide correct unconditional coverage, will they all have correct conditional coverage? In particular, will they cover δ_{i,T_0+h} with the nominal coverage probability of $1-\alpha$, conditionally on some information set $\mathcal{H}$? To answer this question, consider again the AR(1) case when $e_{it} = \phi_i e_{i,t-1} + v_{it}$, $v_{it} \overset{d}{\sim} N(0, \sigma^2_{v,i})$, and $\mathcal{H}$ is an information set containing e_{i,T_0}. Under the assumption of normality, $e_{i,T_0+1} \overset{d}{\sim} N(\phi_i e_{i,T_0}, \sigma^2_{v,i})$ conditionally on e_{i,T_0} and hence $\frac{e_{i,T_0+1} - \phi_i e_{i,T_0}}{\sigma_{v,i}} \overset{d}{\sim} N(0,1)$. But as $\frac{e_{i,T_0+1}}{\sigma_{e,i}} \overset{d}{\sim} N(\phi_i \frac{e_{i,T_0}}{\sigma_{e,i}}, \frac{\sigma^2_{v,i}}{\sigma^2_{e,i}}) \neq N(0,1)$, the standard prediction will not usually have the correct coverage unless ϕ_i or e_{i,T_0} are zero. Indeed, PI_{i1} will not have the correct conditional coverage probability even asymptotically because $P\left(\delta_{i,T_0+1} \in \text{PI}_{i1} | e_{i,T_0}\right)$ is asymptotically equal to

$$\Phi\left(-\frac{\phi_i}{\sigma_{v,i}} e_{i,T_0} + \frac{\sigma_{e,i}}{\sigma_{v,i}} z_{1-\alpha/2}\right) - \Phi\left(-\frac{\phi_i}{\sigma_{v,i}} e_{i,T_0} - \frac{\sigma_{e,i}}{\sigma_{v,i}} z_{1-\alpha/2}\right).$$

When $\phi_i \neq 0$, the first term will not return the normal cdf at level $1-\alpha/2$ unless e_{i,T_0} is zero.

The problem of distorted inference extends to a multi-period ahead conditional inference. For any $h > 1$, we see that

$$e_{i,T_0+h} = \phi^h_i e_{i,T_0} + u_{i,T_0+h} | e_{i,T_0} \overset{d}{\sim} N\left(\phi^h_i e_{i,T_0}, \omega^2_{h,i}\right),$$

where $u_{i,T_0+h} \overset{d}{\sim} (0, \omega^2_{h,i})$, $\omega^2_{h,i} = \gamma_{0,i}\left(1-\phi_i^{2h}\right)$,[6] with $\gamma_{0,i} \equiv \sigma^2_{e,i} = \sigma^2_{v,i}\left(1-\phi_i^2\right)^{-1}$. The problem arises because the standard prediction error is not centered at zero when we condition on e_{i,T_0}. Thus for any $h \geq 1$, $P(\delta_{i,T_0+h} \in \text{PI}_{ih} | e_{i,T_0})$ is asymptotically equal to

$$\Phi\left(-\frac{\phi_i^h}{\omega_{h,i}} e_{i,T_0} + z_{1-\alpha/2}\frac{\sigma_{e,i}}{\omega_{h,i}}\right) - \Phi\left(-\frac{\phi_i^h}{\omega_{h,i}} e_{i,T_0} - z_{1-\alpha/2}\frac{\sigma_{e,i}}{\omega_{h,i}}\right)$$

where $\omega^2_{h,i}$ is the h-period forecast error variance defined above. For fixed h, conditional inference is distorted unless $e_{i,T_0} = 0$, though the distortion decreases with h because ϕ_i^h tends to zero.

To illustrate the extent of conditional bias, consider two models for e_{it}: one when $e_{it} = \phi_i e_{i,t-1} + v_{it}$ is an AR(1), and one when $e_{it} = v_{it} + \theta_i v_{i,t-1}$ is an MA(1). In both cases, the conditional forecast is biased. Analytically evaluating actual coverage for any i for a nominal 95% interval, we have

Standard Prediction with $(e_{1,T_0}, \sigma_{v,1}) = (-2.0, 0.5)$

h	coverage	bias	coverage	bias
	AR(1): $\phi_1 = 0.8$		MA(1): $\theta_1 = 0.8$	
1	0.84	-2.26	0.58	-1.77
2	0.87	-1.41	0.95	-0.00
3	0.90	-1.01	0.95	-0.00
4	0.92	-0.76	0.95	-0.00
5	0.93	-0.59	0.95	-0.00

The coverage probability is 0.95 at $\phi_1 = \theta_1 = 0$ for all values of h because there is no conditional bias. When $\phi_1 \neq 0$ in the AR(1) case, bias decreases with $\sigma^2_{v,1}$ and h. When $\theta_1 \neq 0$ in the MA(1) case, the bias is limited to $h = 1$ because the process has a memory of one period. Coverage is distorted by bias, as suggested by theory.

In contrast, PUP has error $\widehat{\delta}^+_{i,T_0+h} - \delta_{i,T_0+h} = v_{i,T_0+h} + o_p(1)$. It is conditionally centered at zero with conditional variance $\sigma^2_{v,i}$. Let $\widehat{\sigma}^+_{i,h}$ denote a consistent estimator of this variance (which will depend on whether an iterative or a direct estimator is used). Under normality, the PUP prediction interval $\text{PI}^+_{ih} = \widehat{\delta}^+_{i,T_0+h} \pm \widehat{\sigma}^+_{i,h} z_{1-\alpha/2}$ has the correct coverage asymptotically. At $h = 1$ when iterative and direct PUP coincide, $P\left(\delta_{i,T_0+1} \in \text{PI}^+_{i1} | e_{i,T_0}\right)$ is asymptotically

$$\Phi\left(\frac{\widehat{\sigma}^+_{i,1}}{\sigma_{v,i}} z_{1-\alpha/2}\right) - \Phi\left(-\frac{\widehat{\sigma}^+_{i,1}}{\sigma_{v,i}} z_{1-\alpha/2}\right) = 1 - \alpha + o_p(1),$$

[6] This follows because $u_{i,T_0+h} = v_{i,T_0+h} + \ldots + \phi_i^{h-1} v_{i,T_0}$. Thus, $\omega^2_{h,i} = \sigma^2_{v,i}\sum_{j=0}^{h-1}\phi_i^{2h} = \sigma^2_{v,i}\frac{1-\phi_i^{2h}}{1-\phi_i^2} \equiv \gamma_{0,i}\left(1-\phi_i^{2h}\right)$, $\gamma_{0,i} = \sigma^2_{v,i}\left(1-\phi_i^2\right)^{-1}$.

where the last equality uses the fact that $\widehat{\sigma}^{+}_{i,1} \xrightarrow{p} \sigma_{v,i}$.

Analogous to conditional bias due to time dependence, a similar bias occurs when the errors are cross-sectionally correlated. In particular, suppose that e_{it} is serially uncorrelated for all i, but $E[e_{it}e_{jt}] \neq 0$ for at least one $j \neq i$. If $i = 1$ is the only treated unit, the model of interest is $Y_{1t}(0) = m_{1t} + e_{1t}$ with $e_{1t} = \theta_1' e_{2:N,t} + v_{1t}$. Consider two treatment effects estimators, one based on the standard prediction of $Y_{1t}(0)$ and another based on PUP . The standard prediction yields $\widehat{\delta}_{1,T_0+1} - \delta_{1,T_0+1} = e_{1,T_0+1} + o_p(1)$ which has asymptotic mean zero and asymptotic variance $\sigma^2_{e,1} \equiv E(e^2_{1t})$. Instead, the PUP prediction yields a prediction error $\widehat{\delta}^{+}_{1,T_0+1} - \delta_{1,T_0+1} = e_{1,T_0+1} - \theta_1' e_{2:N,T_0+1} + o_p(1)$, whose asymptotic variance is $\sigma^2_{v,1} = \sigma^2_{e,1} - \Sigma_{01}'\Sigma_{00}^{-1}\Sigma_{01} < \sigma^2_{e,1}$. With cross-sectionally correlated errors, the standard prediction is conditionally biased because $e_{1,T_0+1}|e_{2:N,T_0+1} \overset{d}{\sim} N(\theta_1' e_{2:N,T_0+1}, \sigma^2_{v,1})$ is not centered at zero. Conditional coverage under normality is asymptotically determined by

$$\Phi\bigg(-\frac{\theta_1'}{\sigma_{v,1}} e_{2:N,T_0+1} + z_{1-\alpha/2}\frac{\sigma_{e,1}}{\sigma_{v,1}}\bigg) - \Phi\bigg(-\frac{\theta_1'}{\sigma_{v,1}} e_{2:N,T_0+1} - z_{1-\alpha/2}\frac{\sigma_{e,1}}{\sigma_{v,1}}\bigg).$$

The conditional bias arising from $\theta_1 \neq 0$ and $\sigma_{e,1} \neq \sigma_{v,1}$ will distort inference. Notably, $-\frac{\theta_1'}{\sigma_{v,1}}$ here plays the role of $-\frac{\phi_1}{\sigma_{v,1}}$ in the AR(1) setting. But in contrast to the case of serial correlation, the size distortion does not diminish with h. The PUP prediction is more efficient because $e_{1t}|e_{2:N,t} \overset{d}{\sim} (\theta_1' e_{2:N,t}, \sigma^2_{v,1})$ which has a smaller variance.

To illustrate, suppose that $N_1 = 1$ and $e_{1t} = \theta_1 e_{2t} + v_{1t}$, where we assume that σ_{12} is the only non-zero cross sectional covariance. In this case, the relevant parameters are $[\Sigma_{00}]_{11} = \sigma^2_{e,2}$, and $[\Sigma_{01}]_{11} = \text{cov}(e_{1t}, e_{2t}) \equiv \sigma_{12}$. Unlike in the time series case when we only condition on e_{1,T_0}, we now condition on e_{2,T_0+h} for each h. In the following example, we draw e_{2,T_0+h} once from the normal distribution using the RNDN function in MATLAB with seed 1234, and $\text{cov}(e_{1t}, e_{2t})$ from the uniform distribution using the RAND function with seed 57.

Standard Prediction when $e_{1,T_0+h} = \theta_1 e_{2,T_0+h} + v_{1,T_0+h}$, e_{2,T_0+h} fixed
$\Sigma = \text{diag}(\sigma^2_{e,1}, \Sigma_{00}) = \text{diag}(0.5, 0.841)$.

		$\Sigma_{01} = 0.613$		$\Sigma_{01} = -0.613$	
h	$\bar{e}_{2,T_0+h}$	coverage	bias	coverage	bias
1	0.68	0.63	-1.64	0.89	-0.70
2	-0.83	0.81	1.07	0.86	0.85
3	-0.92	0.93	-0.38	0.84	0.95
4	0.09	0.95	0.17	0.95	-0.09
5	0.86	0.94	0.32	0.86	-0.89

The parameterizations result in $\theta_1 = \text{sgn} \times 0.72$ where the sign depends on whether $\Sigma_{01} \equiv \sigma_{12}$ is positive or negative. The sign affects the magnitude of the bias which in turn affects the extent of size distortion. However, unlike in the time series case when predictability falls with h by nature of stationarity, the cross-section correlation does not decrease with h. As a consequence, the effect on coverage can vary significantly across horizons.

5.2 Simulations: PI vs PI$^+$

Tables 1 and 2 above showed that PUP corrections reduce bias and mean-squared prediction error significantly. But do these improvements lead to more accurate inference? We first return to Table 1 when point prediction is based on the simple linear model is $y_t = X_t'\beta + e_t$. The corresponding results for coverage are reported in Table 1(b). Note first that there is no gain in unconditional coverage regardless of the error structure because $E[e_{it}] = 0$ by assumption. Hence we focus on conditional coverage. In the AR(1) case, coverage is 40% at $h = 1$ without PUP correction, but is at the desired level of 95% with correction. PUP coverage in the AR(2) case is 91%, which is less accurate, but still better than the OLS coverage of 0.69.

Evaluation of coverage based on δ_{it} is more involved because the sampling error depends on the estimator. Furthermore, given estimates of $\widehat{\delta}_{i,T_0+h}$ for some $i \in [1, N_1]$, any of the following hypothesis can be considered.

$$\begin{array}{llll}
H_0^A : \delta_{i,T_0+h} & = \delta^0_{i,T_0+h} & & \\
H_0^B : \delta_{i,T_0+h} & = 0 & \forall h & \\
H_0^C : \Delta_{i,T_0+1:T} & = \Delta^0_{i,T_0+1:T} & \Delta_{i,T_0+1:T} & = \frac{1}{T_1}\sum_{h=1}^{T_1} \delta_{i,T_0+h} \\
H_0^D : \Delta_{1:N_1,T_0+h} & = \Delta^0_{1:N_1,T_0+h}, & \Delta_{1:N_1,T_0+h} & = \frac{1}{N_1}\sum_{i=1}^{N_1} \delta_{i,T_0+h} \\
H_0^E : \Delta_{1:N_1,T_0+1:T} & = \Delta^0_{1:N_1,T_0+1:T_1}, & \Delta_{1:N_1,T_0+1:T} & = \frac{1}{T_1}\frac{1}{N_1}\sum_{h=1}^{T_1}\sum_{i=1}^{N_1} \delta_{i,T_0+h}
\end{array}$$

While A and B are pointwise hypotheses, hypotheses C,D,E concern the average treatment effect of the treated, where the average can be taken over time, over units, or both. The interpretation of each test depends on T_1 and N_1. Consider H_0^C. If T_1 is small so that $\Delta_{i,T_0+1:T} = \frac{1}{T_1}\sum_{t=T_0+1}^{T}\delta_{it}$ is random, we construct a *prediction* interval for $\Delta_{i,T_0+1:T}$. When T_1 is large, $\Delta_{i,T_0+1:T}\xrightarrow{p}\Delta_{i,\infty} = E[\delta_{it}]$ is non-random. In this case, we construct a *confidence* interval for $\Delta_{1,T_0+1:T}$. Considerations of N_1 are likewise needed for testing H_0^D.[7]

To give a flavor of the results, we only consider H_0^A and H_0^C using the same data generating process in Table 2. We estimate (F_t, Λ_i) using the TALL-WIDE procedure in Bai and Ng (2021). With this estimator, the asymptotic distribution of $\widehat{\delta}_{i,T_0+h}$ is determined by the distribution of e_{i,T_0+h}. Hence the distribution of $\widehat{\delta}_{i,T_0+h}$ is normal only if e_{it} is Gaussian. However, the treatment effect averaged over T_1 periods can be asymptotically normal with a convergence rate of $\min(T_0,T_1)$ if the CLT $\frac{1}{\sqrt{T_0}}\sum_{t=1}^{T_0} F_t e_{it}\xrightarrow{d}N(0,\Phi_i)$ holds. In contrast, the average treatment effect over N_1 units can be asymptotically normal with a convergence rate of $\min(N_0,N_1)$ if a CLT for $\frac{1}{\sqrt{N_0}}\sum_{i=1}^{N_0}\Lambda_i e_{it}\xrightarrow{d}N(0,\Gamma_t)$ holds. These distinctions will help understand the coverage results reported in Table 2(b).

Turning first to the time series case in the top panel of Table 2(b), PUP coverage for δ_{i,T_0+h} improves over PCA, but only for $h = 1, 2$, suggesting that for serially correlated errors, PUP is most effective when h is small, especially when T_0 is small. For cross-sectionally correlated errors reported in the bottom panel, improved coverage of PUP can occur at any h, but is more systematic when N_0 is large.

As explained above, the limiting distribution of the Bai and Ng (2021) estimator of $\widehat{\bar{\delta}}_i$ depends on adequacy of central limit theory, while it is normality of e_{it} that renders $\widehat{\delta}_{it}$ normally distributed. It is thus not surprising that coverage of δ_{it} does not provide a good guide to the coverage of $\bar{\delta}_i = \frac{1}{T_1}\sum_{h=1}^{T_1}\delta_{i,T_0+h}$. At $(T_0, N_0) = (50, 20)$, PUP conditional coverage of $\bar{\delta}_i$ is too low, though no more inaccurate than the standard PCA prediction. Whether the errors are serially or cross-sectionally correlated, coverage of $\bar{\delta}_i$ is more reliable as T_0 increases because we average over T_1 variables that become increasingly Gaussian. Sampling methods could be useful when T_0 is small, see e.g. Li, Shen, and Zhou (2023).

6 Conclusion

Goldberger (1962) shows that if the errors of the prediction model are non-spherical, they can be exploited to improve prediction. Motivated by this result, this paper has suggested PUP, a simple way to adjust existing estimates of counterfactual outcomes for dependence in the errors. The adjustment consists of adding a term that exploits the presence of serial or

[7] For large T_1, $\frac{1}{T_1}\sum_{h=1}^{T_1} m_{i,T_0+h} - \widehat{m}_{i,T_0+h} + \frac{1}{T_1}\sum_{h=1}^{T} e_{i,T_0+h} + \delta_{i,T_0+h} - E[\delta_{i,T_0+h}]$, which equals $\widehat{\Delta}_{i,T_0+1:T} - \Delta_{i,T_0+1:T} + \delta_{i,T_0+h} - E[\delta_{i,T_0+h}]$.

cross sectional correlation in the prediction errors of the model used to obtain the estimated counterfactual outcomes.

We showed that improved mean-squared errors are possible without knowledge of the true error structure, and simple corrections often suffice. We also showed that omitting the PUP adjustment term when the error is predictable can result in conditional bias, thus leading to prediction intervals that are not conditionally asymptotically valid. In contrast, a prediction interval based on PUP is conditionally unbiased, resulting in valid inference both conditionally and unconditionally.

While improved predictions are possible, it should be pointed out that when serial correlation is a concern, we can focus on corrections at small h because ρ^h will be small for stationary mixing processes. Furthermore, dependence in the residuals is necessary but not sufficient for improved prediction. This is because the PUP adjustment depends not just on $\widehat{\rho}_{is}$ or $\widehat{\rho}_{js}$, but also on the values of $\widehat{e}_{i,T_0-s}$ and $\widehat{e}_{j,T_0+j-s}$ which can take on values close to zero. In the German unification example, the cross-section PUP correction does not make much difference. However, the residuals exhibit serial dependence and the AR(1) PUP correction changed the imputed growth rate for 1991 from -1.722 to -1.537 and for 1996 from -1.528 to -1.662. If take log GDP as the outcome variable instead of GDP growth, the residuals are still serially correlated with $\widehat{\rho}_1 = 0.72$. But $(\widehat{e}_{1,T_0}, \widehat{e}_{1,T_0-1}) = (0.025, -0.0024)$ which are small relative to $y_{1,T_0} = 9.26$. Hence in this case, the PUP corrections did not make appreciable difference. Ultimately, whether the PUP corrections are large is an empirical matter. Our goal is simply to draw awareness to the possibility of improvements. A practical first step could be to use the in-sample residuals $\widehat{e}_{it}$ to construct an LM test for no dependence using the auxiliary regression

$$\widehat{e}_{it} = X_{it}'\delta_0 + \sum_{s=0}^{p_i} \delta_{i,s}\widehat{e}_{i,t-s} + \sum_{j=1}^{N} \sum_{s=-p_j}^{p_j} \delta_{j,s}\widehat{e}_{j,t-s}.$$

We can entertain PUP corrections if the null hypothesis is rejected, keeping in mind that the significance of the corrections depend not only on the hypothesis to be tested, but also on the estimator used.

A Appendix: Proofs

Proof of Lemma 1. The proof of part (i) follows from standard arguments, see e.g., Hayashi (2000), p. 145. We provide a brief sketch here. Write $\widehat{\rho}_1 = \frac{\widehat{\gamma}_1}{\widehat{\gamma}_0}$, with $\widehat{\gamma}_j = \frac{1}{T_0}\sum_{t=1}^{T_0} \widehat{e}_t \widehat{e}_{t-j}$, for $j = 0, 1$. Let $\tilde{\gamma}_j = \frac{1}{T_0}\sum_{t=1}^{T_0} e_t e_{t-j}$ and note that $\tilde{\gamma}_j \to_p \gamma_j$ under Assumption A1. Since $\widehat{e}_t = y_t - X_t'\widehat{\beta} = e_t + X_t'(\beta - \widehat{\beta})$, we can write

$$\widehat{\gamma}_j = \tilde{\gamma}_j - T_0^{-1}\sum_{t=1}^{T_0}(X_{t-j}e_t + X_t e_{t-j})'(\widehat{\beta} - \beta) + (\widehat{\beta} - \beta)'(T_0^{-1}\sum_{t=1}^{T_0} X_t X_{t-j}')^{-1}(\widehat{\beta} - \beta).$$

The last two terms are $o_p(1)$ under the assumption that $\widehat{\beta} - \beta = o_p(1)$ and assuming that $E(X_{t-j}e_t)$, $E(X_t e_{t-j}')$ and $E\left(X_t X_{t-j}'\right)$ are all finite for $j = 0, 1$. It follows that $\widehat{\gamma}_j = \tilde{\gamma}_j + o_p(1) = \gamma_j + o_p(1)$, which implies the result. To prove (ii), note that $\widehat{e}_{T_0+1|T_0} = y_{T_0+1} - X_{T_0+1}'\widehat{\beta} = e_{T_0+1} - X_{T_0+1}'(\widehat{\beta} - \beta) = e_{T_0+1} + o_p(1)$ under the assumption that $\widehat{\beta} - \beta = o_p(1)$. It follows immediately that $E(e_{T_0+1}) = 0$ and $Var(e_{T_0+1}) \equiv \gamma_0 = \sigma_e^2$. For (iii), note that we can write $\widehat{e}^{+}_{T_0+1|T_0} = e_{T_0+1} - \rho_1 e_{T_0} + o_p(1)$, given (i) and the assumption that $\widehat{\beta} - \beta = o_p(1)$. Since $E(e_t) = 0$ for all t, we have $E(e_{T_0+1} - \rho_1 e_T) = 0$. In addition, $Var(e_{T_0+1} - \rho_1 e_T) = \gamma_0 + \rho_1^2\gamma_0 - 2\rho_1\gamma_1 = \gamma_0 - \gamma_1^2/\gamma_0 \leq \gamma_0$, where the second equality follows by replacing $\rho_1 = \gamma_1/\gamma_0$.

Proof of Lemma 2. The standard OLS prediction error is

$$\widehat{e}_{T_0+h|T_0} = y_{T_0+h} - \widehat{y}_{T_0+h|T_0} = e_{T_0+h} - X_{T_0+h}'(\widehat{\beta} - \beta) = e_{T_0+h} + o_p(1).$$

The mean of e_{T_0+h} is zero and the variance is γ_0. The h-step ahead direct PLUP based on an AR(1) approximation is

$$\widehat{e}^{+d}_{T_0+h|T_0} = y_{T_0+h} - \widehat{y}^{+d}_{T_0+h|T_0} = e_{T_0+h} - \widehat{\rho}_h\widehat{e}_{T_0} + o_p(1) = e_{T_0+h} - \rho_h e_{T_0} + o_p(1),$$

where $\widehat{e}_{T_0}$ is the OLS residual for observation T_0, and $\widehat{\rho}_h$ is the h-th order sample autocorrelation of $\widehat{e}_t$. It follows immediately that the asymptotic MSE of the direct PLUP is smaller or equal than that of OLS since the variance of $e_{T_0+h} - \rho_h e_{T_0}$ is $\gamma_0(1 - \rho_h^2) \leq \gamma_0$, since $\rho_h^2 \leq 1$. To compare direct PLUP with iterated PLUP , note that the h-step ahead iterated PLUP based on an AR(1) approximation is

$$\widehat{y}^{+I}_{T_0+h|T_0} = \widehat{y}_{T_0+h|T_0} + \widehat{\rho}_1^h\widehat{e}_{T_0},$$

where $\widehat{\rho}_1$ is defined in (5) and $\widehat{e}_{T_0}$ is the OLS residual for observation T_0. The iterated PLUP error is

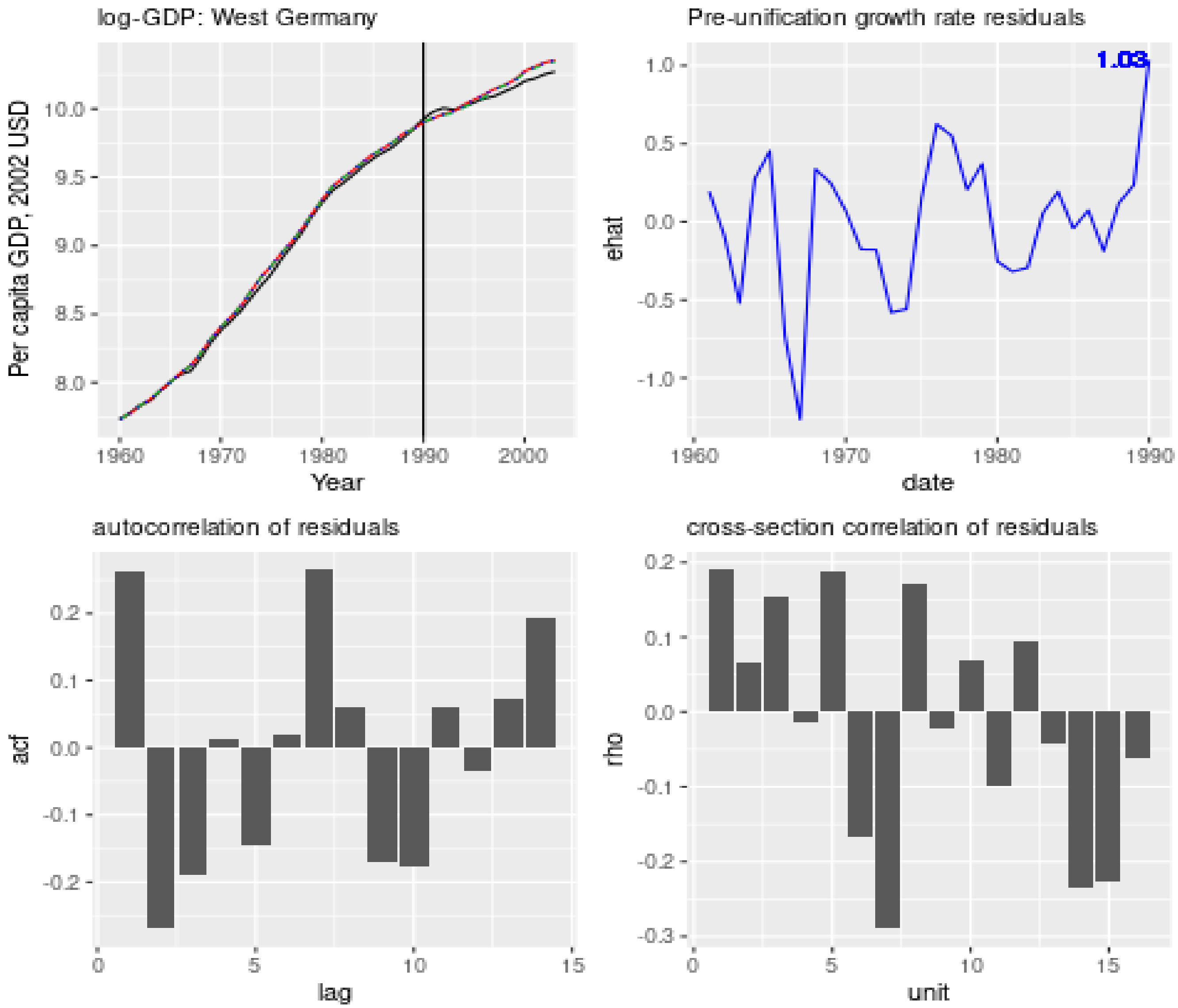
$$\widehat{e}^{+I}_{T_0+h|T_0} = y_{T_0+h} - \widehat{y}^{+I}_{T_0+h|T_0} = e_{T_0+h} - \widehat{\rho}_1^h\widehat{e}_{T_0} + o_p(1) = e_{T_0+h} - \rho_1^h e_{T_0} + o_p(1).$$

Both versions of PLUP have mean zero but the variance of iterated PLUP is larger or equal than that of direct PLUP since we can show that the difference between these two variances is $\gamma_0(\rho_h - \rho_1^h)^2 \geq 0$. A comparison between iterated PLUP and the standard prediction reveals that both prediction errors have mean zero, but the variance of the standard OLS prediction is $Var\left(e_{T_0+h}\right) = \gamma_0$ whereas that of the iterated PLUP error is equal to

$$
\begin{aligned}
Var\left(e_{T_0+h} - \rho_1^h e_{T_0}\right) &= \gamma_0 + \rho_1^{2h}\gamma_0 - 2\rho_1^h\gamma_h \\
&= \gamma_0[1 + \rho_1^{2h} - 2\rho_1^h(\frac{\gamma_h}{\gamma_0})] \\
&= \gamma_0[1 + \rho_1^{2h} - 2\rho_1^h\rho_h)],
\end{aligned}
$$

since $\rho_h = \frac{\gamma_h}{\gamma_1}$, where $\gamma_h = E\left(e_{t+h}e_t\right)$. When $1 + \rho_1^{2h} - 2\rho_1^h\rho_h \leq 1$ we can conclude that the mean square prediction error of PLUP is smaller than the mean square prediction error of OLS.

Figure 1: Effect of German Unification, GDP growth



The data are downloaded from `https://doi.org/10.7910/DVN/24714`

Figure 2: Effect of German Unification on level of GDP

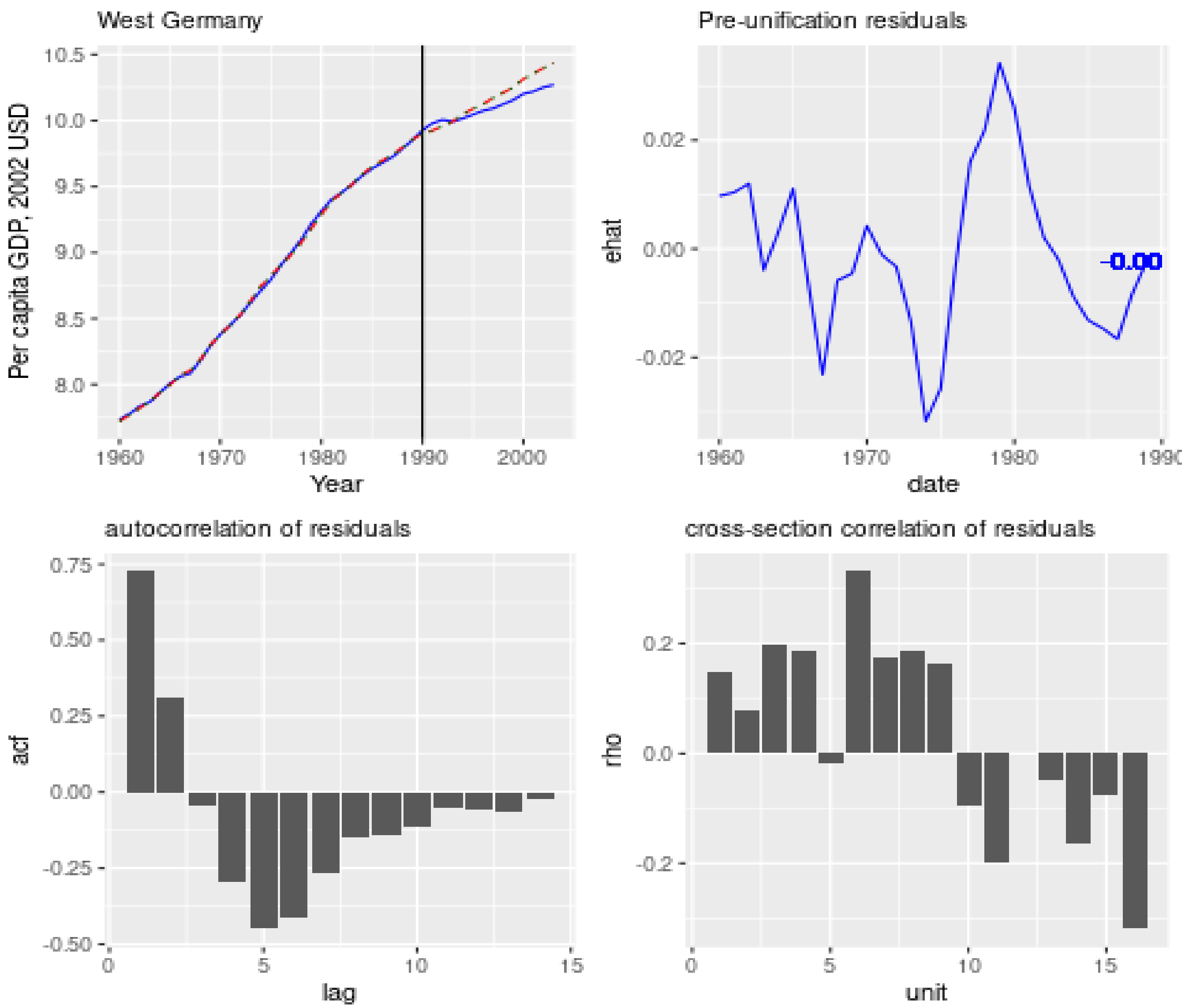


Table 1: Prediction Errors

$$
\begin{aligned}
y_t &= X_{1t}\beta_1 + X_{2t}\beta_2 + e_t, \quad \beta_1 = \beta_2 = 1, \quad T = 50 \\
X_{1t} &= 0.6X_{1,t-1} + u_{1t}, \quad u_{1t} \sim N(0,1), \qquad X_{2t} = u_{2t}, \quad u_{2t} \sim N(0,1) \\
e_t &= \phi_1 e_{t-1} + \phi_2 e_{t-2} + v_t, \quad v_t \sim N(0, 0.05)
\end{aligned}
$$

	Unconditional									
h	best	noadj	ols	plupI	plupD	best	noadj	ols	plupI	plupD
	bias					mse				
	(ϕ_1,ϕ_2)=(0.80,0.00)									
1	-0.00	-0.01	-0.00	-0.00	-0.00	0.05	0.14	0.14	0.05	0.05
2	0.00	-0.00	0.00	0.00	0.00	0.05	0.14	0.14	0.08	0.08
5	0.00	0.00	0.00	0.00	0.00	0.05	0.13	0.14	0.12	0.12
10	0.00	-0.00	-0.00	-0.00	0.00	0.05	0.14	0.14	0.14	0.14
avg	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.07	0.08	0.06	0.06
	(ϕ_1,ϕ_2)=(1.30,-0.40)									
1	-0.00	-0.01	-0.00	-0.00	-0.00	0.05	0.43	0.43	0.06	0.06
2	0.00	-0.00	-0.00	-0.00	-0.00	0.05	0.43	0.43	0.16	0.16
5	0.00	-0.00	0.00	-0.00	0.00	0.05	0.42	0.43	0.37	0.35
10	0.00	-0.01	-0.00	-0.00	-0.00	0.05	0.43	0.44	0.48	0.44
avg	-0.00	-0.00	-0.00	-0.00	-0.00	0.00	0.28	0.29	0.21	0.20
	Conditional on $e_{T_0} = 1.0$ and $e_{T_0-1} = 0.5$									
	bias					mse				
	(ϕ_1,ϕ_2)=(0.80,0.00)									
1	-0.00	0.80	0.79	0.01	0.01	0.05	0.68	0.68	0.05	0.05
2	0.00	0.64	0.64	0.01	0.02	0.05	0.49	0.49	0.09	0.09
5	0.00	0.33	0.33	0.01	0.02	0.05	0.23	0.23	0.13	0.13
10	0.00	0.11	0.10	0.00	0.02	0.05	0.15	0.15	0.14	0.15
avg	-0.00	0.36	0.35	0.01	0.02	0.00	0.18	0.18	0.06	0.06
	(ϕ_1,ϕ_2)=(1.30,-0.40)									
1	-0.00	1.10	1.09	0.18	0.18	0.05	1.25	1.26	0.09	0.09
2	0.00	1.03	1.03	0.18	0.24	0.05	1.19	1.20	0.17	0.20
5	0.00	0.62	0.62	-0.04	0.20	0.05	0.72	0.73	0.34	0.38
10	0.00	0.21	0.21	-0.23	0.08	0.05	0.47	0.48	0.49	0.45
avg	-0.00	0.61	0.61	-0.04	0.17	0.00	0.56	0.56	0.19	0.22

5000 replications using Matlab R2019a, seed = rng(1234,'twister')

]

Table 1(b) Coverage

h	best	noadj	ols	plupI	plupD	best	noadj	ols	plupI	plupD
	T=50									
	Unconditional					Conditional				
	(ϕ_1, ϕ_2)=(0.80,0.00)									
1	0.96	0.95	0.95	0.95	0.95	0.96	0.40	0.41	0.95	0.95
2	0.95	0.95	0.95	0.95	0.95	0.95	0.64	0.63	0.95	0.95
5	0.96	0.95	0.95	0.95	0.95	0.96	0.88	0.88	0.95	0.94
10	0.95	0.95	0.95	0.94	0.94	0.95	0.94	0.94	0.94	0.94
avg	0.95	0.99	0.99	0.94	0.94	0.96	0.94	0.93	0.94	0.94
	(ϕ_1, ϕ_2)=(1.30,-0.40)									
1	0.96	0.95	0.95	0.95	0.95	0.96	0.79	0.77	0.92	0.92
2	0.95	0.95	0.95	0.95	0.95	0.95	0.76	0.75	0.94	0.92
5	0.96	0.95	0.95	0.95	0.95	0.96	0.87	0.87	0.96	0.93
10	0.95	0.95	0.94	0.95	0.94	0.95	0.94	0.94	0.94	0.94
avg	0.95	0.98	0.98	0.94	0.94	0.96	0.94	0.93	0.96	0.92
	T=200									
	Unconditional					Conditional				
	(ϕ_1, ϕ_2)=(0.80,0.00)									
1	0.96	0.95	0.95	0.95	0.95	0.96	0.39	0.40	0.95	0.95
2	0.95	0.94	0.94	0.94	0.94	0.96	0.63	0.63	0.94	0.93
5	0.95	0.95	0.94	0.94	0.94	0.96	0.87	0.86	0.93	0.92
10	0.95	0.94	0.93	0.93	0.92	0.96	0.93	0.92	0.94	0.91
avg	0.94	0.99	0.98	0.92	0.92	0.95	0.93	0.91	0.91	0.89
	(ϕ_1, ϕ_2)=(1.30,-0.40)									
1	0.96	0.95	0.94	0.95	0.95	0.96	0.73	0.69	0.91	0.91
2	0.95	0.94	0.94	0.94	0.94	0.96	0.72	0.70	0.93	0.90
5	0.95	0.94	0.93	0.94	0.93	0.96	0.85	0.84	0.94	0.91
10	0.95	0.93	0.92	0.93	0.91	0.96	0.93	0.91	0.93	0.91
avg	0.94	0.98	0.97	0.94	0.91	0.95	0.92	0.90	0.94	0.89

Table 2: Errors in Estimated Treatment Effect : $T = 50, N = 20$

$$\begin{aligned} Y_{it}(0) &= c_i + \Lambda_i' F_t + e_{it}, \quad \delta_1 = 0.1, 1 \\ e_{it} &= \phi_i e_{it-1} + v_{it}, \quad v_{it} \sim N(0, 0.25), \quad \phi_i = 0.6, \phi_j = 0, j > 1 \\ F_{1t} &= 0.8 F_{1,t-1} + e^F_{1t}, \quad e^F_{1t} \sim N(0, 0.5), \quad \Lambda_{1i} \sim N(0,1) \\ F_{2t} &= 0.5 F_{2,t-1} + e^F_{2t}, \quad e^F_{2t} \sim N(0, .3), \quad \Lambda_{2i} \sim N(0,1). \end{aligned}$$

h	best	noadj	pca	plupI	plupD	best	noadj	pca	plupI	plupD
	$e_{1t} = 0.6 e_{1t-1} + \epsilon_{1t}$									
	Unconditional bias of $\widehat{\delta}_{i,T_0+h}$					mse				
1	0.00	-0.00	0.00	0.01	0.01	0.25	0.37	0.43	0.33	0.33
2	-0.01	-0.01	-0.00	-0.00	-0.00	0.33	0.37	0.45	0.42	0.42
5	-0.00	-0.00	0.01	0.01	0.01	0.37	0.39	0.48	0.47	0.48
10	-0.01	-0.01	-0.00	-0.00	-0.00	0.39	0.40	0.50	0.50	0.51
avg	-0.00	-0.00	0.00	0.00	0.00	0.12	0.13	0.18	0.17	0.17
	$e_{1t} = 0.6 e_{1t-1} + \epsilon_{1t}$, e_{T_0}=1.00									
	Conditional bias of $\widehat{\delta}_{i,T_0+h}$					mse				
1	0.00	0.60	0.56	0.15	0.15	0.25	0.61	0.66	0.36	0.36
2	-0.01	0.35	0.32	0.12	0.14	0.33	0.45	0.53	0.45	0.47
5	-0.00	0.21	0.19	0.09	0.13	0.37	0.42	0.51	0.48	0.52
10	-0.01	0.12	0.09	0.04	0.09	0.39	0.40	0.51	0.50	0.54
avg	-0.00	0.15	0.12	0.04	0.09	0.12	0.14	0.19	0.17	0.19
	$e_{1t} = 0.5 e_{2t} + \epsilon_{1t}$									
	Unconditonal bias of $\widehat{\delta}_{i,T_0+h}$					mse				
1	0.00	-0.00	0.00	0.00	0.00	0.15	0.20	0.33	0.28	0.28
2	-0.01	-0.01	-0.00	-0.01	-0.01	0.15	0.21	0.35	0.29	0.29
5	0.00	-0.00	0.01	0.01	0.01	0.15	0.21	0.35	0.30	0.30
10	-0.01	-0.01	-0.01	0.00	0.00	0.15	0.21	0.34	0.29	0.29
avg	-0.00	-0.00	0.00	0.01	0.01	0.01	0.02	0.05	0.04	0.04
	$e_{1t} = 0.5 e_{2t} + \epsilon_{1t}$, $e_{2T_0+1} = 0.27539$									
	Conditional bias of $\widehat{\delta}_{i,T_0+h}$					mse				
1	0.00	0.14	0.14	0.04	0.04	0.15	0.17	0.29	0.27	0.27
2	-0.01	0.49	0.49	0.12	0.12	0.15	0.38	0.52	0.35	0.35
5	0.00	0.47	0.48	0.13	0.13	0.15	0.37	0.52	0.35	0.35
10	-0.01	-0.45	-0.44	-0.14	-0.14	0.15	0.35	0.48	0.34	0.34
avg	-0.00	-0.12	-0.12	-0.06	-0.06	0.01	0.03	0.06	0.05	0.05

5000 replications using Matlab 2019a with seed rng(456,'twister'). For cross-section correlation, $e_{2,T_0:1:T}$=(0.6361, -0.6386, 0.7118, -1.7044, -1.2992, 1.6402, 0.1395, 0.9348, 0.5051, 0.9692).

Table 2(b), Coverage

h	best	noadj	pca	plupI	plupD	best	noadj	pca	plupI	plupD
	$e_{1t} = \phi e_{1t-1} + v_{1t}$									
	(T_0, N_0)=(50,20):Unconditional					Conditional				
1	0.95	0.94	0.92	0.93	0.93	0.96	0.87	0.86	0.92	0.92
2	0.91	0.94	0.91	0.91	0.91	0.92	0.92	0.90	0.91	0.90
5	0.88	0.93	0.90	0.88	0.87	0.90	0.94	0.91	0.89	0.87
10	0.89	0.94	0.90	0.84	0.83	0.91	0.94	0.91	0.85	0.84
avg	0.98	0.85	0.75	0.77	0.75	0.98	0.83	0.73	0.76	0.72
	(T_0, N_0)=(200,50): Unconditional					Conditional				
1	0.95	0.95	0.94	0.95	0.95	0.96	0.90	0.89	0.95	0.95
2	0.91	0.95	0.94	0.94	0.94	0.91	0.92	0.92	0.94	0.94
5	0.89	0.95	0.94	0.93	0.93	0.89	0.95	0.94	0.94	0.93
10	0.89	0.95	0.94	0.94	0.93	0.90	0.95	0.94	0.94	0.93
avg	0.99	0.93	0.92	0.92	0.92	0.99	0.92	0.90	0.92	0.91
	$e_{1t} = \theta_1 e_{2t} + v_{1t}$									
	(T_0, N_0)=(50,20): Unconditional					Conditional				
1	0.95	0.95	0.94	0.93	0.93	0.95	0.97	0.95	0.93	0.93
2	0.95	0.95	0.93	0.93	0.93	0.95	0.86	0.87	0.90	0.90
5	0.95	0.94	0.93	0.91	0.91	0.95	0.91	0.91	0.90	0.90
10	0.95	0.95	0.93	0.88	0.88	0.95	0.98	0.96	0.89	0.89
avg	0.88	0.88	0.79	0.78	0.78	0.88	0.81	0.75	0.77	0.77
	(T_0, N_0)=(200,50): Unconditional					Conditional				
1	0.95	0.96	0.95	0.95	0.95	0.95	0.94	0.94	0.95	0.95
2	0.95	0.95	0.94	0.94	0.94	0.95	0.94	0.93	0.94	0.94
5	0.95	0.95	0.95	0.95	0.95	0.95	0.75	0.78	0.93	0.93
10	0.95	0.95	0.95	0.94	0.94	0.95	0.86	0.87	0.93	0.93
avg	0.94	0.94	0.92	0.92	0.92	0.94	0.92	0.91	0.92	0.92

References

ABADIE, A., A. DIAMOND, AND J. HAINMUELLER (2010): "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program," *Journal of the American Statistical Association*, 105:490, 493–505, DOI: 10.1198/jasa.2009.ap08746.

ABADIE, A., AND J. GARDEAZABAL (2003): "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93:1.

ARKHANGELSKY, D., S. ATHEY, D. A. HIRSHBERG, G. W. IMBENS, AND S. WAGER (2021): "Syntheticx Difference-in-Differences," *American Economic Review*, 111-(12), 4088–4118, DOI: 10.1257/aer.20190159.

ATHEY, S., M. BAYATI, N. DOUDCHENKO, G. IMBENS, AND K. KHOSRAVI (2021): "Matrix Completion Methods for Causal Panel Data Models," *Journal of the American Statistical Association*, 116:536, 1716–1730, DOI: 10.1080/01621459.2021.1891924.

BAI, J., AND S. NG (2021): "Matrix Completion, Counterfactuals, and Factor Analysis of Missing Data," *Journal of the American Statistical Association*, 116:536, 1746–1763.

BALTAGI, B. (2008): "Forecasting with Panel Data," *Journal of Forecasting*, 27, 153–173.

——— (2013): "Panel Data Forecasting," *Handbook of Forecasting*, 2B, 995–2021.

BEN-MICHAEL, E., A. FELLER, AND J. ROTHSTEIN (2021): "The Augmented Synthetic Control Method," *Journal of the American Statistical Associaton*, 116:536, 1789–1803.

BRODERSEN, K., F. GALLUSSER, J. KOEHLER, N. REMY, AND S. SCOTT (2015): "Inferring Causal Impact Using Bayesian Structural Time-Series Models," *Annals of Applied Statistics*, 9:1, 247–274.

CARVALHO, C., R. MASINI, AND M. MEDEIROS (2018): "ArCo: An Artificial Counterfactual Approach for High-Diemsnional Panel Time Series Data," *Journal of Econometrics*, 207, 352–380.

CATTANEO, M. D., Y. FENG, AND R. TITIUNIK (2021): "Prediction Intervals for Synthetic Control Methods," *Journal of the American Statistical Association*, 116:536, 1865–1880, DOI: 10.1080/01621459.2021.1979561.

CHERNOZHUKOV, V., K. WÜTHRICH, AND Y. ZHU (2021): "An Exact and Robust Conformal Inference Method for Counterfactual and Synthetic Controls," *Journal of the American Statistical Association*, 116:536, 1849–1864.

CHOW, G. C., AND A. LIN (1971): "Best Linear Unbiased Interpolation, Distribution, and Extrapolation of Time Series by Related Series," *The Review of Economics and Statistics*, 53:4, 372–375.

COCHRANE, D., AND G. H. ORCUTT (1949): "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Assocation*, 44, 32–64.

FAN, J., R. MASINI, AND M. MEDEIROS (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.

Ferman, B. (2023): "Inference in Difference-in-Differences: How Much Should Should We Trust in Independent Clusters," *Journal of Applied Econometrics*, 38:3, 270–449.

Ferman, B., and C. Pinto (2021): "Synthetic Controls with Imperfect Pretreatment Fit," *Quantitative Economics*, 12, 1197–1221.

Goldberger, A. (1962): "Best Linear Unbiased Prediction in the Generalized Linear Regression Model," *Journal of the American Statistical Association*, 57, 369–375.

Henderson, C. R. (1950): "Estimation of Genetic Parameters," *Annals of Mathematical Statistics*, 21, 309–310.

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

Li, K. (2020): "Statistical Inference for Average Treatment Effects Estimated by Synthetic Control Methods," *Journal of the American Statistical Association*, 115:532, 2068–2083.

Li, X., Y. Shen, and Q. Zhou (2023): "Confidence Intervals of Treatment Effects in Panel Data Models with Interactive Fixed Effects," *Journal of Econometrics*, 240:1, Article 105684.

Little, R., and D. B. Rubin (2019): *Statistical Analysis with Missing Data*. Wiley, New York, third edn.

Masini, R., and M. Medeiros (2022): "Counterfactual Analysis and Inference with Nonstationary Data," *Journal of Business and Economic Statistics*, 40:1, 227–239.

Masini, R., and M. C. Medeiros (2021): "Counterfactual Analysis With Artificial Controls: Inference, High Dimensions, and Nonstationarity," *Journal of the American Statistical Association*, 116:536, 1773–1788, DOI: 10.1080/01621459.2021.1964978.

Ng, S., and S. Scanlan (2024): "Constructing High Frequency Economic Indicators by Imputation," *Econometrics Journal*, 27:1, 1–30.

Phillips, P. (1979): "The Sampling Distribution of Forecasts From a First-Order Autoregression," *Journal of Econometrics*, 9, 241–261.

Robinson, G. K. (1991): "That BLUP is a Good Thing: The Estimation of Random Effects," *Statistical Science*, 6(1), 15 – 32, https://doi.org/10.1214/ss/1177011926.

Spall, J. (1991): "Comment: The Kalman Filter and BLUP," *Statistical Science*, 6:1, 39–41.

Taub, A. (1979): "Prediction in Context of Variance-Components Model," *Journal of Econometrics*, 10, 103–107.

Xu, Y. (2017): "Generalized Synthetic Control Methods: Causal Inference with Interactive Fixed Effects Models," *Political Analysis*, pp. 56–75.