ISSN 1936-5330



A Note on the Finite Sample Bias in Time Series Cross-Validation

Amaze Lusompa
November 2025; updated December 2025
RWP 25-17
https://doi.org/10.18651/RWP2025-17



A Note on the Finite Sample Bias in Time Series Cross-Validation

Amaze Lusompa *
Federal Reserve Bank of Kansas City

December 5, 2025

Abstract

It is well known that model selection via cross validation can be biased for time series models. However, many researchers have argued that this bias does not apply when using cross-validation with vector autoregressions (VARs) or with time series models whose errors follow a martingale-like structure. I show that even under these circumstances, performing cross-validation on time series data will still generate bias in general.

*I thank Johnson Oliyide for excellent research assistance. The views expressed are those of the author and do not necessarily reflect the positions of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

1 Introduction

Cross-validation (CV) is a widely used tool for model/hyperparameter selection, model comparison, model averaging/combination, or to evaluate the predictive ability of models and is an important tool in statistics, econometrics, machine learning, and many other fields. It is considered by many to be the gold standard of model selection. Though CV is flexible and robust, it was developed for independent data.

It is well known that model selection via CV can be biased for time series data (Altman, 1990, Hart, 1991, Opsomer et al., 2001). It was first argued by Burman and Nolan (1992) that this bias does not apply if the residuals of the models follow a martingale difference sequence (MDS) or martingale-like structure, and it has since become conventional wisdom that CV will work well if this is the case (see for example Burman et al. (1994), Racine (2000), Arlot and Celisse (2010), Gelman et al. (2014), Bergmeir et al. (2018), Coulombe et al. (2022), Petropoulos et al. (2022)). This argument has more recently picked up steam due to Bergmeir et al. (2018) using a reformulated proof specific for Vector Autoregressions (VARs). In this note I will show that CV with time series data will, in general, be biased even if the residuals are MDS.

2 How Cross-Validation Works

Let

$$y_i = \mu_i + \varepsilon_i \; for \; i = 1, 2, \ldots, T,$$

where y_i is the dependent variable, μ_i is the conditional mean, and ε_i is the residual. In the special case of linear regressions

$$y_i = x_i\beta + \varepsilon_i \; for \; i = 1, 2, \ldots, T,$$

where x_i is a $1 \times p$ vector of regressors, β is a $p \times 1$ vector of regression coefficients. The Leave-one-out (LOO) residual for a regression model is calculated by estimating β for all but one observation, and then predicting the residual for the left out observation. More formally, for observation i,

$$\widetilde{\beta}_{-i} = (\sum_{j \neq i} x_j' x_j)^{-1} \sum_{j \neq i} x_j' y_j, \qquad \widetilde{\mu}_{-i} = x_i \widetilde{\beta}_{-i}, \qquad \widetilde{\varepsilon}_i = y_i - \widetilde{\mu}_{-i},$$

where $\widetilde{\beta}_{-i}$, $\widetilde{\mu}_{-i}$, and $\widetilde{\varepsilon}_i$ are the LOO estimates of β, μ_i, and ε_i, respectively. LOO CV is generally calculated as the mean-squared error (MSE) of the LOO residuals, that is, $T^{-1}\sum_{i=1}^{T}\widetilde{\varepsilon}_i^2$, though other loss functions could be used. The Burman and Nolan (1992) and Bergmeir et al. (2018) arguments are with respect to the MSE case, so that is the case that will be used throughout.

Define $\mathcal{A}_T$ as the set of all models being compared, and α is the model index. Generally, the goal of CV (or any model selection criterion) is to find the model with the smallest average squared error (ASE) or mean average squared error (MASE). The ASE is a distance or loss function between the true conditional mean and the estimated conditional mean for model α and is a measure of how well the conditional mean for model α approximates the true conditional mean. A model with a smaller ASE means that model is closer to the truth in the squared error loss sense. The ASE is defined as

$$\widetilde{L}_T(\alpha) = \frac{1}{T}\sum_{i=1}^{T}(\mu_i - \widetilde{\mu}_{-i}(\alpha))^2 \qquad and \qquad L_T(\alpha) = \frac{1}{T}\sum_{i=1}^{T}(\mu_i - \hat{\mu}_i(\alpha))^2$$

for the LOO and full sample estimates, respectively.

$$\widetilde{R}_T(\alpha) = E(\widetilde{L}_T(\alpha)) \qquad and \qquad R_T(\alpha) = E(L_T(\alpha))$$

are the LOO and full sample estimates of the MASE, respectively. CV is considered to be a good model selection criterion in small samples if CV can accurately choose the model with the smallest ASE or MASE on average.

3 Bias in Time Series Cross-Validation

To illustrate the intuition behind CV, I will make the following assumption:

Assumption 1. *The following conditions are satisfied for all models, α, in the set A_T :*

(a) For all i, $E(\varepsilon_i) = 0$, $E(\varepsilon_i^2) = \sigma^2$, $E(\mu_i\varepsilon_i) = 0$.

(b) For each model, α, and sample size T, $\frac{1}{T}\sum_{i=1}^{T}E(\mu_i - \hat{\mu}_i(\alpha))^2 \approx \frac{1}{T}\sum_{i=1}^{T}E(\mu_i - \widetilde{\mu}_{-i}(\alpha))^2$.

Remark 1. This assumption should be uncontroversial and is standard (Opsomer et al., 2001). Assumption 1(a) comprises of standard moment conditions for regressions. Assumption 1(b), assuming the full sample MASE and the LOO MASE being approximately the same, should hold in general unless some observations have extremely high leverage.

Note that since the following equation holds for the LOO residual for model α: $\widetilde{\varepsilon}_{-i}(\alpha) = y_i - \widetilde{\mu}_{-i}(\alpha) = \mu_i + \varepsilon_i - \widetilde{\mu}_{-i}(\alpha)$, it can be shown via algebra that

$$\widetilde{\varepsilon}^2_{-i}(\alpha) = \varepsilon_i^2 + 2(\mu_i - \widetilde{\mu}_{-i}(\alpha))\varepsilon_i + (\mu_i - \widetilde{\mu}_{-i}(\alpha))^2.$$

Averaging over all observations, we have the following formula for the CV MSE

$$\frac{1}{T}\sum_{i=1}^{T}\widetilde{\varepsilon}^2_{-i}(\alpha) = \frac{1}{T}\sum_{i=1}^{T}\varepsilon_i^2 + \frac{1}{T}\sum_{i=1}^{T}2\mu_i\varepsilon_i - \frac{1}{T}\sum_{i=1}^{T}2\widetilde{\mu}_{-i}(\alpha)\varepsilon_i + \frac{1}{T}\sum_{i=1}^{T}(\mu_i - \widetilde{\mu}_{-i}(\alpha))^2.$$

CV is thought to perform well in finite samples if the expectation of the CV MSE satisfies

$$E(\frac{1}{T}\sum_{i=1}^{T}\widetilde{\varepsilon}^2_{-i}(\alpha)) \approx \sigma^2 + \frac{1}{T}\sum_{i=1}^{T}E(\mu_i - \hat{\mu}_i(\alpha))^2$$

(Altman, 1990, Hart, 1991, Opsomer et al., 2001). If this formula holds, CV will, on average, choose the model with the smallest MASE since the only thing differentiating the expectation of the CV MSE across models is the MASE for the respective models. By Assumption 1(a), $E(\frac{1}{T}\sum_{i=1}^{T}\varepsilon_i^2) = \sigma^2$ and $E(\frac{1}{T}\sum_{i=1}^{T}2\mu_i\varepsilon_i) = 0$, so in order for CV to perform well on average in finite samples, it essentially needs to be the case that $E(\frac{1}{T}\sum_{i=1}^{T}\widetilde{\mu}_{-i}(\alpha)\varepsilon_i) = 0$.[1] If $E(\frac{1}{T}\sum_{i=1}^{T}\widetilde{\mu}_{-i}(\alpha)\varepsilon_i) \neq 0$, then the bias can change the ordering of the CV MSE and cause CV not to choose the model with the smallest MASE on average. Note that outside of razor's edge cases, in order for $E(\frac{1}{T}\sum_{i=1}^{T}\widetilde{\mu}_{-i}(\alpha)\varepsilon_i) = 0$, it needs to be the case that $E(\widetilde{\mu}_{-i}(\alpha)\varepsilon_i) = 0$ for all i.

If the data is i.i.d. and $E(\varepsilon_i|x_i(\alpha)) = 0$ for all α, then CV would be unbiased since

$$E[\widetilde{\mu}_{-i}(\alpha)\varepsilon_i] = E[x_i(\alpha)\widetilde{\beta}_{-i}(\alpha)\varepsilon_i] = E[x_i(\alpha)\varepsilon_i]E[\widetilde{\beta}_{-i}(\alpha)] = 0,$$

where the second equality is due to independence. For time series data, it will generally be the case that $E(\widetilde{\mu}_{-i}(\alpha)\varepsilon_i) \neq 0$ since $\widetilde{\mu}_{-i}(\alpha)$ and ε_i will generally be correlated in finite samples, and contrary to conventional wisdom in the literature, ε_i being a MDS is not sufficient to guarantee that $E(\widetilde{\mu}_{-i}(\alpha)\varepsilon_i) = 0$ for all i.

As a counterexample, let us assume the true model is a stationary AR(1) where

$$y_i = \rho y_{i-1} + \varepsilon_i,$$

[1] CV could also perform well if $E(\frac{1}{T}\sum_{i=1}^{T}\widetilde{\mu}_{-i}(\alpha)\varepsilon_i) \approx 0$, but the performance of CV would still depend on the magnitude of the MASE, since $E(\frac{1}{T}\sum_{i=1}^{T}\widetilde{\mu}_{-i}(\alpha)\varepsilon_i) \approx 0$ could still affect the ordering of the models.

$|\rho| < 1$, $\rho \neq 0$. Assume Assumption 1 still holds, so ε_i is a mean zero random variable with finite variance. Also assume that ε_i is a MDS, that is, $E(\varepsilon_i|\varepsilon_{i-1}, \varepsilon_{i-2}, ...) = 0$ for all i. Lastly, let us also assume that $E[\widetilde{\mu}_{-i}(\alpha)\varepsilon_i]$ exists and is finite for all i and T.[2] For simplicity, let us also assume that the candidate model being analyzed is the true model so that $\varepsilon_i(\alpha) = \varepsilon_i$ and $x_i(\alpha) = x_i$ and therefore

$$\widetilde{\mu}_{-i}(\alpha)\varepsilon_i = x_i(\alpha)\widetilde{\beta}_{-i}(\alpha)\varepsilon_i = x_i(\sum_{j\neq i} x_j'x_j)^{-1}(\sum_{j\neq i} x_j'y_j)\varepsilon_i.$$

Since

$$y_i = \rho y_{i-1} + \varepsilon_i = \sum_{m=0}^{\infty} \rho^m \varepsilon_{i-m},$$

then

$$E[\widetilde{\mu}_{-i}(\alpha)\varepsilon_i] = E[x_i(\alpha)\widetilde{\beta}_{-i}(\alpha)\varepsilon_i] = E[y_{i-1}(\sum_{j\neq i} y_{j-1}'y_{j-1})(\sum_{j\neq i} y_{j-1}'y_j)\varepsilon_i]$$

$$= E\{[\sum_{m=0}^{\infty} \rho^m \varepsilon_{i-1-m}](\sum_{j\neq i}[\sum_{m=0}^{\infty} \rho^m \varepsilon_{j-1-m}]'[\sum_{m=0}^{\infty} \rho^m \varepsilon_{j-1-m}])^{-1}(\sum_{j\neq i}[\sum_{m=0}^{\infty} \rho^m \varepsilon_{j-1-m}]'[\sum_{m=0}^{\infty} \rho^m \varepsilon_{j-m}])\varepsilon_i\},$$

which clearly will not be equal to zero in general for all i since the MDS assumption is not sufficient to guarantee it.[3,4] This is because even though the MDS assumption ensures the error terms are uncorrelated, it allows for dependence in the error terms, so the expectation of functionals of the error terms will not be zero in general. In this specific counterexample, the bias is somewhat reminiscent on the finite sample bias for autocovariances or VAR parameters (see for example Kendall (1954), Marriott and Pope (1954), Shaman and Stine (1988)) except things are further complicated by the additional multiplication by $x_i\varepsilon_i$.[5] The above counterexample assumes the true model is an AR(1) model, and the candidate model being examined is the true model, but in general, the bias will depend on the candidate model and the true data generating process, so it will generally

[2] One could make more explicit distribution assumptions for ε_i instead, but the argument would lose its generality.

[3] An exception would be when $i = T$. When $i = T$, $E[\widetilde{\mu}_{-i}(\alpha)\varepsilon_i] = 0$ due to the MDS assumption that $E(\varepsilon_T|\varepsilon_{T-1}, \varepsilon_{T-2}, \ldots) = 0$ and since $x_T = y_{T-1}$ and $\widetilde{\beta}_{-T}$ are functions of ε up to time $T-1$, $E[\widetilde{\mu}_{-i}(\alpha)\varepsilon_i] = E(E[\varepsilon_i|\widetilde{\mu}_{-i}(\alpha)]) = 0$. More generally, the value of $E[\widetilde{\mu}_{-i}(\alpha)\varepsilon_i]$ can depend on i.

[4] Note that the bias goes away asymptotically under general conditions (see for example Theorem 1 in Lusompa (2025) for the proof). Intuitively, this is because $\widetilde{\beta}_{-i}(\alpha)$ will converge in probability to $\beta(\alpha)$, which is simply the convergence of pseudo true parameters (see for example White (1982, 1994) and references therein for more details). Since $\beta(\alpha)$ is a constant vector and assuming the standard moment condition $E[x_i(\alpha)\varepsilon_i] = 0$ holds, $E[x_i(\alpha)\varepsilon_i\beta(\alpha)] = E[x_i(\alpha)\varepsilon_i]\beta(\alpha) = 0$.

[5] For more general cases, the bias would be somewhat reminiscent of the more general time series bias formula (see for example Anatolyev (2005), Bao and Ullah (2007)).

be the case that the bias can vary from one candidate model to the next and can affect the model chosen by CV.

Since most time series models include lagged dependent variables and are assumed to have a Wold type representation or a Wold type component (see for example Priestley (1981), Brockwell and Davis (1991)), it will generally be the case that $E(\widetilde{\mu}_{-i}(\alpha)\varepsilon_i) \neq 0$ in finite samples. So just like lagged dynamics in the dependent variable cause bias in regression coefficients for time series models, it would also cause bias in CV. If strict exogeneity was satisfied for the true and candidate regressors, that is, for each candidate model $E(\varepsilon_i|x_1(\alpha), x_1, \ldots, x_T(\alpha), x_T) = 0$ for all i in addition to a martingale type assumption such as the residuals satisfying $E(\varepsilon_i|\varepsilon_j) = 0$ for $i > j$, then the bias would be zero and time series CV could carry on as if the data were independent. Unfortunately, strict exogeneity is an unrealistic assumption for time series that only applies in a small number of situations at best (Stock and Watson, 2007).

To clarify what Burman and Nolan (1992) and Bergmeir et al. (2018) do, their arguments are with respect to what is known in the nonparametric literature as the correlation effect in the variance of the MASE decomposition (Altman, 1990, Hart, 1991, Opsomer et al., 2001). This is separate from the bias that is coming from $\frac{1}{T}\sum_{i=1}^{T}\widetilde{\mu}_{-i}(\alpha)\varepsilon_i$. Burman and Nolan (1992) and Bergmeir et al. (2018) assume that

$$E(\frac{1}{T}\sum_{i=1}^{T}\widetilde{\varepsilon}^2_{-i}(\alpha)) = \sigma^2 + \frac{1}{T}\sum_{i=1}^{T}E(\mu_i - \widetilde{\mu}_{-i}(\alpha))^2,$$

which we know is not true in general since $E[\frac{1}{T}\sum_{i=1}^{T}\widetilde{\mu}_{-i}(\alpha)\varepsilon_i] \neq 0$ in general, even if the residuals follow a martingale-like structure.[6,7] Burman and Nolan (1992) and Bergmeir et al. (2018) then decompose $E(\mu_i - \widetilde{\mu}_{-i}(\alpha))^2$ via the standard bias-variance decomposition of the estimator. That is,

$$E(\mu_i - \widetilde{\mu}_{-i}(\alpha)) = \underbrace{E([\widetilde{\mu}_{-i}(\alpha) - E(\widetilde{\mu}_{-i}(\alpha)]^2)}_{variance} + \underbrace{(E(\widetilde{\mu}_{-i}(\alpha)) - \mu_i)^2}_{squared\,bias}.$$

What Burman and Nolan (1992) and Bergmeir et al. (2018) analyze is the impact that er-

[6] Burman and Nolan (1992) write the proofs in terms of the full sample estimates $E(\frac{1}{T}\sum_{i=1}^{T}\hat{\varepsilon}^2_i(\alpha)) = \sigma^2 + \frac{1}{T}\sum_{i=1}^{T}E(\mu_i - \hat{\mu}_i(\alpha))^2$, as opposed to using the LOO estimates as I do, but the same arguments still hold since $\frac{1}{T}\sum_{i=1}^{T}\hat{\varepsilon}^2_i(\alpha)$ can be decomposed the same way as $\frac{1}{T}\sum_{i=1}^{T}\widetilde{\varepsilon}^2_{-i}(\alpha)$ with $\widetilde{\mu}_{-i}(\alpha)$ being replaced by $\hat{\mu}_i(\alpha)$. It is still the case that $E[\frac{1}{T}\sum_{i=1}^{T}\hat{\mu}_i(\alpha)\varepsilon_i] \neq 0$ in general.

[7] Bergmeir et al. (2018) assume this formula is approximately true instead of being exactly true.

rors have on the variance component in the bias-variance decomposition of $E(\mu_i - \widetilde{\mu}_{-i}(\alpha))$ when they follow a martingale-like structure versus when they do not. It is debatable whether one should even care about this impact if what one actually cares about is choosing the model with the smallest ASE or MASE since the magnitude of the ASE and MASE already account for the impact or lack there from the errors. It is also besides the point since even if there is no correlation effect on the variance component in the bias-variance decomposition, $E(\frac{1}{T}\sum_{i=1}^{T}\widetilde{\mu}_{-i}(\alpha)\varepsilon_i) \neq 0$ in general for finite samples, so CV with time series data will be biased in general.

4 Conclusion

I show that contrary to conventional wisdom, model errors following a martingale-like structure is not sufficient for model selection via CV to be unbiased for time series data. If one is concerned with the finite sample bias, the block CV methods of Chu and Marron (1991) and Burman et al. (1994) which were designed to alleviate the impact of the bias or alternative methods (see Opsomer et al. (2001) or Arlot and Celisse (2010) for reviews) may be better alternatives in terms of finite sample performance.

References

Altman, N. S. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association 85*(411), 749–759. 2, 4, 6

Anatolyev, S. (2005). Gmm, gel, serial correlation, and asymptotic bias. *Econometrica 73*(3), 983–1002. 5

Arlot, S. and A. Celisse (2010). A survey of cross-validation procedures for model selection. *Statistical Survey 4*, 40–79. 2, 7

Bao, Y. and A. Ullah (2007). The second-order bias and mean squared error of estimators in time-series models. *Journal of Econometrics 140*(2), 650–669. 5

Bergmeir, C., R. J. Hyndman, and B. Koo (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics Data Analysis 120*, 70–83. 2, 3, 6

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods* (Second ed.). Springer-Verlag New York. 6

Burman, P., E. Chow, and D. Nolan (1994). A cross-validatory method for dependent data. *Biometrika 81*(2), 351–358. 2, 7

Burman, P. and D. Nolan (1992). Data-dependent estimation of prediction functions. *Journal of Time Series Analysis 13*(3), 189–207. 2, 3, 6

Chu, C.-K. and J. S. Marron (1991). Comparison of two bandwidth selectors with dependent errors. *The Annals of Statistics 19*(4), 1906–1918. 7

Coulombe, P. G., M. Leroux, D. Stevanovic, and S. Surprenant (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics 37*(5), 920–964. 2

Gelman, A., J. Hwang, and A. Vehtari (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing 24*, 997–1016. 2

Hart, J. D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society. Series B 52*(1), 173–187. 2, 4, 6

Kendall, M. G. (1954). Note on bias in the estimation of autocorrelation. *Biometrika 41*(3/4), 403–404. 5

Lusompa, A. (2025). Regression model selection under general conditions. *Federal Reserve Bank of Kansas City Working Paper No. RWP 25-12*. 5

Marriott, F. H. C. and J. A. Pope (1954). Bias in the estimation of autocorrelations. *Biometrika 41*(3/4), 390–402. 5

Opsomer, J., Y. Wang, and Y. Yang (2001). Nonparametric regression with correlated errors. *Statistical Science 16*(2), 134–153. 2, 3, 4, 6, 7

Petropoulos, F., D. Apiletti, V. Assimakopoulos, M. Z. Babai, D. K. Barrow, S. B. Taieb, C. Bergmeir, R. J. Bessa, J. Bijak, J. E. Boylan, J. Browell, C. Carnevale, J. L. Castle, P. Cirillo, M. P. Clements, C. Cordeiro, F. L. C. Oliveira, S. D. Baets, A. Dokumentov, J. Ellison, P. Fiszeder, P. H. Franses, D. T. Frazier, M. Gilliland, M. S. Gönül, P. Goodwin, L. Grossi, Y. Grushka-Cockayne, M. Guidolin, M. Guidolin, U. Gunter, X. Guo, R. Guseo, N. Harvey, D. F. Hendry, R. Hollyman, T. Januschowski, J. Jeon, V. R. R. Jose, Y. Kang, A. B. Koehler,

S. Kolassa, N. Kourentzes, S. Leva, F. Li, K. Litsiou, S. Makridakis, G. M. Martin, A. B. Martinez, S. Meeran, T. Modis, K. Nikolopoulos, D. Önkal, A. Paccagnini, A. Panagiotelis, I. Panapakidis, J. M. Pavía, M. Pedio, D. J. Pedregal, P. Pinson, P. Ramos, D. E. Rapach, J. J. Reade, B. Rostami-Tabar, M. Rubaszek, G. Sermpinis, H. L. Shang, E. Spiliotis, A. A. Syntetos, P. D. Talagala, T. S. Talagala, L. Tashman, D. Thomakos, T. Thorarinsdottir, E. Todini, J. R. T. Arenas, X. Wang, R. L. Winkler, A. Yusupova, and F. Ziel (2022). Forecasting: theory and practice. *International Journal of Forecasting 38*(3), 705–871. 2

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. Academic Press. 6

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics 99*(1), 39–61. 2

Shaman, P. and R. A. Stine (1988). The bias of autoregressive coefficient estimators. *Journal of the American Statistical Association 83*(403), 842–848. 5

Stock, J. and M. Watson (2007). *Introduction to Econometrics*. Addison Wesley Longman. 6

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25. 5

White, H. (1994). *Estimation, inference and specification analysis*. Cambridge University Press. 5