The Canadian Journal of Statistics | La revue canadienne de statistique
**Vol. , No. , , 1–2**

# On the Bias of Cross-Validated Variance Estimators in Small Samples

**Ada Lovelace** @ , **Second Author**
Department of Statistics, Example University, Canada

**Abstract** We study the finite-sample bias of $k$-fold cross-validated estimators of prediction error variance under small sample sizes typical of applied survey work. Simulation results show that the bias is negligible for $k \geq 10$ but can exceed 15% for $k = 5$ when the sample size is below 50, and we propose a simple bias-correction factor that removes most of this effect.

**Résumé** Nous étudions le biais à distance finie des estimateurs de la variance de l'erreur de prévision validés par validation croisée à $k$ plis, pour des tailles d'échantillon réduites typiques des enquêtes appliquées. Les résultats de simulation montrent que ce biais est négligeable pour $k \geq 10$, mais peut dépasser 15 % pour $k = 5$ lorsque la taille de l'échantillon est inférieure à 50 ; nous proposons un facteur de correction simple qui élimine la majeure partie de cet effet.

**Keywords** cross-validation; variance estimation; small-sample bias; resampling.

**@ Corresponding author** ada@example.edu

## 1 Introduction

Cross-validation is widely used to estimate out-of-sample prediction error, but its variance estimator is known to be biased in small samples (Efron and Tibshirani, 1997). We revisit this bias for $k$-fold cross-validation with small $k$ and small total sample size $n$, a regime common in applied survey statistics where data collection is expensive.

## 2 Simulation Study

We simulated 10,000 replicate datasets under a linear model with Gaussian errors, for $n \in \{20, 50, 100\}$ and $k \in \{5, 10, n\}$ (leave-one-out). Table 1 reports the empirical bias of the cross-validated variance estimator relative to its Monte Carlo target.

**Table 1:** Empirical relative bias (%) of the CV variance estimator.

| $n$ | $k = 5$ | $k = 10$ | $k = n$ (LOO) |
|---|---|---|---|
| 20 | 18.4 | 6.1 | 2.0 |
| 50 | 11.2 | 3.4 | 0.9 |
| 100 | 4.7 | 1.2 | 0.3 |

## 3 A Bias-Correction Factor

Following the approach of Davison and Hinkley (1997), we propose a correction factor $c(n, k) = 1 + (k - 1)/n$ applied multiplicatively to the raw cross-validated variance estimate. Applying this correction to the $k = 5$ simulations reduces the residual bias to under 3% across all sample sizes considered.

*Licence statement to appear here.*

## 4 Discussion

The correction proposed here is simple to compute and requires no additional resampling, making it attractive for routine use in applied work with small $k$ and small $n$.

## References

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.