When Should You Adjust Standard Errors for Clustering—Simulation Study with a Continuous Covariate

Masashi Yoshioka

May 22, 2022

1. Introduction

A traditional textbook cross section inference often assumes random sampling of variables. That is, dependent/independent variables are often assumed to be independent and identically distributed. This assumption is unlikely to hold when there is a group structure. For example, children in the same class or citizens in the same state are expected to have similar characteristics in common (e.g., family background, industry structure), which violates the random sampling assumption. In this situation, clustering adjustments are commonly used. Moulton (1986) suggests a cluster adjustment and applies it to empirical problems such as hedonic house pricing. Liang and Zeger (1986) propose an adjusted variance estimator that allows for an unrestricted within-cluster covariance matrix.

This paper is inspired by Abadie et al. (2017, 2021). They write: "contrary to common wisdom, correlations between residuals within clusters are neither necessary, nor sufficient, for cluster adjustments to matter. Similarly, correlations between regressors within clusters are neither necessary, not sufficient, for cluster adjustments to matter or to justify clustering." (Abadie et al., 2017, p.2). They argue that clustering is essentially a design problem, either a sampling design or an experimental design. Abadie et al. (2021) propose a new variance estimator that nests the standard heteroskedasticity-robust variance estimator and the cluster-robust variance estimator suggested by Liang and Zeger (1986). However, since Abadie et al. (2017, 2021) assume a model with a binary covariate, it is not clear whether their discussion applies to a general case.

In this paper, I simulate a model with a continuous covariate and show that the discussion of Abadie et al. (2017, 2021) could apply to the case with a continuous covariate. Following Abadie et al. (2021), I do not consider design issues and just assume that all the clusters are observed (I would need other sources to infer the fraction of observed clusters). I show that the heteroskedasticity-robust variance estimator can be too small and the cluster-robust variance estimator too large under some conditions. Then, I propose an appropriate clustering adjustment based on the simulation results. In particular, I suggest in which dimensions we should adjust for clustering. I also examine the asymptotic performance of the cluster-robust variance estimator depending on whether the number of clusters is fixed or large.

2. Simulation Setting

I replicate a similar simulation to Abadie et al. (2021) with a continuous covariate. I assume samples of N units and M clusters. For simplicity, the cluster size N_m is assumed to be identical: $N_m = N/M$. Each unit i is assumed to belong to a cluster $m_i \in \{1, \ldots, M\}$. I define a continuous covariate W_{mi}, the distribution of which is given by

$$W_{mi} = \mu_W + \zeta_m + \xi_{mi},$$

where μ_W is a constant, $\zeta_m \sim N(0, \sigma_\zeta^2)$ and $\xi_{mi} \sim N(0, \sigma_\xi^2)$. Note that ζ_m is a cluster-specific error whereas ξ_{mi} is an individual error. Since $\text{Var}(W_{mi}) = \text{Var}(\zeta_m) + \text{Var}(\xi_{mi}) = \sigma_\zeta^2 + \sigma_\xi^2$ and $\text{Cov}(W_{mi}, W_{mj}) = \text{Var}(\zeta_m) = \sigma_\zeta^2$ for $i \neq j$, the within-cluster correlation of W_{mi} is given by $\rho_W = \sigma_\zeta^2/(\sigma_\zeta^2 + \sigma_\xi^2)$.

First, I assume the following fixed coefficient model:

$$Y_{mi} = \alpha + \beta W_{mi} + u_{mi}, \tag{1}$$

where the error term u_{mi} can be separated into a cluster-specific error and an individual error $u_{mi} = \eta_m + \varepsilon_{mi}$ with $\eta_m \sim N(0, \sigma_\eta^2)$ and $\varepsilon_{mi} \sim N(0, \sigma_\varepsilon^2)$. Thus, the within-cluster correlation of u_{mi} is given by $\rho_u = \sigma_\eta^2/(\sigma_\eta^2 + \sigma_\varepsilon^2)$.

I run an OLS on equation (1) and compute the following two variance estimates:

$$\hat{V}^{EHW} = \frac{\sum_{i=1}^N \hat{u}_{mi}^2 (W_{mi} - \overline{W})^2}{\left(\sum_{i=1}^N (W_{mi} - \overline{W})^2\right)^2} \text{ and } \hat{V}^{LZ} = \frac{\sum_{l=1}^M \left(\sum_{i|m_i=l} \hat{u}_{mi}(W_{mi} - \overline{W})\right)^2}{\left(\sum_{i=1}^N (W_{mi} - \overline{W})^2\right)^2}.$$

$\hat{V}^{EHW}$ is the standard heteroskedasticity-robust variance estimator or Eicker–Huber–White (EHW) variance estimator. $\hat{V}^{LZ}$ is Liang–Zeger (LZ) cluster-robust variance estimator. I also compute a "true" variance of $\hat{\beta}$ by a Monte Carlo simulation and compare it with EHW and LZ variance estimators.

Subsequently, I consider a random coefficient model

$$Y_{mi} = \alpha + \beta_{mi} W_{mi} + u_{mi},$$

where $\beta_{mi} = \beta + \upsilon_m + \nu_{mi}$ with $\upsilon_m \sim N(0, \sigma_\upsilon^2)$ and $\nu_{mi} \sim N(0, \sigma_\nu^2)$. The within-cluster correlation of β_{mi} is given by $\rho_\beta = \sigma_\upsilon^2/(\sigma_\upsilon^2 + \sigma_\nu^2)$. W_{mi} and u_{mi} are defined in the same way as in the fixed coefficient model. The OLS gives us the estimate for β.

The simulation process can be described as follows. First, I generate the coefficients β_{mi} by picking υ_m and ν_{mi} randomly from $N(0, \sigma_\upsilon^2)$ and $N(0, \sigma_\nu^2)$, respectively. I assume $\alpha \equiv 1$ and $\beta \equiv 2$. In the case of the fixed coefficient, I assume $\text{Var}(\beta_{mi}) = \sigma_\upsilon^2 + \sigma_\nu^2 \equiv 0$, which corresponds to $\beta_{mi} = \beta$ for all $m = 1, \ldots, M$ and $i = 1, \ldots, N$. In the case of the random coefficient, I assume $\text{Var}(\beta_{mi}) = \sigma_\upsilon^2 + \sigma_\nu^2 \equiv 1$ and assume different values of $\rho_\beta \in [0, 1]$ to see how the results would change.

After determining the coefficients, I repeatedly ($R = 3,000$ times) generate variables W_{mi} and u_{mi} by picking ζ_m, ξ_{mi}, η_m and ε_{mi} randomly from respective normal distributions. For simplicity, I assume $E[W_{mi}] = \mu_W \equiv 2$, $\text{Var}(W_{mi}) = \sigma_\zeta^2 + \sigma_\xi^2 \equiv 1$ and $\text{Var}(u_{mi}) = \sigma_\eta^2 + \sigma_\varepsilon^2 \equiv 1$. Then, I generate outcomes Y_{mi} by $Y_{mi} = \alpha + \beta_{mi} W_{mi} + u_{mi}$. The different values of $\rho_W \in [0, 1]$ and $\rho_u \in [0, 1]$ are assumed to examine if within-cluster correlation of covariate and that of error term matter for standard error estimation. For each iteration, I obtain the OLS estimate $\hat{\beta}_r, r = 1, \ldots, R$ with two kinds of standard errors se^{EHW} and se^{LZ} based on $\hat{V}^{EHW}$ and $\hat{V}^{LZ}$, respectively. Then, I compare the standard deviation of the simulated values of $\hat{\beta}_r$ with the average of those of se^{EHW} and the average of those of se^{LZ}.

In addition, I try two different cases where the number of clusters M is fixed and where it is large. The former corresponds to $N = 250,000$, $M = 50$ and $N_m = 5,000$ and the latter $N = 250,000$, $M = 5,000$ and $N_m = 50$. These experiments could provide some insight into how two kinds of asymptotics matter, i.e., $N_m \to \infty$ and $M \to \infty$, for the accuracy of standard error estimation.

3. Results

3.1 Fixed Coefficient Model

First, I consider a fixed coefficient model $Y_{mi} = \alpha + \beta W_{mi} + u_{mi}$. I assume different values of $\rho_W \in \{0, 0.5, 1\}$ and $\rho_u \in \{0, 0.5, 1\}$. Tables 1 and 2 summarize the results. Standard errors are shown in bold if they seem to be correct henceforth.

Tables 1 and 2 show that se^{LZ} always estimates the standard deviation relatively correctly regardless of the within-correlations ρ_W and ρ_u. When $\rho_W = 0$ **or** $\rho_u = 0$, se^{EHW} is quite close to se^{LZ} and also estimates the standard deviation correctly. However, when $\rho_W \neq 0$ **and** $\rho_u \neq 0$, it underestimates the true standard deviation.

We can also see that se^{LZ} performs better when M is large (Table 2). When M is fixed (Table 1), it tends to underestimate the true standard deviation. Figures 1 and 2 show the ratio of $\text{se}^{LZ}/\text{std}(\hat{\beta}_r)$ for different values of N_m and M, respectively. Figure 1 assumes $M = 50$ and Figure 2 assumes $N_m = 50$. These figures graphically show that se^{LZ} converges to $\text{std}(\hat{\beta}_r)$ as M gets larger (Figure 2) but does not with large N_m (Figure 1).

3.2 Random Coefficient Model

3.2.1 No Correlation of Coefficient Within Clusters ($\rho_\beta = 0$)

Now I consider a random coefficient model $Y_{mi} = \alpha + \beta_{mi} W_{mi} + u_{mi}$. First, I consider the case where β_{mi} is uncorrelated within clusters, i.e., $\text{Var}(\beta_{mi}) = 1$ and $\text{Cov}(\beta_{mi}, \beta_{lj}) = 0$ for $i \neq j$ **even if** $m = l$. Tables 3 and 4 summarize the results. These tables show roughly the same implications as in the fixed coefficient model. That is, se^{LZ} estimates the standard deviation more correctly than se^{EHW} regardless of the values of ρ_W and ρ_u. se^{EHW} can be also correct when $\rho_W = 0$ or $\rho_u = 0$, but otherwise it underestimates the true standard deviation. In terms of asymptotics, se^{LZ} converges to true value when M is large, but fails to do so when M is fixed.

3.2.2 Some Correlation of Coefficient Within Clusters ($\rho_\beta = 0.5$)

Now I consider a random coefficient model with $\rho_\beta = 0.5$. Tables 5 and 6 show the results. The implications are quite different from those of the fixed coefficient model now. The point is that se^{EHW} can be better than se^{LZ} under some conditions unlike the fixed coefficient model where se^{LZ} always outperforms se^{EHW}. Specifically, we can observe that:

1. The within-cluster correlation of the covariate ρ_W matters for the appropriate choice of standard errors. When ρ_W is close to 0, se^{EHW} estimates the standard deviation correctly whereas se^{LZ} overestimates it. On the other hand, when ρ_W is close to 1, se^{LZ} is relatively correct whereas se^{EHW} underestimates the true standard deviation. Both fail to estimate the true standard deviation correctly when ρ_W is neither small nor large.

2. The within-cluster correlation of error term ρ_u does not matter for the relative performances of se^{EHW} and se^{LZ}.

3. se^{LZ} still tends to underestimate the true standard deviation when M is fixed.

3.2.3 Perfect Correlation of Coefficient Within Clusters ($\rho_\beta = 1$)

Finally, I consider a random coefficient model where β_{mi} is cluster-specific, or equivalently, perfectly correlated coefficient within clusters, i.e., $\rho_\beta = 1$. Tables 7 and 8 show the results. The conclusions are basically the same as when $\rho_\beta = 0.5$ (Tables 5 and 6). That is, we should use se^{EHW} when ρ_W is close to 0 and se^{LZ} when ρ_W is close to 1. Unfortunately, both are incorrect when ρ_W is neither small nor large.

3.2.4 Relative Performance of se^{EHW} and se^{LZ}

I consider only the fixed M, i.e., $N = 250,000$, $M = 50$ and $N_m = 5,000$ henceforth. In order to evaluate the relative accuracy of se^{EHW} and se^{LZ}, I define a relative performance parameter λ such that

$$\text{std}(\hat{\beta}_r)^2 = (1-\lambda)(\text{se}^{EHW})^2 + \lambda(\text{se}^{LZ})^2.$$

By design, when λ is close to 0, se^{EHW} is relatively correct. When λ is close to 1, se^{LZ} is relatively correct.

I plot the values of λ for different values of $\rho_W \in \{0, 0.1, \ldots, 1\}$ and $\rho_u \in \{0, 0.1, \ldots, 1\}$ when $\rho_\beta = 1$. Figure 3 shows the values of λ across ρ_W and Figure 4 shows across ρ_u. Figure 3 implies that the relative performance parameter λ is close to 0 when ρ_W is small and gets closer to 1 as ρ_W gets larger. On the other hand, Figure 4 shows that the relative performance appears to be insensitive to the value of ρ_u.

Comparing the case of $\rho_\beta = 0.5$ (Tables 5 and 6) with that of $\rho_\beta = 1$ (Tables 7 and 8), the differences between se^{EHW} and se^{LZ} appear to get larger as ρ_β gets larger. Figure 5 shows this tendency graphically by computing standard errors for different values of $\rho_\beta \in \{0.1, 0.2, \ldots, 1\}$. However, even when se^{LZ} gets farther from se^{EHW}, their relative performance does not apparently depend on ρ_β. Figure 6 shows the values of λ for different values of $\rho_\beta \in \{0.1, 0.2, \ldots, 1\}$. The figure shows that λ is stable.

4. Conclusion

The simulation results give us insight into when and how to adjust standard errors for clustering with a continuous covariate. First of all, when the coefficient is fixed, we can always use Liang–Zeger cluster-robust variance estimator for samples that have a group structure. The cluster-robust variance estimator works even when clusters do not matter, i.e., either covariate or error term is uncorrelated within clusters.

However, if the coefficient is random and correlated within clusters, Liang–Zeger cluster-robust variance estimator can be too conservative. The within-cluster correlation of covariate ρ_W matters in this situation. If ρ_W is close to 0, the Eicker–Huber–White heteroskedasticity-robust variance estimator will be correct. If ρ_W is close to 1, Liang–Zeger cluster-robust variance estimator will be correct. If ρ_W is neither small nor large, both estimators will fail. On the other hand, the within-cluster correlation of error term ρ_u does not matter as far as I have examined. This observation suggests the possibility of a new variance estimator that nests EHW and LZ estimators with weights that depend on ρ_W.

Researchers are often tempted to use LZ variance estimator when there is a group structure in samples. However, it is not a good idea to always use LZ estimator because it can lead to an unnecessarily conservative estimate. The appropriate standard error depends on within-cluster correlation of covariate ρ_W. Thus, I propose one potential strategy for clustering adjustments as follows:

1. Estimate the within-cluster correlation of ρ_W by $\mathrm{Var}(\overline{W}_m)/\mathrm{Var}(W_{mi})$ where $\overline{W}_m \equiv \sum_{i|m_i=m} W_{mi}/N_m$ is a cluster-averages of covariates.

2. If the estimate $\hat{\rho}_W$ is close enough to 0, use EHW estimator. If $\hat{\rho}_W$ is close enough to 1, use LZ estimator.

3. If $\hat{\rho}_W$ is neither small nor large, report both estimators.

Note that I have assumed all the clusters to be observed in the simulations. The above strategy applies only when researchers can be confident that all the clusters are observed. Thus, for example, it is obviously wrong to assume all the samples have their own unique clusters, which leads to $\hat{\rho}_W = 1$ and suggests always using LZ estimator based on the strategy. The strategy can provide insight into in which dimensions we should adjust for clustering. If researchers suspect that a certain group structure (e.g., gender, race etc.) matters for standard error estimation, they can first estimate ρ_W assuming that cluster. If the estimate is large enough, it is worth considering adjusting for the cluster.

Future work could involve identifying a new standard error formula that can be used regardless of the value of ρ_W or applying my proposed strategy to the empirical analysis.

References

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). *When should you adjust standard errors for clustering?* (NBER Working Paper No. 24003). National Bureau of Economic Research. https://www.nber.org/papers/w24003

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2021, October 8). *When should you adjust standard errors for clustering?* [Conference presentation]. The Gary Chamberlain Online Seminar in Econometrics. https://www.chamberlainseminar.org/past-seminars/autumn-2021#h.i4t9eedzteeb

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22.

Moulton, B. R. (1986). Random group effects and the precision of regression estimates. *Journal of Econometrics*, 32(3), 385–397.

Table 1: Fixed coefficient model when M is fixed

	$\rho_u = 0$	$\rho_u = 0.5$	$\rho_u = 1$
$\rho_W = 0$	*std: .0020* **EHW: .0020** **LZ: .0020**	*std: .0020* **EHW: .0020** **LZ: .0019**	*std: .0020* **EHW: .0020** **LZ: .0019**
$\rho_W = 0.5$	*std: .0020* **EHW: .0020** **LZ: .0020**	*std: .0698* EHW: .0020 **LZ: .0672**	*std: .1013* EHW: .0020 **LZ: .0954**
$\rho_W = 1$	*std: .0020* **EHW: .0021** **LZ: .0019**	*std: .1021* EHW: .0020 **LZ: .0967**	*std: .1436* EHW: .0019 **LZ: .1378**

Table 2: Fixed coefficient model when M is large

	$\rho_u = 0$	$\rho_u = 0.5$	$\rho_u = 1$
$\rho_W = 0$	*std: .0020* **EHW: .0020** **LZ: .0020**	*std: .0020* **EHW: .0020** **LZ: .0020**	*std: .0020* **EHW: .0020** **LZ: .0020**
$\rho_W = 0.5$	*std: .0020* **EHW: .0020** **LZ: .0020**	*std: .0073* EHW: .0020 **LZ: .0073**	*std: .0102* EHW: .0020 **LZ: .0101**
$\rho_W = 1$	*std: .0020* **EHW: .0020** **LZ: .0020**	*std: .0100* EHW: .0020 **LZ: .0101**	*std: .0138* EHW: .0020 **LZ: .0141**

Table 3: Random coefficient model with $\rho_\beta = 0$ when M is fixed

	$\rho_u = 0$	$\rho_u = 0.5$	$\rho_u = 1$
$\rho_W = 0$	*std: .0053* **EHW: .0057** **LZ: .0055**	*std: .0054* **EHW: .0057** **LZ: .0056**	*std: .0052* **EHW: .0056** **LZ: .0056**
$\rho_W = 0.5$	*std: .0053* **EHW: .0057** **LZ: .0055**	*std: .0700* EHW: .0057 **LZ: .0673**	*std: .1011* EHW: .0056 **LZ: .0950**
$\rho_W = 1$	*std: .0052* **EHW: .0057** **LZ: .0049**	*std: .1049* EHW: .0057 **LZ: .0978**	*std: .1444* EHW: .0057 **LZ: .1383**

Table 4: Random coefficient model with $\rho_\beta = 0$ when M is large

	$\rho_u = 0$	$\rho_u = 0.5$	$\rho_u = 1$
$\rho_W = 0$	*std: .0053* **EHW: .0057** **LZ: .0057**	*std: .0052* **EHW: .0057** **LZ: .0057**	*std: .0053* **EHW: .0057** **LZ: .0056**
$\rho_W = 0.5$	*std: .0053* **EHW: .0056** **LZ: .0056**	*std: .0088* EHW: .0056 **LZ: .0090**	*std: .0114* EHW: .0057 **LZ: .0114**
$\rho_W = 1$	*std: .0052* **EHW: .0057** **LZ: .0057**	*std: .0112* EHW: .0056 **LZ: .0114**	*std: .0147* EHW: .0057 **LZ: .0151**

Table 5: Random coefficient model with $\rho_\beta = 0.5$ when M is fixed

	$\rho_u = 0$	$\rho_u = 0.5$	$\rho_u = 1$
$\rho_W = 0$	*std: .0053* **EHW: .0056** LZ: .0975	*std: .0051* **EHW: .0056** LZ: .0964	*std: .0055* **EHW: .0059** LZ: .1100
$\rho_W = 0.5$	*std: .1436* EHW: .0054 LZ: .1610	*std: .2016* EHW: .0061 LZ: .2194	*std: .1627* EHW: .0052 LZ: .1744
$\rho_W = 1$	*std: .2492* EHW: .0055 **LZ: .2459**	*std: .2906* EHW: .0057 **LZ: .2779**	*std: .3106* EHW: .0057 **LZ: .3020**

Table 6: Random coefficient model with $\rho_\beta = 0.5$ when M is large

Large M	$\rho_u = 0$	$\rho_u = 0.5$	$\rho_u = 1$
$\rho_W = 0$	*std: .0054* **EHW: .0057** LZ: .0115	*std: .0053* **EHW: .0057** LZ: .0115	*std: .0052* **EHW: .0056** LZ: .0113
$\rho_W = 0.5$	*std: .0165* EHW: .0057 LZ: .0195	*std: .0182* EHW: .0057 LZ: .0208	*std: .0191* EHW: .0057 LZ: .0217
$\rho_W = 1$	*std: .0254* EHW: .0057 **LZ: .0269**	*std: .0269* EHW: .0057 **LZ: .0286**	*std: .0289* EHW: .0057 **LZ: .0307**

Table 7: Random coefficient model with $\rho_\beta = 1$ when M is fixed

	$\rho_u = 0$	$\rho_u = 0.5$	$\rho_u = 1$
$\rho_W = 0$	*std: .0050* **EHW: .0054** LZ: .1345	*std: .0050* **EHW: .0053** LZ: .1318	*std: .0051* **EHW: .0055** LZ: .1372
$\rho_W = 0.5$	*std: .2325* EHW: .0058 LZ: .2618	*std: .2388* EHW: .0056 LZ: .2636	*std: .2315* EHW: .0052 LZ: .2537
$\rho_W = 1$	*std: .3678* EHW: .0055 **LZ: .3614**	*std: .3287* EHW: .0049 **LZ: .3279**	*std: .4404* EHW: .0061 **LZ: .4345**

Table 8: Random coefficient model with $\rho_\beta = 1$ when M is large

	$\rho_u = 0$	$\rho_u = 0.5$	$\rho_u = 1$
$\rho_W = 0$	*std: .0054* **EHW: .0056** LZ: .0150	*std: .0053* **EHW: .0057** LZ: .0151	*std: .0053* **EHW: .0057** LZ: .0152
$\rho_W = 0.5$	*std: .0226* EHW: .0057 LZ: .0269	*std: .0240* EHW: .0057 LZ: .0280	*std: .0251* EHW: .0057 LZ: .0287
$\rho_W = 1$	*std: .0355* EHW: .0056 **LZ: .0371**	*std: .0356* EHW: .0057 **LZ: .0387**	*std: .0369* EHW: .0057 **LZ: .0401**

Figure 1: $\mathrm{se}^{LZ}/\mathrm{std}$ across N_m when $\rho_W = \rho_u = 1$

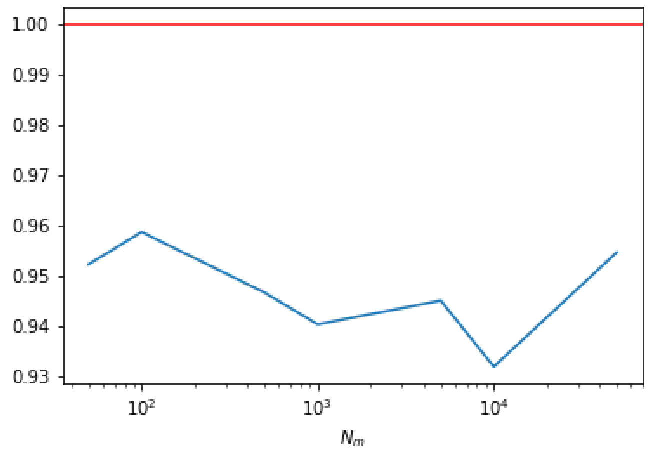

Figure 2: $\mathrm{se}^{LZ}/\mathrm{std}$ across M when $\rho_W = \rho_u = 1$

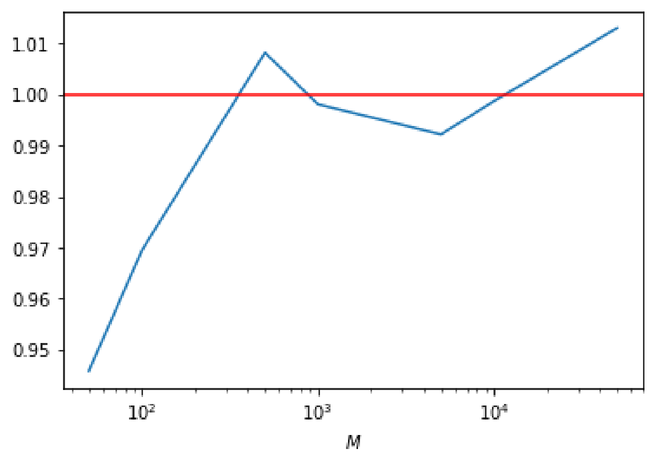

Figure 3: The relative performance λ across ρ_W when $\rho_\beta = 1$ and $\rho_u = 1$

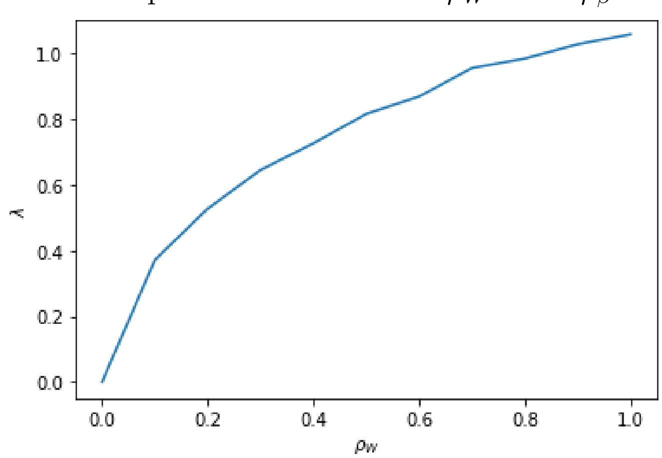

Figure 4: The relative performance λ across ρ_u when $\rho_\beta = 1$ and $\rho_W = 0.5$

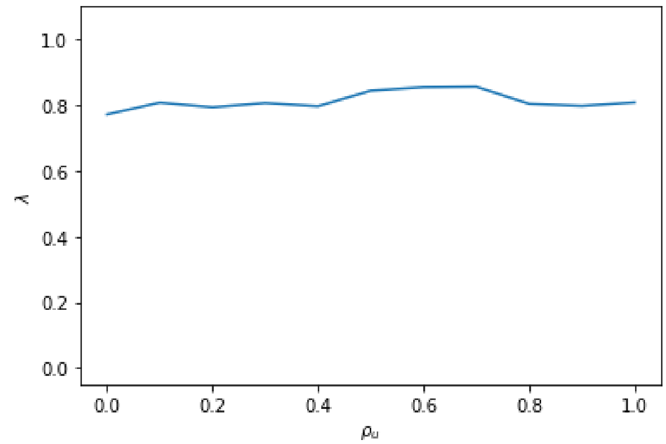

Figure 5: Standard errors across ρ_β when $\rho_W = 0.2$ and $\rho_u = 0$

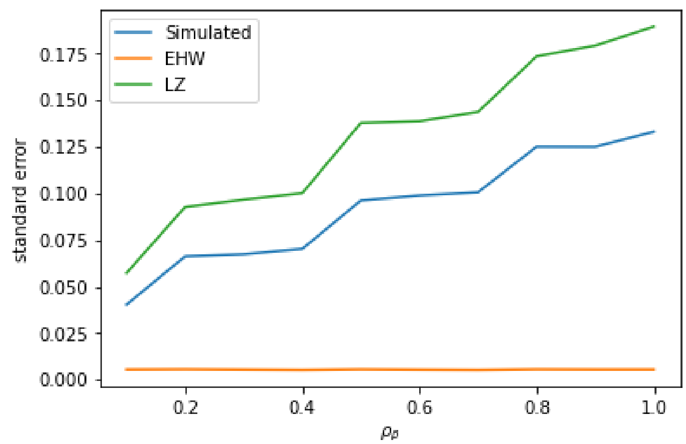

Figure 6: The relative performance λ across ρ_β when $\rho_W = 0.2$ and $\rho_u = 0$