Final Project Report: Endogenous Stratification in Randomized Experiments

STA 640
Gaojia Xu, Guanqi Zeng

April 24, 2022

1 Introduction

Randomized experiments in the social sciences field have become more and more popular in recent years, as these types of experiments can eliminate the effect of unobserved difference between the treatment group and the control group that may not be able to explained by the study. In addition, for policy makers, as they are typically more interested in the effect of the treatment on particular groups, for example, those who are most in need, subgroups are often created based on the predicted outcome of the units without treatment. In other words, the potential outcomes are generated based on the model built by the out-of-sample untreated units. These predicted outcomes can be seen as the estimated outcomes have the units are untreated, and thus are used as the metrics to stratify the experimental units for average treatment effect calculation.

However, in most of the social sciences experiments settings, such a reliable model may not be obtained. Therefore, a similar approach, **endogenous stratification**, is employed. Instead of using the out-of-sample data, endogenous stratification applies in-sample-data to build the predictive outcome model. This method first regresses the outcome variable on the baseline covariates using all units from the experimental control group, and then uses the regression coefficients to predict the potential outcomes without treatment for all experimental units. These predicted outcomes are then sorted and split into intervals (e.g. bottom, medium, and top) to create strata for the units. Finally, the average treatment effects are calculated within each stratus.

As endogenous stratification incorporates the information and the relationship between the outcome variable and the covariates for full sample experimental control units, the average treatment effect is biased with a predicted pattern due to the predicted outcome generation process. As demonstrated by Abadie et. al. in 2018, the ATE of the bottom group has a positive bias while the ATE of the top group has a negative bias.

In this project, we demonstrated the bias problem by implementing endogenous stratification on the Project Star data set and compared the performance of ATE estimation of the two other endogenous stratification methods, leave-one-out and sample splitting, proposed by Abadie et. al. In addition, a third endogenous stratification method, 10 Fold, is proposed and implemented on the STAR data set. Then, we tested the bias problem on a STAR-based

simulated data set with these four methods. Finally, we explored the effect of sample size and the number of covariates on the performance of these four methods on the computer-generated data set. Possible explanations of the cause of the bias issue is discussed along with the algorithm explanation of these four methods.

2 Methods

In this section, we explain the algorithms of the four endogenous stratification methods: full sample endogenous stratification, leave-one-out endogenous stratification, sample splitting endogenous stratification, and 10-fold endogenous stratification. Then, we describe the process of implementing these four methods on the Project STAR data set, the STAR-based simulation data set, and the computer-generated simulation data set.

2.1 Algorithm

2.1.1 Full Sample Endogenous Stratification

The full sample endogenous stratification uses the full sample units in the experimental control group to generate predicted outcome model. With this model, the predicted outcome for all experimental units are predicted and used for stratification. The mathematical formulation is as follows:
1. Get coefficients for predicted outcomes by using the full sample units:

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{N} \boldsymbol{x}_i(1-w_i)\boldsymbol{x}_i^{'})^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i(1-w_i)y_i$$

where w_i is the indicator of the treatment group.
2. Obtain the predicted outcomes without treatment for all experimental units, $\boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}$ and sort by ascending order.
3. Based on the predicted outcomes quantiles $(\frac{1}{3}, \frac{2}{3})$, split the experimental units into three strata, c_k for $j = 1, 2$, and calculate the unadjusted ATE for each stratus:

$$\hat{\tau_k} = \frac{\sum_{i=1}^{N} y_i I_{[w_{i=1}, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}} \leq c_k]}}{\sum_{i=1}^{N} I_{[w_{i=1}, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}} \leq c_k]}} - \frac{\sum_{i=1}^{N} y_i I_{[w_{i=0}, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}} \leq c_k]}}{\sum_{i=1}^{N} I_{[w_{i=0}, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}} \leq c_k]}}$$

For adjusted ATE, we regress the outcome variables on the covariates for each stratification. The coefficient of the treatment is the adjusted ATE for that stratus.

An intuitive explanation of the production of bias problem can be seen in the predicted outcome generating process. As we use the full sample of control units for predicting the predicted outcome of the control units, the result is actually overfitted. Denote the coefficients for the untreated units in the finite sample $\boldsymbol{\beta}$, the residual $e_i = y_i - \boldsymbol{x}_i^{'}\boldsymbol{\beta}$, the untreated observations with large negative values for the residuals tend to be overfitted, as we expect the predicted outcome of the control units in the experimental group underestimates the predicted outcome of the finite untreated group, producing negative bias. These observations

are thus pushed to the lower interval of the predicted outcomes. The positive bias is produced similarly for those observation with large positive values of the residuals.

2.1.2 Leave-one-out Endogenous Stratification

The leave-one-out endogenous stratification uses the leave-out-out coefficients of the predicted outcome model to generate predicted outcomes without treatment for the experimental control units. The predicted outcomes without treatment for the experimental treatment units, however, are generated using the full sample predicted outcome model demonstrated in Section 2.1.1. The justification is for the predicted outcomes without treatment of the experimental treatment units, the estimation is in fact an 'out-of-sample' estimation, as none of the units in the treatment group is employed for the estimation. The leave-one-out predicted outcomes for the experimental control units are calculated as follows:

$$\boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}_{(-1)} = \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}} - \frac{h_{N_i}}{1-h_{N_i}}(y_i - \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}})$$

where $\hat{\boldsymbol{\beta}}$ is the coefficients of the full sample predicted outcome model, and h_{N_i} is the leverage. The unadjusted ATE is calculated as follows:

$$\hat{\tau_k}^{LOO} = \frac{\sum_{i=1}^{N} y_i I_{[w_{i=1}, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}} \leq c_k]}}{\sum_{i=1}^{N} I_{[w_{i=1}, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}} \leq c_k]}} - \frac{\sum_{i=1}^{N} y_i I_{[w_{i=0}, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}_{(-1)} \leq c_k]}}{\sum_{i=1}^{N} I_{[w_{i=0}, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}_{(-1)} \leq c_k]}}$$

As the τ_k^{LOO} adjusted the predicted outcome of the experimental control units by predicting each of them with the rest of the experimental control units, it prevents overfitting and thus alleviating the bias problem.

2.1.3 Sample Splitting Endogenous Stratification

The sample splitting endogenous stratification repeatedly uses a random subset of half of the experimental control units as the prediction group to estimate the coefficients of the model for predicted outcome without treatment. The rest of the control units are the estimation group whose predicted outcomes without treatment are estimated by the model. When calculating the ATE, the prediction group are not included. Suppose we repeat the sample splitting for M times, the coefficients of the m^{th} predicted outcome model is calculated as follows:

$$\hat{\boldsymbol{\beta}}_m = (\sum_{i=1}^{N} \boldsymbol{x}_i(1-w_i)(1-v_{im})\boldsymbol{x}_i^{'})^{-1} \sum_{i=1}^{N} \boldsymbol{x}_i(1-w_i)(1-v_{im})y_i$$

and the unadjusted ATE is calculated by averaging the ATEs for each stratification:

$$\hat{\tau}_{km}^{SS} = \frac{\sum_{i=1}^{N} y_i I_{[w_{i=1}, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}_m \leq c_k]}}{\sum_{i=1}^{N} I_{[w_{i=1}, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}_m \leq c_k]}} - \frac{\sum_{i=1}^{N} y_i I_{[w_{i=0}, v_{im}=1, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}_m \leq c_k]}}{\sum_{i=1}^{N} I_{[w_{i=0}, v_{im}=1, c_{k-1} < \boldsymbol{x}_i^{'}\hat{\boldsymbol{\beta}}_m \leq c_k]}}$$

$$\hat{\tau}_k^{RSS} = \frac{1}{M}\sum_{m=1}^{M}\hat{\tau}_{km}^{SS}$$

The sample splitting method prevents the overfitting problem by only including a part of the control group for predicted outcome model and exclude these observations in the ATE calculation. By averaging all the ATEs, the variance of the estimation should also be reduced.

2.1.4 10-Fold Endogenous Stratification

10-Fold endogenous stratification is similar to leave-one-out method, except that for each time, we take a size of $\frac{9}{10}$ of the control group units to fit the predicted outcome model and use the rest $\frac{1}{10}$ for estimation of the predicted outcome without treatment for the control group. Compared to the leave-one-out method, the variance of the estimated ATE is expected to be smaller. The coefficients of the predicted outcome model is calculated as follows:

$$\hat{\boldsymbol{\beta}}_{f=j} = (\sum_{i=1}^{N}\boldsymbol{x}_i(1-w_i)I(i \notin fold_j)\boldsymbol{x}_i^{'})^{-1}\sum_{i=1}^{N}\boldsymbol{x}_i(1-w_i)I(i \notin fold_j)y_i$$

The unadjusted ATE is calculated as follows:

$$\hat{\tau_k}^{10fold} = \frac{\sum_{i=1}^{N} y_i I_{[w_{i=1}, c_{k-1}<\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\leq c_k]}}{\sum_{i=1}^{N} I_{[w_{i=1}, c_{k-1}<\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}\leq c_k]}} - \frac{\sum_{i=1}^{N} y_i I_{[w_{i=0}, c_{k-1}<\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{(10fold)}\leq c_k]}}{\sum_{i=1}^{N} I_{[w_{i=0}, c_{k-1}<\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_{(10fold)}\leq c_k]}}$$

2.2 Data Sets

In this section, we introduce the data sets on which we implement the four methods. The first data set is the Project STAR data set, and the rest two data sets are simulated.

The STAR data set collects both standardized math test scores and related covariates of 3764 students in 79 Tennessee schools, who were randomly assigned to 3 class settings, small, regular-size and regular-size with a teacher's aide. The related covariates include African-American, female, eligibility for the free lunch program, and school attended. Our project only uses data in former two types of class setting.

2.3 Simulation: STAR-based Simulation

The STAR-based simulation data set is generated based on a linear model with standard normal errors. First, we obtain the outcome variable of the original STAR data. Then, we regress the outcome on the control group baseline covariates to obtain the model coefficients and the residuals. With the residuals, we obtain the estimated standard error of the model. With the original STAR data, we eliminated the original outcome and treatment, and instead

reassign the new treatment groups and the new control groups with the same sizes as the original STAR data.

For each round of simulation, a bootstrap sample of the STAR-based simulated data is selected. Then, we generate the outcome variable by adding the standard normal noise scaled by the estimated standard error of the previous model to the fitted values. The outcome variables, the treatment variable, and the bootstrap covariates are combined as our simulated data set.

2.4 Simulation: Computer-generated Simulation

To demonstrate how bias of different estimators change with different combinations of sample sizes and numbers of covariates, we implement the four methods of endogeous stratification on a computer-generated data set. The data is generated by following model:

$$y_i = 1 + \sum_{l=1}^{40} z_{li} + v_i$$

where v_i has independent normal distribution with variance equal to 60 and z_{li} has independent standard normal distribution. Samples are randomized into treated or control group with equal sizes.

3 Results and Conclusion

In this section, we comment on the results and make conclusions. The plots and tables are attached in appendix.

The unadjusted and adjusted default ATE estimation results are shown in the Figure 1. The unadjusted effect of small classes is 0.1659 and the adjusted effect of small classes is 0.1892. We present the standard error of both estimations using the Neyman formula of variance and output from the linear regression. Thus, we see both estimators are significant at the 5

Compared to the default estimation values, the full-sample endogenous stratification estimator $\hat{\tau}_k$ (Figure 2) generates approximately doubled values in the low and the medium group estimation. Digging into the $\hat{\tau}_k$ estimation in different groups, the high group always produces negative estimations, unlike the other two groups. This indicates assigning to a small class can be harmful to students predicted to have high math scores when in a regular class, which is counterintuitive. Similarly, $\hat{\tau}_k^{LOO}$ and $\hat{\tau}_k^{10fold}$ also have negative signs for high group students but are not statistically significant. $\hat{\tau}_k^{RSS}$ has positive signs for high group students but is not significant as well. Generally, the low and medium groups produce positive and significant effects of the small class setting.

From the STAR-based simulation, we obtain the distribution of 4 types of estimations in both unadjusted and adjusted form (Figure 3, 4). For the $\hat{\tau}_k$, in both forms, the bias lie around ± 0.5, substantial to its estimation value. Also we have a clear view that the distribution of bottom group (solid line) is the right-most one whereas top group (dot-dashed line) is the left-most. Even though the middle group distribution centered at 0, the wide gap between

groups indicate the full-sample estimator is negatively biased. The other 3 estimators, $\hat{\tau}_k^{LOO}$, $\hat{\tau}_k^{RSS}$ and $\hat{\tau}_k^{10fold}$ all centered at 0 and $\hat{\tau}_k^{LOO}$, $\hat{\tau}_k^{10fold}$ are more stable.

In the bias table (Figure 5, 6, 7) of all estimators, full-sample estimator $\hat{\tau}_k$ leads to the highest bias in low group and high group. Moreover, coverage rates of this estimator results in most distortion in low gorup and high group as well. Moreover, $\hat{\tau}_k^{LOO}$, $\hat{\tau}_k^{10fold}$ always have lowest RMSE value while $\hat{\tau}_k$, $\hat{\tau}_k^{RSS}$ generates higher ones.

Leveraging the computer-generated dataset, we perform simulation according to different combination of K number of covariates and N number of observations and the bias are displayed in the table (Figure 8). $\hat{\tau}_k$ produces higher bias when sample size decreases or number of regressor is higher, which has the tendency to overfit. The bias of the other 3 estimators are consistently smaller than $\hat{\tau}_k$ in most cases. Bias of leave-one-out estimator $\hat{\tau}_k^{LOO}$ is higher than $\hat{\tau}_k^{RSS}$ and $\hat{\tau}_k^{10fold}$ and $\hat{\tau}_k^{RSS}$ has the lowest bias.

Thus $\hat{\tau}_k$ is the worst estimator with highest bias, lowest coverage and comparably high RMSE due to the overfitting problem on control sample. $\hat{\tau}_k^{LOO}$, $\hat{\tau}_k^{RSS}$ and $\hat{\tau}_k^{10fold}$ generally produces neglectable bias. $\hat{\tau}_k^{LOO}$, $\hat{\tau}_k^{10fold}$ have higher coverage and lower RMSE than $\hat{\tau}_k^{RSS}$ if we do the simulation on estimators. In high dimensional and low sample setting, where overfitting can easily occur, $\hat{\tau}_k^{RSS}$ performs better than $\hat{\tau}_k^{LOO}$ and $\hat{\tau}_k^{10fold}$ in most of times. On the contrary, when overfitting is not a problem, $\hat{\tau}_k^{LOO}$ can execute badly as $\hat{\tau}_k$.

4 Citation

Alberto Abadie, Matthew M. Chingos, Martin R. West; Endogenous Stratification in Randomized Experiments. The Review of Economics and Statistics 2018; 100 (4): 567–580. doi: https://doi.org/10.1162/rest$_{a0}$0732

Appendix

STAR Default Estimation Results

	Unadjusted	Adjusted
$\hat{\tau}$	0.1659	0.1892
se $\hat{\tau}$	0.0334	0.0295

Figure 1: Unstratified ATE

STAR Estimation Results

	Unadj low	Unadj medium	Unadj high	Adj low	Adj medium	Adj high
$\hat{\tau}_k$	0.3705	0.2688	-0.1330	0.3908	0.3023	-0.1242
se $\hat{\tau}_k$	0.0593	0.0618	0.0636	0.0562	0.0621	0.0583
$\hat{\tau}_k^{LOO}$	0.3277	0.2499	-0.0486	0.3440	0.2730	-0.0660
se $\hat{\tau}_k^{LOO}$	0.0603	0.0659	0.0661	0.0565	0.0648	0.0611
$\hat{\tau}_k^{RSS}$	0.1829	0.2083	0.2203	0.1737	0.1933	0.2754
se $\hat{\tau}_k^{RSS}$	0.0446	0.0429	0.0427	0.0389	0.0363	0.0365
$\hat{\tau}_k^{10fold}$	0.3226	0.2809	-0.0440	0.3201	0.2789	-0.0586
se $\hat{\tau}_k^{10fold}$	0.0611	0.0634	0.0649	0.0569	0.0613	0.0604

Figure 2: Endogenous Stratification ATE

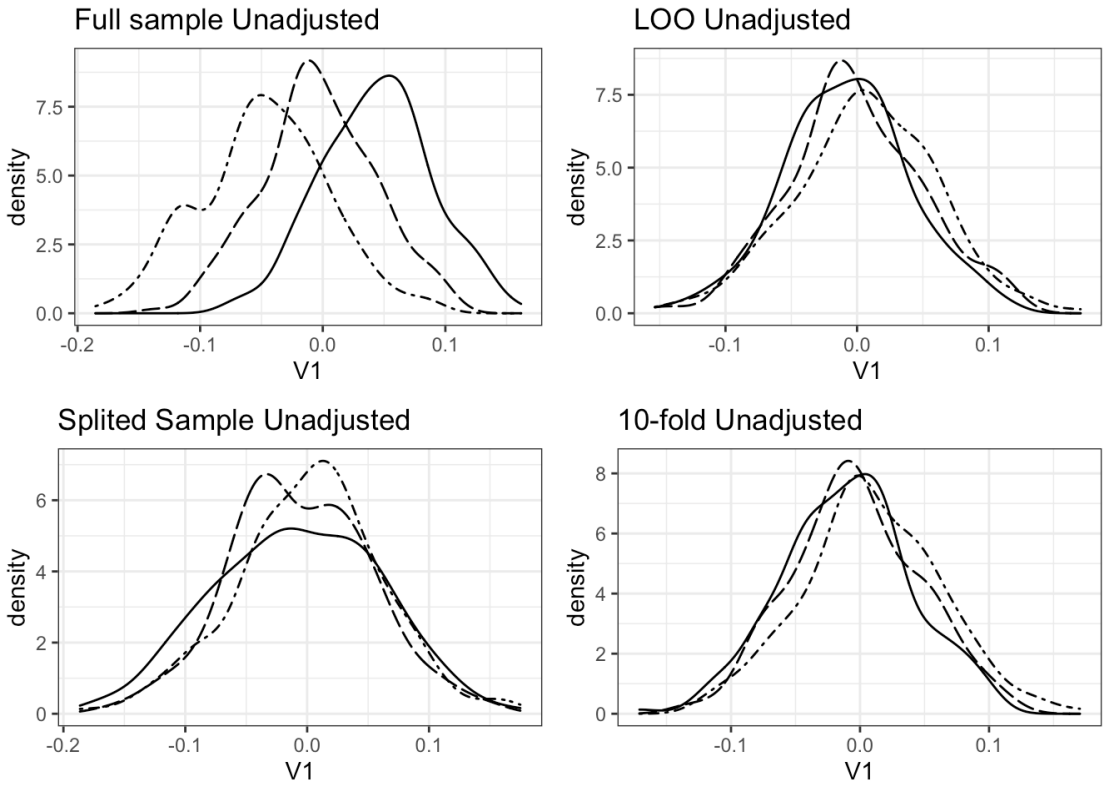

Figure 3: STAR Simulation Unadjusted ATEs

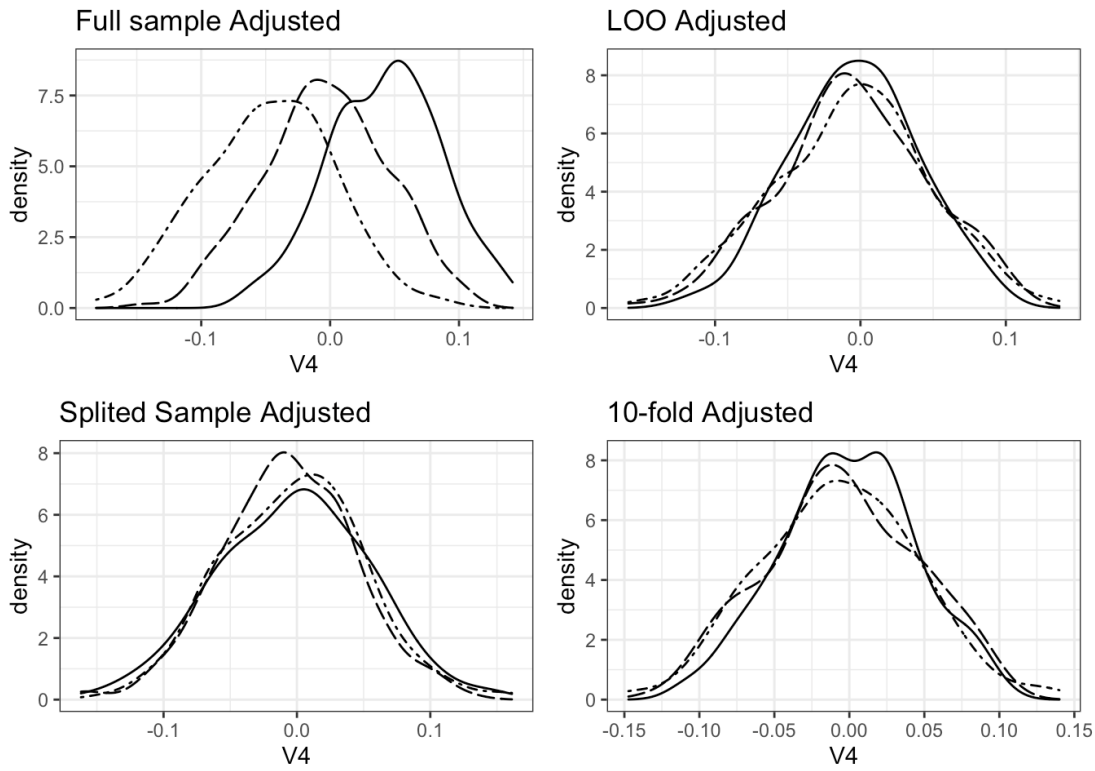

Figure 4: STAR Simulation Adjusted ATEs

Bias in STAR Simulations

	Unadjusted			Adjusted		
	Low	**Medium**	**High**	**Low**	**Medium**	**High**
$\hat{\tau}_k$	0.04	0.00	-0.05	0.04	0.00	-0.05
$\hat{\tau}_k^{LOO}$	-0.01	0.00	0.01	0.00	-0.01	-0.01
$\hat{\tau}_k^{RSS}$	-0.01	-0.01	0.00	0.00	-0.01	0.00
$\hat{\tau}_k^{10fold}$	-0.01	0.00	0.01	0.00	-0.01	-0.01

Figure 5: STAR Simulation Bias

Coverage in STAR Simulations

	Unadjusted			Adjusted		
	Low	**Medium**	**High**	**Low**	**Medium**	**High**
$\hat{\tau}_k$	0.86	0.94	0.80	0.88	0.94	0.80
$\hat{\tau}_k^{LOO}$	0.96	0.94	0.93	0.98	0.96	0.94
$\hat{\tau}_k^{RSS}$	0.90	0.92	0.93	0.95	0.98	0.96
$\hat{\tau}_k^{10fold}$	0.94	0.96	0.93	0.98	0.96	0.94

Figure 6: STAR Simulation Coverage

RMSE in STAR Simulations

	Unadjusted			Adjusted		
	Low	**Medium**	**High**	**Low**	**Medium**	**High**
$\hat{\tau}_k$	0.06	0.05	0.07	0.06	0.05	0.07
$\hat{\tau}_k^{LOO}$	0.05	0.05	0.05	0.04	0.05	0.05
$\hat{\tau}_k^{RSS}$	0.07	0.06	0.06	0.06	0.05	0.05
$\hat{\tau}_k^{10fold}$	0.05	0.05	0.05	0.04	0.05	0.05

Figure 7: STAR Simulation RMSE

Bias in Simulations Using Artificial Data

	K = 10						K = 20						K = 40					
	Unadjusted			Adjusted			Unadjusted			Adjusted			Unadjusted			Adjusted		
	Low	**Medium**	**High**	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**	**Low**	**Medium**	**High**
N=200 $\hat{\tau}_k$	2.54	-0.24	-2.38	2.52	-0.30	-2.32	3.37	0.12	-2.92	3.19	-0.15	-2.80	4.72	-0.08	-4.08	4.34	0.01	-3.83
N=200 $\hat{\tau}_k^{LOO}$	-0.17	-0.01	0.11	0.13	-0.08	-0.02	-0.01	-0.05	0.55	0.31	0.08	-0.10	0.47	0.25	-0.20	1.24	0.15	-1.10
N=200 $\hat{\tau}_k^{RSS}$	-0.01	-0.08	0.13	-0.08	-0.03	0.16	0.32	0.21	0.43	0.25	0.07	0.15	0.14	0.16	0.34	0.19	-0.10	0.11
N=200 $\hat{\tau}_k^{10fold}$	-0.12	-0.19	0.29	0.38	-0.22	-0.16	-0.21	0.15	0.64	0.38	-0.11	-0.15	0.09	0.28	0.05	1.40	0.52	-1.42
N=1000 $\hat{\tau}_k$	0.34	-0.22	-0.43	0.36	-0.22	-0.45	0.78	-0.11	-0.73	0.76	-0.17	-0.76	0.83	0.02	-0.94	0.83	0.05	-0.88
N=1000 $\hat{\tau}_k^{LOO}$	-0.30	-0.21	0.22	-0.27	-0.21	0.19	0.01	-0.04	-0.02	-0.02	-0.10	-0.06	0.01	0.00	-0.14	0.04	0.04	-0.10
N=1000 $\hat{\tau}_k^{RSS}$	0.06	-0.08	-0.20	0.07	-0.10	-0.19	0.04	0.10	-0.15	0.09	0.02	-0.20	-0.17	-0.13	0.04	-0.09	-0.09	-0.05
N=1000 $\hat{\tau}_k^{10fold}$	-0.35	-0.16	0.20	-0.27	-0.14	0.13	-0.04	-0.06	0.03	0.02	-0.12	-0.08	-0.04	0.00	-0.05	0.04	0.01	-0.08
N=5000 $\hat{\tau}_k$	0.06	-0.01	-0.06	0.06	-0.02	-0.05	0.11	-0.05	-0.11	0.11	-0.06	-0.12	0.21	-0.01	-0.26	0.22	0.00	-0.23
N=5000 $\hat{\tau}_k^{LOO}$	-0.06	-0.01	0.06	-0.06	-0.02	0.07	-0.05	-0.05	0.04	-0.05	-0.05	0.03	0.05	-0.01	-0.09	0.06	-0.01	-0.06
N=5000 $\hat{\tau}_k^{RSS}$	0.01	-0.03	-0.04	0.03	-0.03	-0.02	-0.02	-0.08	0.02	-0.04	-0.11	0.02	-0.05	-0.07	0.00	-0.01	-0.07	-0.04
N=5000 $\hat{\tau}_k^{10fold}$	-0.08	0.01	0.06	-0.07	0.00	0.06	-0.05	-0.07	0.07	-0.04	-0.07	0.05	0.05	-0.03	-0.08	0.06	-0.01	-0.06

Figure 8: Computer Simulation Bias