ONLINE APPENDIX
Parallel Experimentation and Competitive Interference on Online Advertising Platforms*

Published in *Marketing Science*

Caio Waisman Navdeep S. Sahni Harikesh S. Nair Xiliang Lin

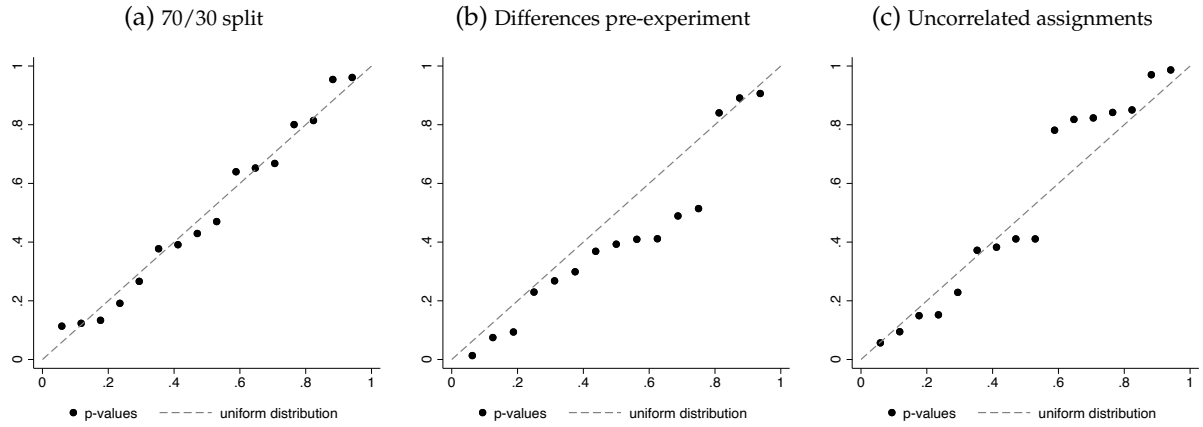
August 2, 2024

* E-mail addresses for correspondence: caio.waisman@kellogg.northwestern.edu (Waisman), navdeep.sahni@stanford.edu (Sahni), hsnair@gmail.com (Nair) or xilianglin@gmail.com (Lin).

A Randomization checks

This section shows a series of randomization checks to ensure that the experiment was implemented correctly. First, for all campaigns 70% of users should have been assigned to the treatment group. To test whether this was the case and handle multiple testing, we perform a proportion test for each campaign, collect the associated p-values and use a KS statistic to test whether their distribution corresponds to a standard uniform distribution, as should be the case. We display this comparison in Figure A.1a. The p-value associated with this KS test for this hypothesis is 0.986, indicating that the test/control splits were correct.

Figure A.1: KS test for p-values of fractions of users in treatment groups



Note: Graphs show quantiles plots of 16, 15 and 16 p-values, respectively, computed using standard errors robust to heteroskedasticity. The p-values of KS tests for whether these p-values are drawn from a standard uniform distribution are 0.986, 0.173 and 0.428, respectively.

We now proceed to investigate possible differences in user behavior before the experiment as a function of whether they belong to the treatment versus the control group of each campaign. If randomization was implemented correctly, there should not be significant differences across groups. To verify whether this is the case, for each campaign, we run regressions of the number of adds to cart, of orders, of visits to the product detail page, and GMV during the three days prior to the start of the experiment on the treatment assignment indicator. The regression coefficients along with standard errors robust to heteroskedasticity are shown in Table A.1. Out of the 64 coefficients, 54 are not statistically significant at the 10% level.

To account for multiple testing, we estimate a SUR model for each campaign, gather the p-values associated with a test for whether all coefficients were equal to zero, and test

Table A.1: Regressions of behavioral variables three days before experiment on treatment

Campaign	# of adds to cart	GMV	# of orders	# of visits	Campaign	# of adds to cart	GMV	# of orders	# of visits
1	-0.0108	-0.1035	0.0008	-0.0121	9	0.0003	0.1096	0.0001	-0.0030
($n = 128,576$)	(0.0073)	(0.3025)	(0.0014)	(0.0180)	($n = 539,702$)	(0.0007)	(0.2891)	(0.0001)	(0.0080)
2	0.0002	-0.0007	0.0002	0.0001	10	0.0025	0.6120	0.0001	0.0226
($n = 535,343$)	(0.0004)	(0.0046)	(0.0001)	(0.0012)	($n = 66,824$)	(0.0014)*	(0.4016)	(0.0001)	(0.0209)
3	0.0094	0.3232	0.0021	0.0144	11	0.0007	-0.0118	0.0004	0.0041
($n = 12,809$)	(0.0059)	(0.2576)	(0.0036)	(0.0349)	($n = 130,071$)	(0.0009)	(0.0254)	(0.0007)	(0.0046)
4	0.0001	0.0010	0.0001	0.0128	12	-0.0047	-0.5814	0.0009	-0.0099
($n = 14,846$)	(0.0001)	(0.0010)	(0.0001)	(0.0071)*	($n = 14,806$)	(0.0137)	(0.9264)	(0.0052)	(0.0855)
5	-0.0006	-0.0777	-0.0003	0.0024	13	0.0059	1.9995	0.0005	0.0357
($n = 179,516$)	(0.0012)	(0.1393)	(0.0006)	(0.0052)	($n = 111,307$)	(0.0034)*	(1.6644)	(0.0004)	(0.0856)
6	0.0013	0.0131	0.0009	0.0028	14	0.0239	0.3669	0.0032	0.0459
($n = 184,689$)	(0.0006)**	(0.0069)*	(0.0004)**	(0.0015)*	($n = 20,907$)	(0.0137)*	(0.4281)	(0.0043)	(0.0660)
7	0.0003	-0.1681	-0.0001	0.0026	15	0.0168	1.4439	0.1330	-0.0294
($n = 302,471$)	(0.0005)	(0.3101)	(0.0002)	(0.0071)	($n = 66,984$)	(0.0104)	(0.6600)**	(0.0663)***	(0.0270)
8	-0.0068	-0.1054	-0.0001	-0.0804	16	-0.0001	0.0025	0.0002	0.0004
($n = 131,743$)	(0.0044)	(2.0914)	(0.0003)	(0.0564)	($n = 339,018$)	(0.0003)	(0.0018)	(0.0001)	(0.0008)

Note: Each cell contains the coefficient associated with the treatment indicator from a linear regression. The unit of observation is a user. Dependent variables are events three days prior to the experiment. Standard errors robust to heteroskedasticity are shown between parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

whether they correspond to draws from a standard uniform distribution using a KS test.[1] Results are shown in Figure A.1b. The p-value associated with the KS test is 0.173, so that we cannot reject the null hypothesis that the p-values follow a standard uniform distribution at the usual significance level.

Finally, we assess whether users were indeed allocated to treatment groups independently across campaigns. To do so, for each campaign we run a regression of an indicator for whether the user is in its treatment group on an indicator for whether the user is in its main competitor's treatment group as determined in Table 2. We then collect all 16 p-values, computed using standard errors robust to heteroskedasticity, and perform a KS test for whether they follow a standard uniform distribution, which should be the case if the treatment assignments are uncorrelated. Results are displayed in Figure A.1c. The p-value associated with this KS test is 0.428, indicating that users were indeed allocated to treatment groups independently across campaigns.

B Alternative estimators

As we noted, our estimands of interest are conditional average treatment effect parameters, so any method that can consistently estimate conditional expectations is applicable in our setting. The use of nonparametric and kernel-based methods to estimate treatment effect parameters is not novel, as attested by Imbens (2004). These methods are often employed to treat continuous variables nonparametrically, which could be incorporated into

[1] We only obtain 15 p-values because the variance matrix associated with the SUR estimates for campaign 4 was not invertible.

our framework by including user- and/or target audience-level variables to the analysis.

Traditional kernel-based methods applied to continuous variables in conjunction with tools that smooth over discrete variables, such as Li et al. (2013), have also been applied. For example, Li et al. (2009) smooth over both continuous and categorical variables to estimate the propensity score and then use it to provide an asymptotically efficient estimator of the unconditional ATE. However, unlike us they do not address estimation of $CATE$s. In a setup where these are the objects of interest and where there are continuous variables over which the econometrician needs to smooth, the approach proposed by Racine and Li (2004), which allows for smoothing over categorical and continuous variables, can be followed. However, the rate of convergence of this estimator is affected by the bandwidths associated with the continuous variables, making it slower.

While the kernel-based estimator we propose has several attractive properties, it also has shortcomings. As noted in page 555 of Li et al. (2013), situations in which the number of variables that are smoothed increases with the sample size are not considered. This is especially relevant with user- and target audience-level variables, which can be numerous, and when many advertisers experiment in parallel. Thus, alternative machine learning methods can become an attractive option.

Given the linear structure of our estimation problem, as shown in equation (5), methods that rely on or apply to such structure are especially attractive. For binary dependent variables, Imai and Ratkovic (2013) provide a method that combines Support Vector Machines with separate LASSO constraints to perform regularization and variable selection, which would be especially appropriate if the outcome of interest was an indicator for purchase, for example. In turn, Tian et al. (2014) combine LASSO constraints with a simple covariate modification; given the correct specification of the regression equation, maximum likelihood estimation can be implemented, ensuring asymptotic efficiency.

Assuming sparsity, Athey et al. (2018) propose approximate residual balancing as a way to debias penalized regression adjustments so that methods like the LASSO and elastic net can be used to perform inference on $CATE$ parameters in high-dimensional settings. Importantly, this method does not require the propensity score to be known or estimable, even though in settings such as ours treatment assignment is randomly determined according to known probabilities. In turn, Bradic et al. (2019) allow the regression equation not be sparse provided that the propensity score is while assuming that it is given by the usual logistic formula. We note that these restrictions on the propensity are usually

satisfied in randomized experiments such as the ones we consider.

Other studies propose methods to deal with potential misspecification of the outcome equation or the propensity score and achieve double robustness in the estimation of conditional average treatment effect parameters, such as Tan (2020) and Ning et al. (2020). Misspecification, however, is not a concern in our setting.

C Simulations

We now present simulations to illustrate of the performance and properties of the estimator introduced by Li et al. (2013). We further compare its performance to that of OLS and to the method proposed by Ye et al. (2024). For simplicity, these simulations focus on scenarios with only three advertisers.

C.1 Setup

In these simulation exercises, we consider the following model:

$$Y_i = \beta_0 + \beta_1 D_f + \beta_2 D_g + \beta_3 D_h + \beta_4 D_f D_g + \beta_5 D_f D_h + \beta_6 D_g D_h + \beta_7 D_f D_g D_h + \epsilon, \quad \text{(C.1)}$$

where: Y is the outcomes of interest; the βs are the coefficients to be estimated, which we collect in a vector, $\boldsymbol{\beta}$; D_ℓ is an indicator for whether the user was in advertiser ℓ's treatment group, in which case they were eligible to be exposed to ℓ's ad, and where $\ell \in \{f, g, h\}$; and ϵ is the error term. For all simulations, $\epsilon \sim N(0, 1)$.

Our objective is to estimate the effects of f's advertising, and whether and how they vary depending on whether f's competitors, g and h, are advertising.

C.2 Data generating processes (DGPs)

We consider five different DGPs, that is, five different vectors $\boldsymbol{\beta}$ in our simulations. The values we use are taken from the results we obtained in our application of the kernel-based estimator, which we present in Section 6.4. We consider situations where only two competitors could be experimenting. We chose specific combinations to create different types of competitive advertising environments. The specific values for $\boldsymbol{\beta}$ we used are given in Table C.1 below.

When f faces no competition, f's $CATE$ is 0.16. The DGPs are such that:

Table C.1: Distribution of users across campaign exposures

Coefficients	DGPs				
	1	2	3	4	5
β_0	0.09	0.09	0.09	0.09	0.09
β_1	0.16	0.16	0.16	0.16	0.16
β_2	0.00	0.00	0.03	0.00	0.00
β_3	0.00	0.05	0.05	0.05	0.01
β_4	0.00	0.00	-0.13	-0.10	-0.10
β_5	0.00	-0.15	-0.15	-0.15	-0.02
β_6	0.00	0.00	-0.07	-0.01	-0.01
β_7	0.00	0.00	0.23	0.09	0.02

- DGP 1: g and h do not impact f's $CATE$ either unilaterally or combined.
- DGP 2: unilaterally, g does not impact f's $CATE$ but h does as the $CATE$ goes from 0.16 to 0.01. With g and h combined, f's $CATE$ is also 0.01.
- DGP 3: g unilaterally reduces f's $CATE$ to 0.03 and h to 0.01. Combined, however, the effects partially cancel each other out, and f's $CATE$ becomes 0.11.
- DGP 4: g unilaterally reduces f's $CATE$ to 0.06 and h to 0.01. When combined, the effects act together and f's $CATE$ becomes 0.
- DGP 5: g unilaterally reduces f's $CATE$ to 0.06 and h only to 0.14. When combined, g's negative impact is not altered by h and the $CATE$ remains 0.06.

C.3 Estimators

We consider and compare three estimators. First, we consider OLS. Second, we consider the kernel-based estimator from Li et al. (2013). Third, we consider the estimator proposed by Ye et al. (2024). Given the absence of covariates other than the treatments and the functional form implied by Assumption 1 from Ye et al. (2024), their estimator collapses to nonlinear least squares (NLS) applied to the following model:

$$\begin{aligned} Y_i = \theta_0 + \theta_0\theta_1 D_f + \theta_0\theta_2 D_g + \theta_0\theta_3 D_h + \theta_0\theta_1\theta_2 D_f D_g + \theta_0\theta_1\theta_3 D_f D_h \\ + \theta_0\theta_2\theta_3 D_g D_h + \theta_0\theta_1\theta_2\theta_3 D_f D_g D_h + \eta. \end{aligned} \tag{C.2}$$

This estimator can also be interpreted simply as OLS applied to (C.1) under specific nonlinear constraints between the parameters. Notice that these constraints are only true under DGP 1.

C.4 Data sets

In the simulations, the Ds are randomly assigned with equal probability independently from one another. We consider three sample sizes—80, 800, and 8,000—to assess how the performance of the estimators vis-à-vis the number of coefficients and how it changes as the sample size increases. We choose these specific values so that the number of observations for each of the eight possible full treatment assignments is the same. This corresponds to allocating observations to each advertiser's treatment or control groups equiprobably and independently across advertisers.

For each DGP-sample size combination, we create 10,000 data sets and implement the three aforementioned estimators on each data set. The results, which we present and detail below, are obtained from these estimates.

We separate our results in two parts. First, we show the performance of the kernel-based in detail because it is the one we use in the main manuscript. Then, we compare its performance to that of OLS and Ye et al. (2024).

C.5 Performance of kernel-based estimator

Consistency The first property of the kernel-based estimator we verify is consistency: as the sample size grows, the estimates of the coefficients should approximate their true values more and more. Tables C.2-C.6 display, for each coefficient-DGP-sample size, the average and standard deviation of the estimates over the 10,000 data sets. Expectedly, given Theorem 2 from Li et al. (2013), the larger the sample size, the closer the estimates are to the true values of the coefficients. This patterns holds for all coefficients and DGPs.

Bandwidths We now assess the behavior of the bandwidths. As explained in Li et al. (2013), whenever a variable is irrelevant, the bandwidth associated with it should be one; otherwise, the bandwidth should converge to zero as the sample size increases. Thus, we should observe the following:

- DGP 1: both bandwidths should be one.
- DGP 2: the bandwidths associated with g and h should be one and zero, respectively.
- DGP 3: both bandwidths should be zero.
- DGP 4: both bandwidths should be zero.

Table C.2: Average bias and standard deviation of kernel estimates under DGP 1

Coefficients	True Value	n	mean($\hat{\beta}^s_{k,n}$)	SD($\hat{\beta}^s_{k,n}$)
β_0	0.09	80	0.098	0.229
		800	0.1	0.074
		8,000	0.101	0.025
β_1	0.16	80	0.18	0.324
		800	0.181	0.106
		8,000	0.181	0.038
β_2	0	80	-0.0004	0.222
		800	0.0007	0.068
		8,000	-0.0002	0.022
β_3	0	80	-0.004	0.216
		800	0.001	0.067
		8,000	-0.0001	0.02
β_4	0	80	0.004	0.303
		800	-0.0004	0.095
		8,000	-0.0001	0.03
β_5	0	80	-0.004	0.306
		800	0.0001	0.094
		8,000	0.0001	0.029
β_6	0	80	-0.002	0.199
		800	-0.0005	0.06
		8,000	-0.0002	0.019
β_7	0	80	0.003	0.406
		800	0.001	0.124
		8,000	0.0003	0.039

Table C.3: Average bias and standard deviation of kernel estimates under DGP 2

Coefficients	True Value	n	mean($\hat{\beta}^s_{k,n}$)	SD($\hat{\beta}^s_{k,n}$)
β_0	0.09	80	0.113	0.23
		800	0.117	0.078
		8,000	0.108	0.029
β_1	0.16	80	0.134	0.329
		800	0.143	0.118
		8,000	0.17	0.04
β_2	0	80	-0.0005	0.222
		800	0.001	0.069
		8,000	-0.0002	0.021
β_3	0.05	80	0.029	0.222
		800	0.03	0.082
		8,000	0.048	0.034
β_4	0	80	0.003	0.303
		800	-0.001	0.097
		8,000	-0.0001	0.031
β_5	-0.15	80	-0.08	0.324
		800	-0.088	0.135
		8,000	-0.144	0.057
β_6	0	80	-0.001	0.202
		800	-0.001	0.066
		8,000	-0.0003	0.026
β_7	0	80	0.029	0.413
		800	0.031	0.141
		8,000	0.048	0.057

Table C.4: Average bias and standard deviation of kernel estimates under DGP 3

Coefficients	True Value	n	mean($\hat{\beta}^s_{k,n}$)	SD($\hat{\beta}^s_{k,n}$)
β_0	0.09	80	0.118	0.23
		800	0.121	0.078
		8,000	0.118	0.033
β_1	0.16	80	0.11	0.328
		800	0.113	0.116
		8,000	0.152	0.053
β_2	0.03	80	0.003	0.225
		800	0.005	0.076
		8,000	0.018	0.039
β_3	0.05	80	0.017	0.22
		800	0.016	0.077
		8,000	0.036	0.041
β_4	-0.13	80	-0.023	0.311
		800	-0.032	0.119
		8,000	-0.086	0.073
β_5	-0.15	80	-0.041	0.316
		800	-0.042	0.121
		8,000	-0.104	0.078
β_6	-0.07	80	-0.013	0.206
		800	-0.015	0.078
		8,000	-0.045	0.053
β_7	0.23	80	0.055	0.431
		800	0.064	0.188
		8,000	0.183	0.141

Table C.5: Average bias and standard deviation of kernel estimates under DGP 4

Coefficients	True Value	n	mean($\hat{\beta}^s_{k,n}$)	SD($\hat{\beta}^s_{k,n}$)
β_0	0.09	80	0.113	0.23
		800	0.117	0.078
		8,000	0.121	0.033
β_1	0.16	80	0.108	0.329
		800	0.113	0.118
		8,000	0.152	0.052
β_2	0	80	-0.002	0.223
		800	-0.001	0.073
		8,000	-0.001	0.033
β_3	0.05	80	0.027	0.221
		800	0.026	0.079
		8,000	0.042	0.039
β_4	-0.1	80	-0.031	0.309
		800	-0.037	0.112
		8,000	-0.067	0.061
β_5	-0.15	80	-0.064	0.319
		800	-0.065	0.125
		8,000	-0.116	0.072
β_6	-0.01	80	-0.003	0.203
		800	-0.002	0.067
		8,000	-0.005	0.038
β_7	0.09	80	0.041	0.416
		800	0.044	0.15
		8,000	0.09	0.09

Table C.6: Average bias and standard deviation of kernel estimates under DGP 5

Coefficients	True Value	n	mean($\hat{\beta}^s_{k,n}$)	SD($\hat{\beta}^s_{k,n}$)
β_0	0.09	80	0.1	0.23
		800	0.105	0.077
		8,000	0.108	0.027
β_1	0.16	80	0.144	0.327
		800	0.149	0.113
		8,000	0.168	0.041
β_2	0	80	-0.002	0.224
		800	-0.001	0.076
		8,000	-0.003	0.033
β_3	0.01	80	0.007	0.217
		800	0.004	0.069
		8,000	0.004	0.022
β_4	-0.1	80	-0.042	0.311
		800	-0.052	0.118
		8,000	-0.085	0.054
β_5	-0.02	80	-0.011	0.307
		800	-0.006	0.096
		8,000	-0.007	0.032
β_6	-0.01	80	-0.003	0.2
		800	-0.002	0.065
		8,000	-0.003	0.026
β_7	0.02	80	0.009	0.408
		800	0.008	0.132
		8,000	0.01	0.052

- DGP 5: the bandwidths associated with g and h should be zero and one, respectively.

Figures C.1-C.5 display the bandwidths across the 10,000 simulations for all DGPs, highlighting the ones from data sets with 80 observations in red, with 800 in green, and with 8,000 in blue. We verify that all expected patterns hold.

Furthermore, we observe that it is not straightforward for the estimator to pick up on the interactions. Notably, when the sample size is just 80 it is very difficult for the estimator to detect the interactions as the majority of bandwidths equal one even when these interactions are significant. This is an important because with only 80 observations there are only 10 per full treatment assignment, a number that is comparable to what our experiments features. This reinforces our confidence that the interactions we were able to detect are indeed significant.

Notice that these results also illustrate the necessity to observe all treatment combinations in the data. To see this, consider DGPs 3 and 4. Under DGP 3, g and h unilaterally reduce f's treatment effect from 0.16 to 0.03 and 0.01, respectively. Together, however, these effects seem to cancel each other out and f's treatment effect falls to 0.11 only. In turn, under DGP 4, g and h unilaterally reduce f's treatment effect to 0.06 and 0.01, respectively, and together they reduce it to zero. Hence, while h always interferes with f's advertising, its interference is somewhat mitigated under DGP 3 relative to DGP 4, so we would expect λ_h to have lower values under DGP 4. The results confirm this intuition.

Now suppose that a partial factorial design was used instead and that this design did not consider the case where $D_g = D_h = 1$. Under both DGPs 3 and 4, h's unilateral effect on f's treatment effect is to reduce it from 0.16 to 0.01. Consequently, using the kernel-based estimator on such data would yield the same value for λ_h. This would potentially lead to a misleading extrapolation of the effect on f's treatment effect when both g and h were present, which highlights that to obtain estimates that account for the different treatment interactions it is necessary to observe all treatment combinations in the data.

Figure C.1: Bandwidths for DGP 1

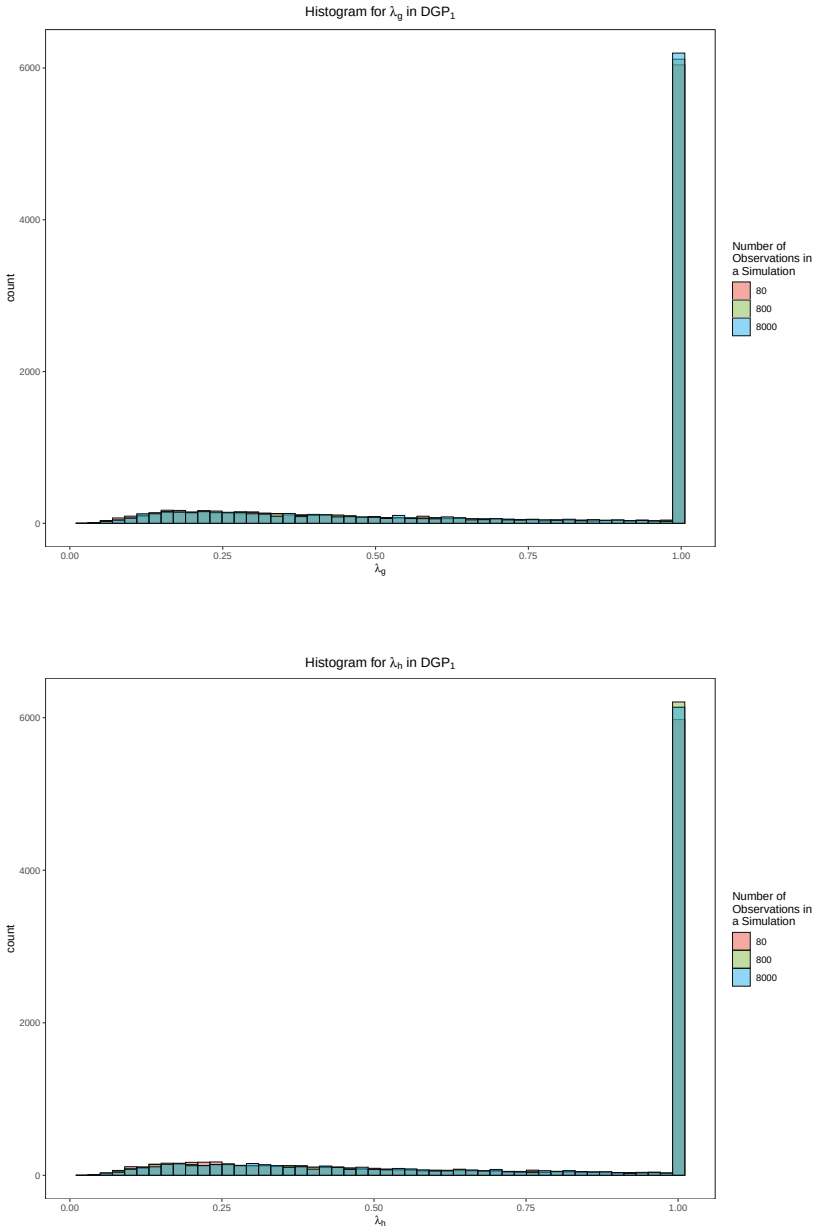


Figure C.2: Bandwidths for DGP 2

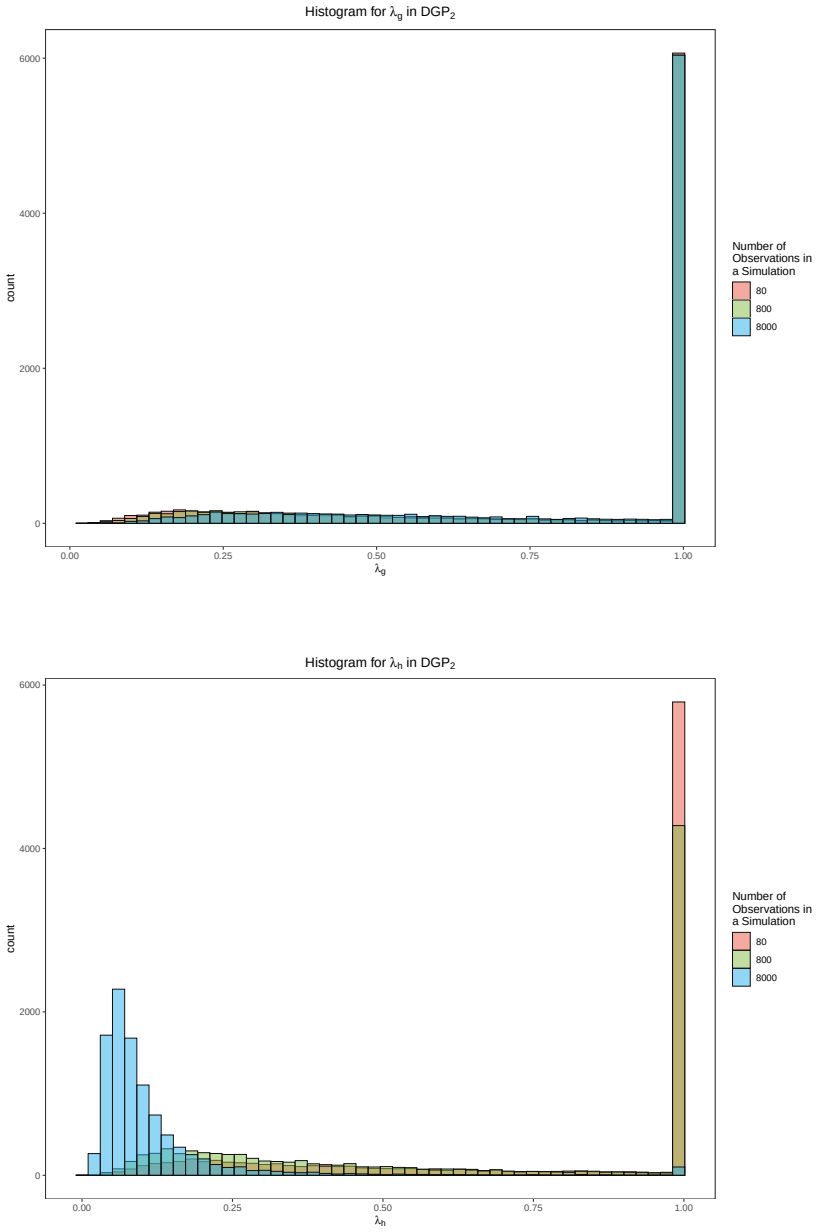


Figure C.3: Bandwidths for DGP 3

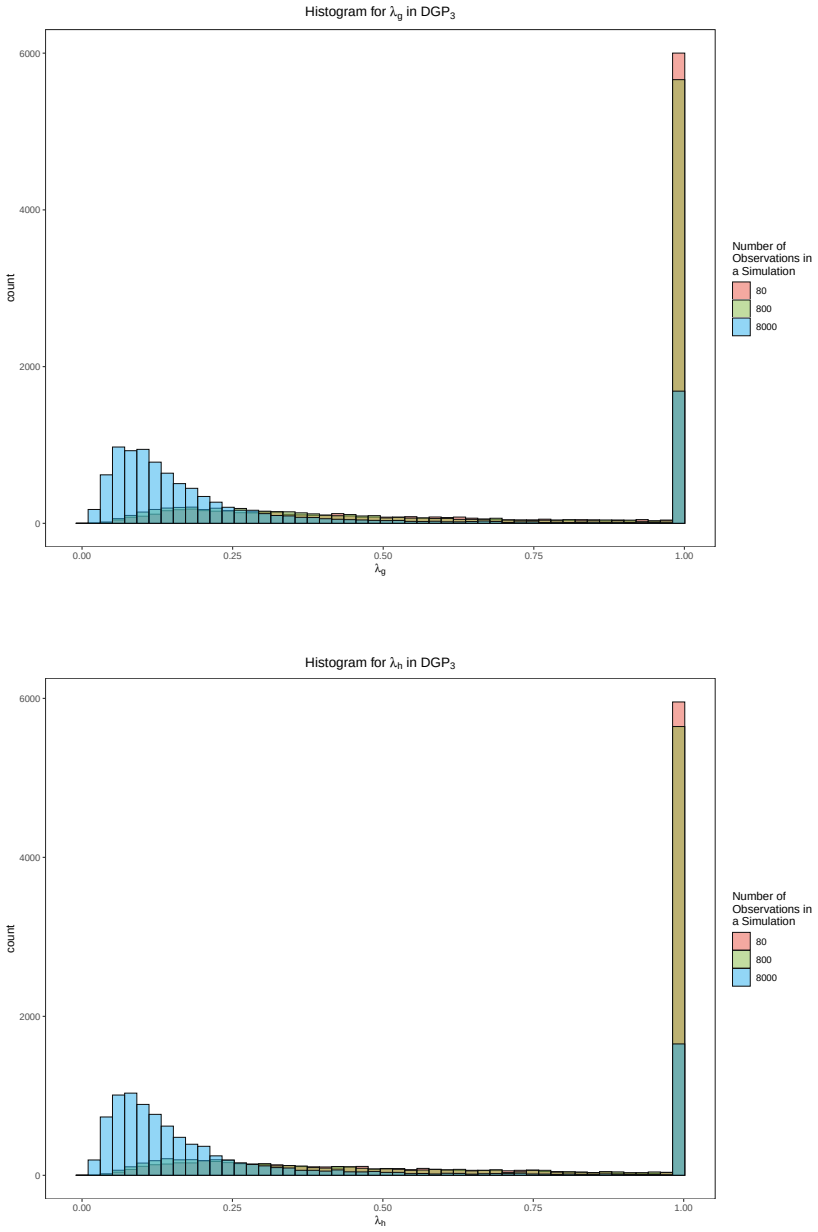


Figure C.4: Bandwidths for DGP 4

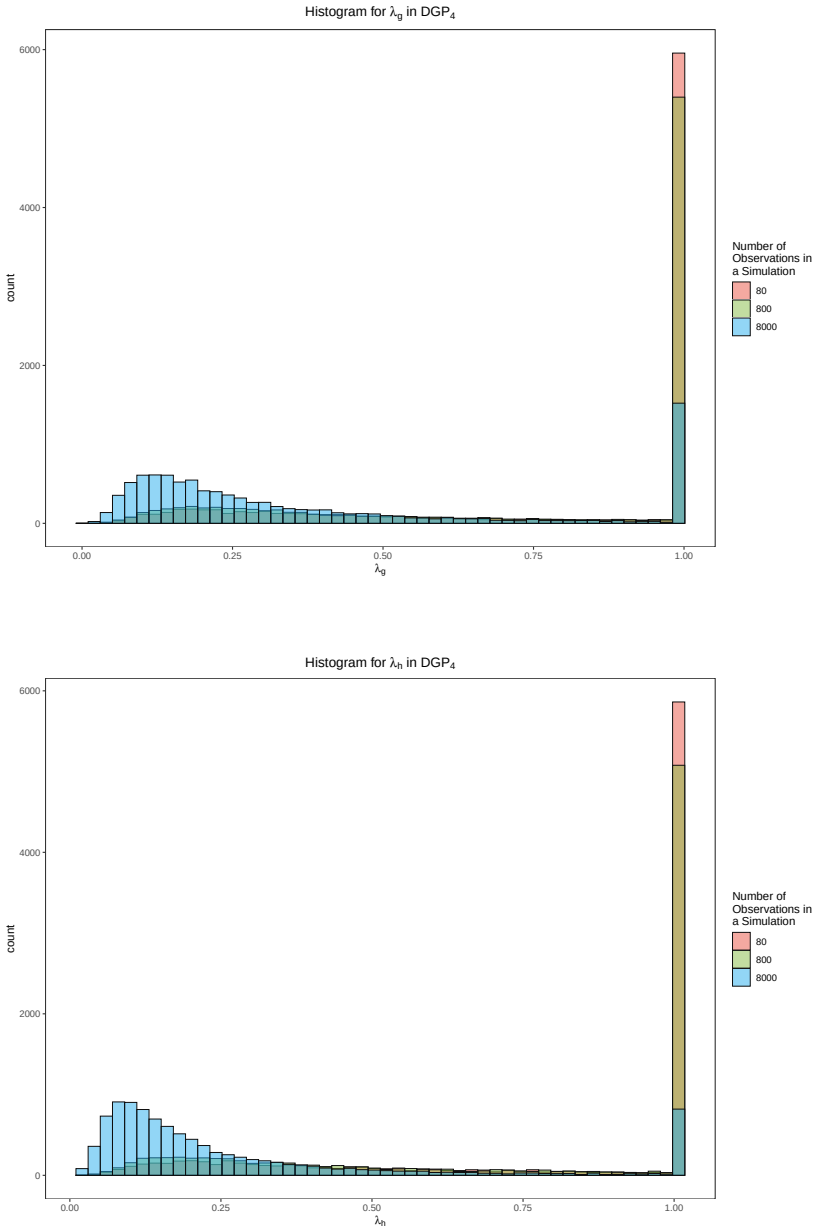
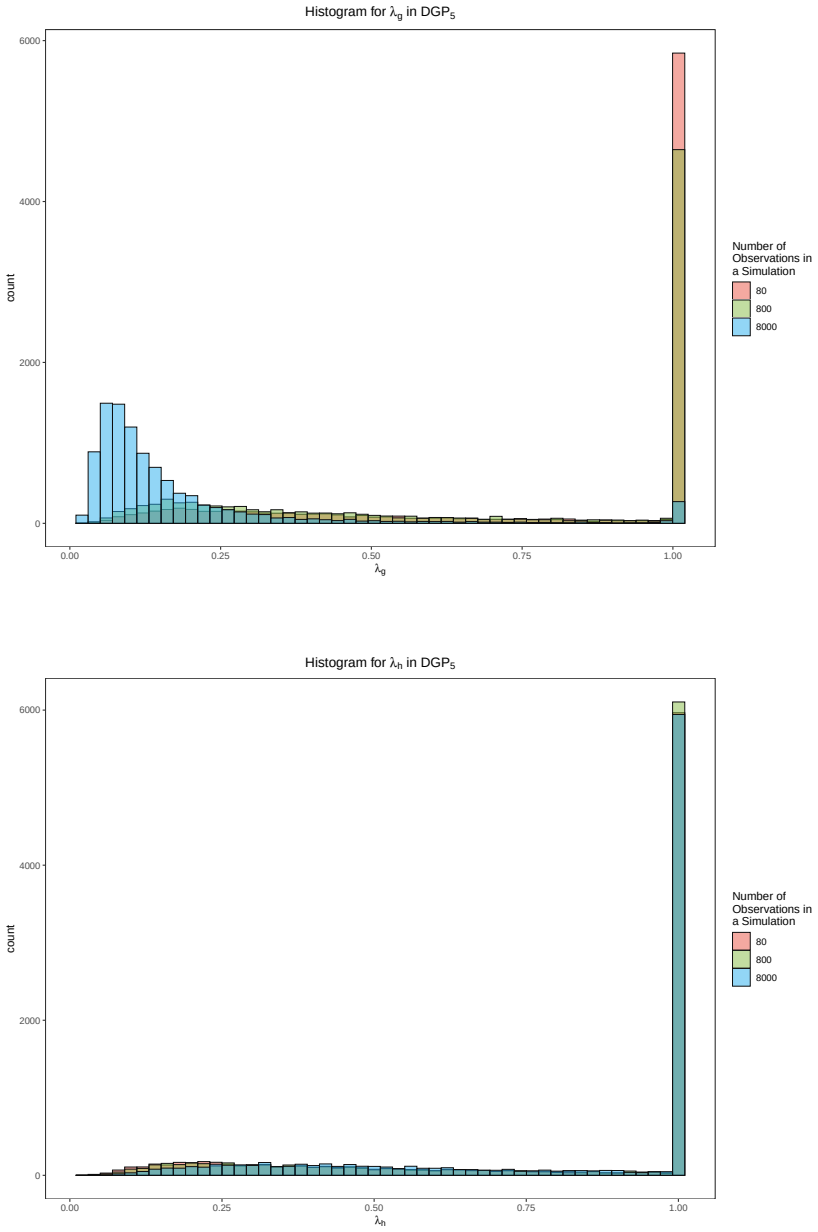


Figure C.5: Bandwidths for DGP 5



C.6 Comparison between methods

Criteria of comparison We compare the performance of the three estimators using three criteria: bias, mean squared error (MSE), and median absolute error (MAE).

We compute MSE as follows. Let $\hat{\beta}_{k,n}^{s,m}$ be the estimate of the $k-$th coefficient from sample

s of size n using method m. We compute the MSE as:

$$\text{MSE}_{k,n}^{m} = \frac{1}{10,000} \sum_{s=1}^{10,000} \left(\hat{\beta}_{k,n}^{s,m} - \beta_k \right)^2. \tag{C.3}$$

To assess the bias-variance tradeoff, which is especially important because of the regularization used by the kernel-based estimator, we assess how much of the MSE is comprised by bias by considering the ratio of bias-squared to MSE where bias squared is:

$$\left(\text{bias}_{k,n}^{m}\right)^2 = \left[\frac{1}{10,000} \sum_{s=1}^{10,000} \left(\hat{\beta}_{k,n}^{s,m} - \beta_k \right) \right]^2. \tag{C.4}$$

Finally, we also consider the median absolute error (MAE). The absolute error of the estimate of the $k-$th coefficient from sample s of size n using method m is $\left| \hat{\beta}_{k,n}^{s,m} - \beta_k \right|$. Using order statistics notation, let the s-th highest absolute error be $\left| \hat{\beta}_{k,n}^{(s),m} - \beta_k \right|$. Then:

$$\text{MAE}_{k,n}^{m} = \frac{\left| \hat{\beta}_{k,n}^{(5,000),m} - \beta_k \right| + \left| \hat{\beta}_{k,n}^{(5,001),m} - \beta_k \right|}{2}. \tag{C.5}$$

Results We begin by comparing the MSE of the methods and the fraction of them that consist of bias squared. Figures C.6-C.15 display these quantities for the all coefficients, DGPs, and sample sizes. Results from the kernel-based estimator are shown in red, from NLS in green, and from OLS in blue.

Figure C.6: MSEs under DGP 1: $\beta_0 - \beta_3$

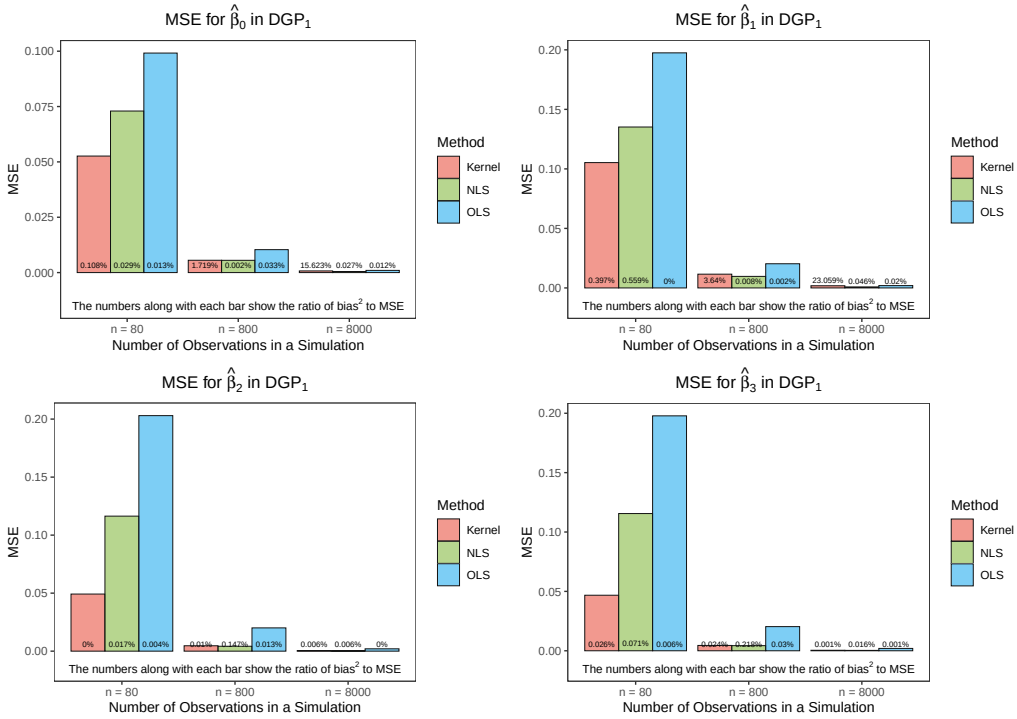


Figure C.7: MSEs under DGP 1: $\beta_4 - \beta_7$

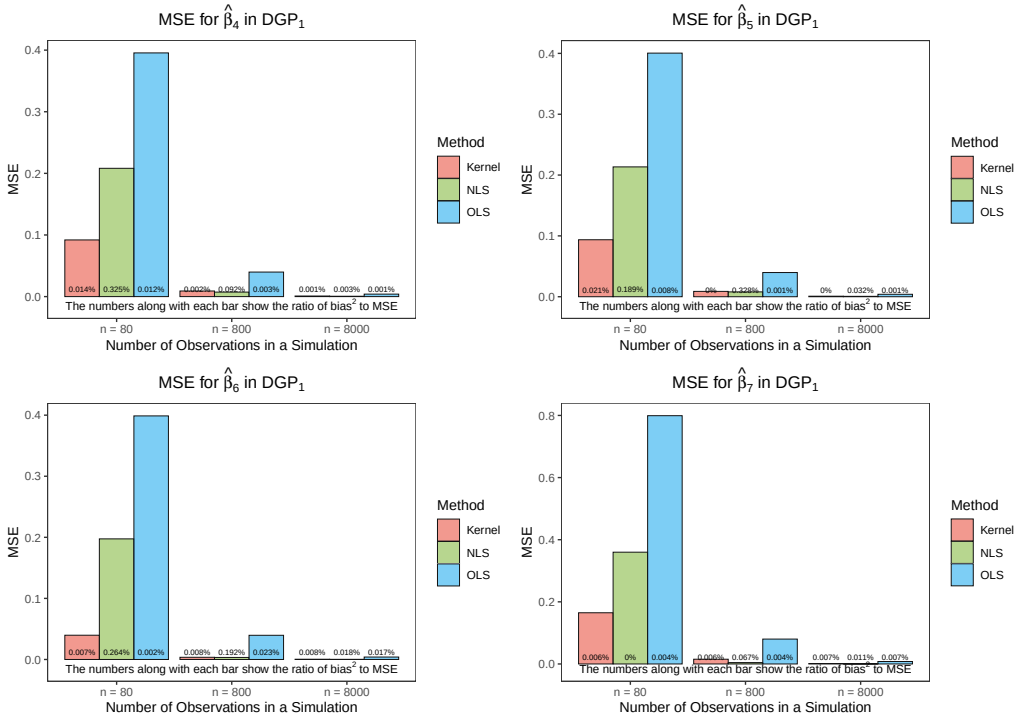


Figure C.8: MSEs under DGP 2: $\beta_0 - \beta_3$

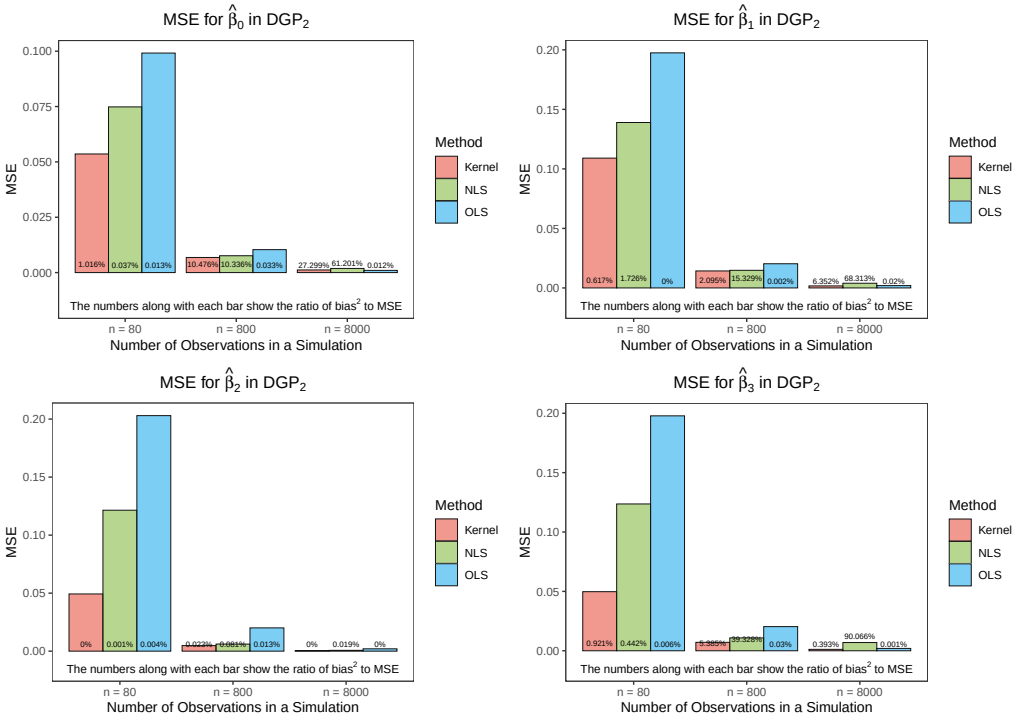


Figure C.9: MSEs under DGP 2: $\beta_4 - \beta_7$

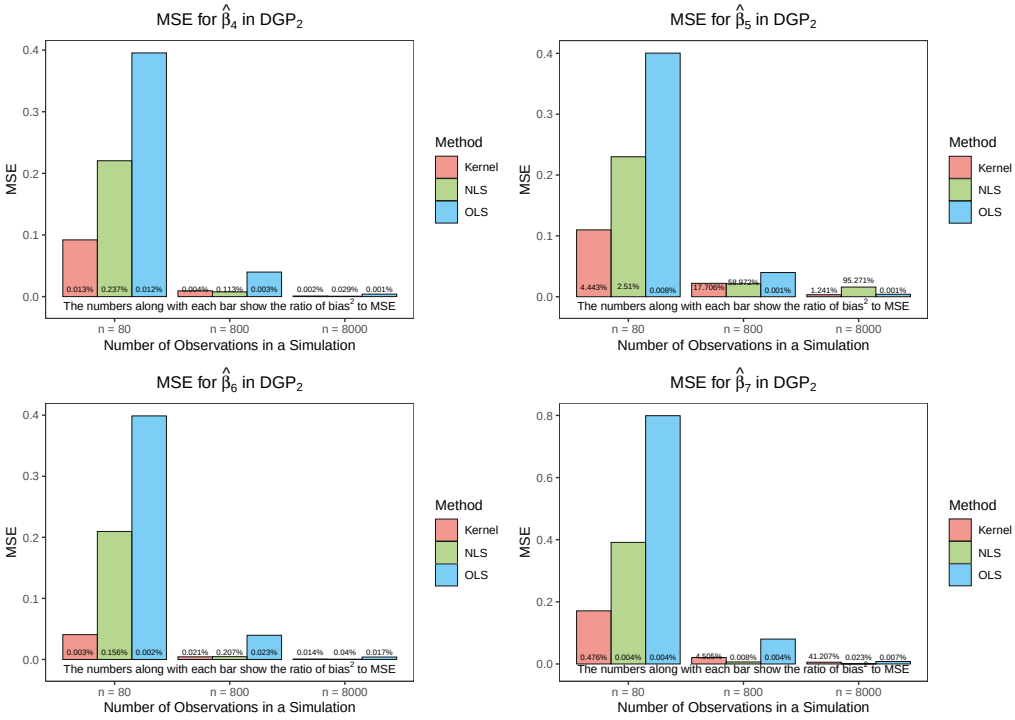


Figure C.10: MSEs under DGP 3: $\beta_0 - \beta_3$

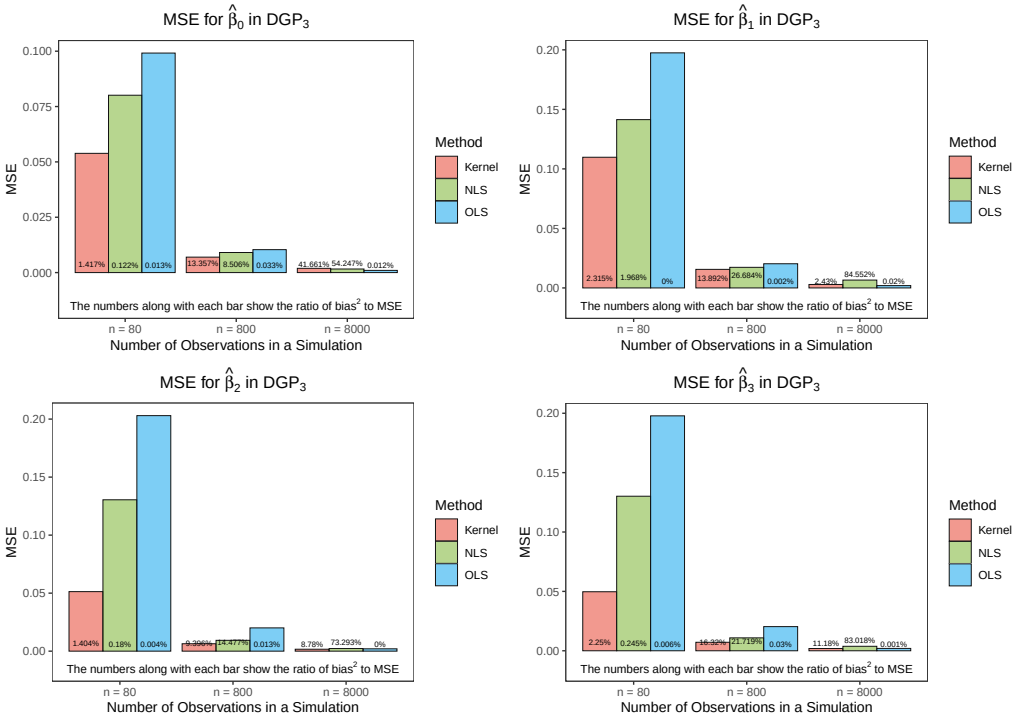


Figure C.11: MSEs under DGP 3: $\beta_4 - \beta_7$

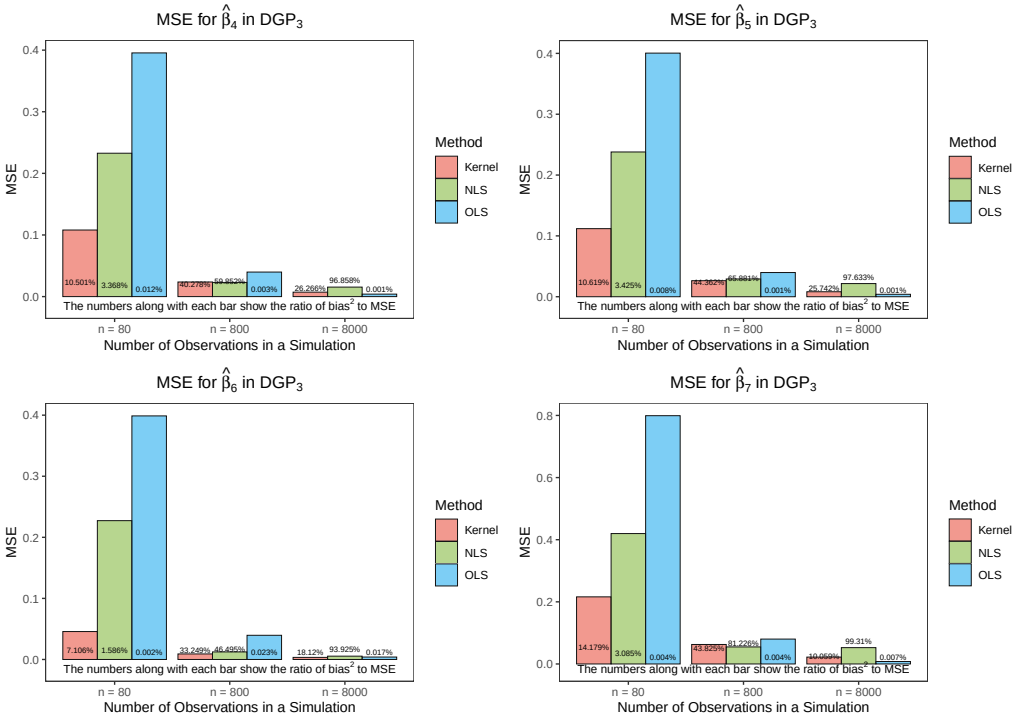


Figure C.12: MSEs under DGP 4: $\beta_0 - \beta_3$

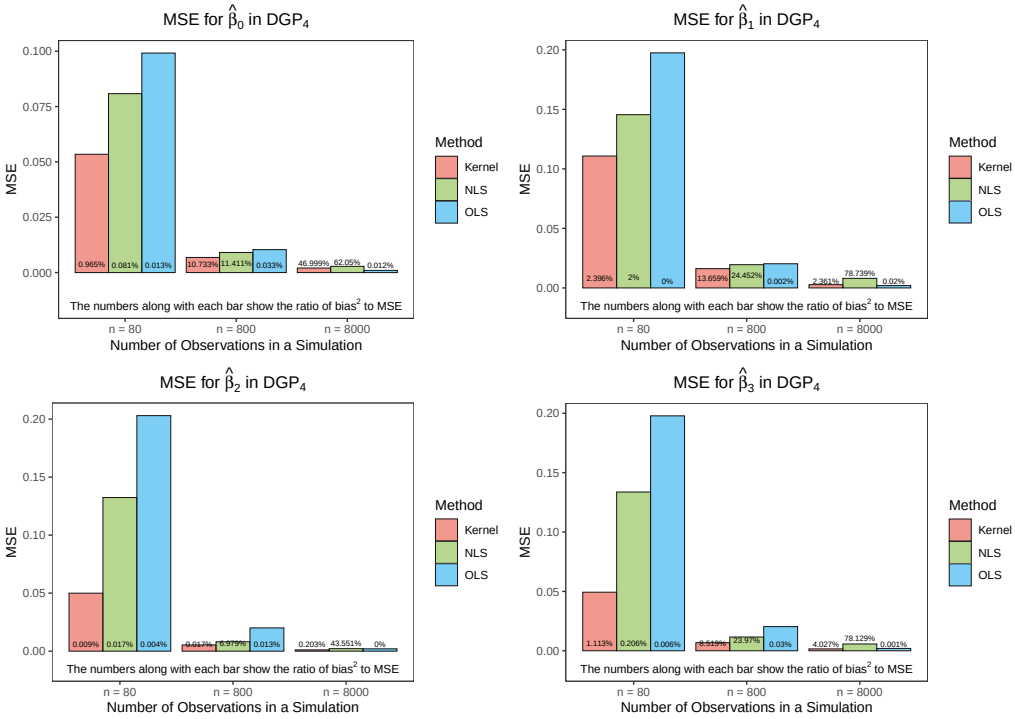


Figure C.13: MSEs under DGP 4: $\beta_4 - \beta_7$

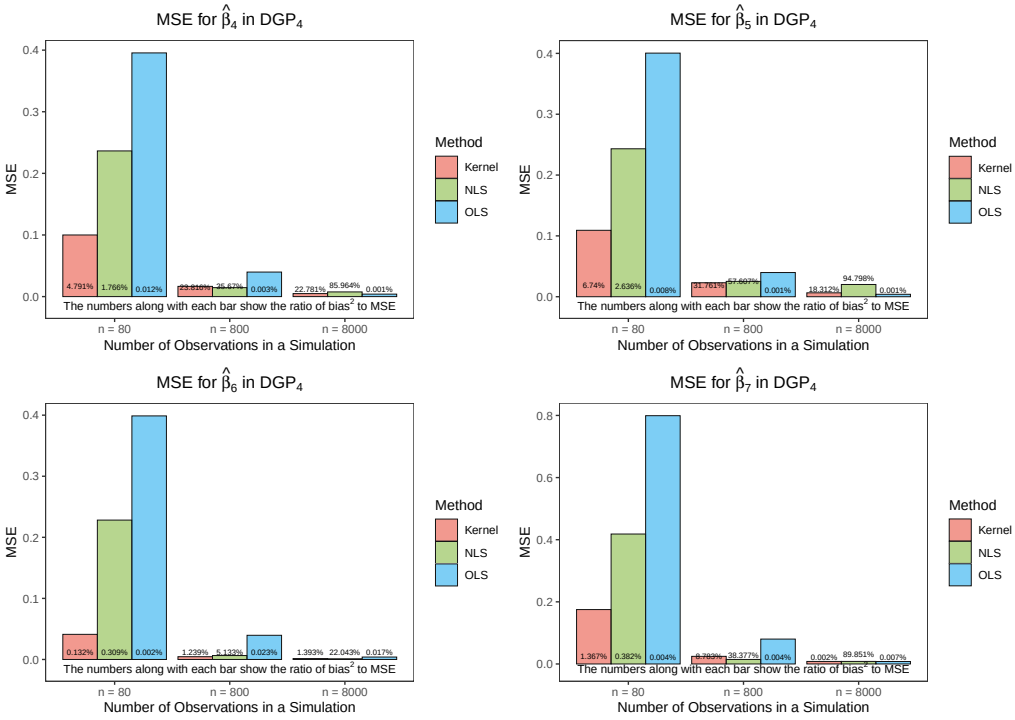


Figure C.14: MSEs under DGP 5: $\beta_0 - \beta_3$

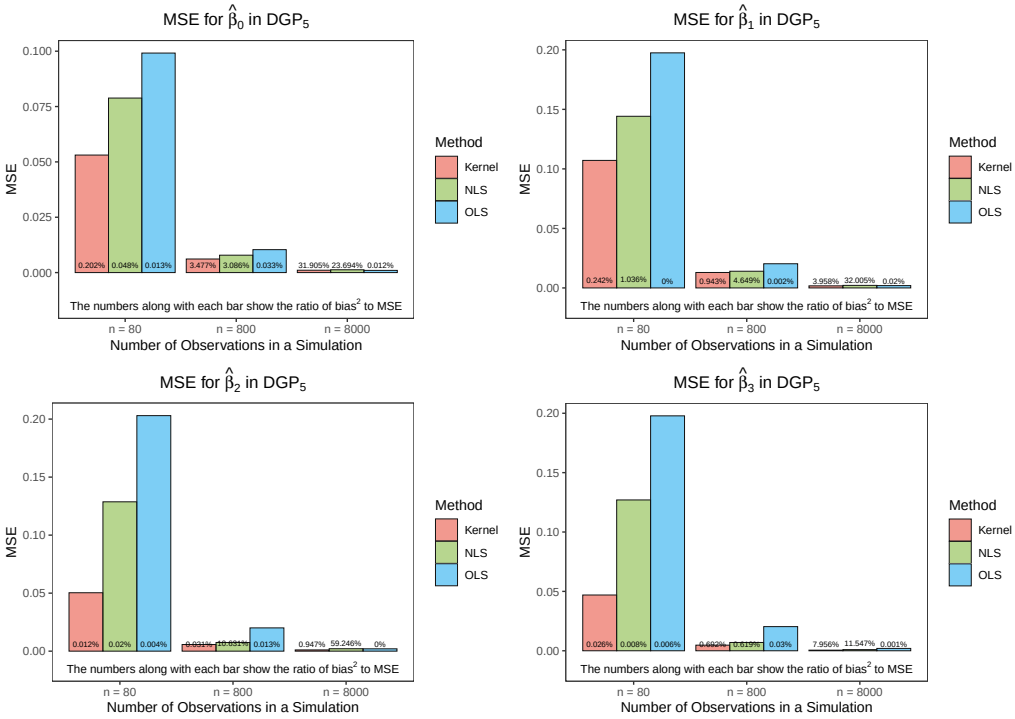
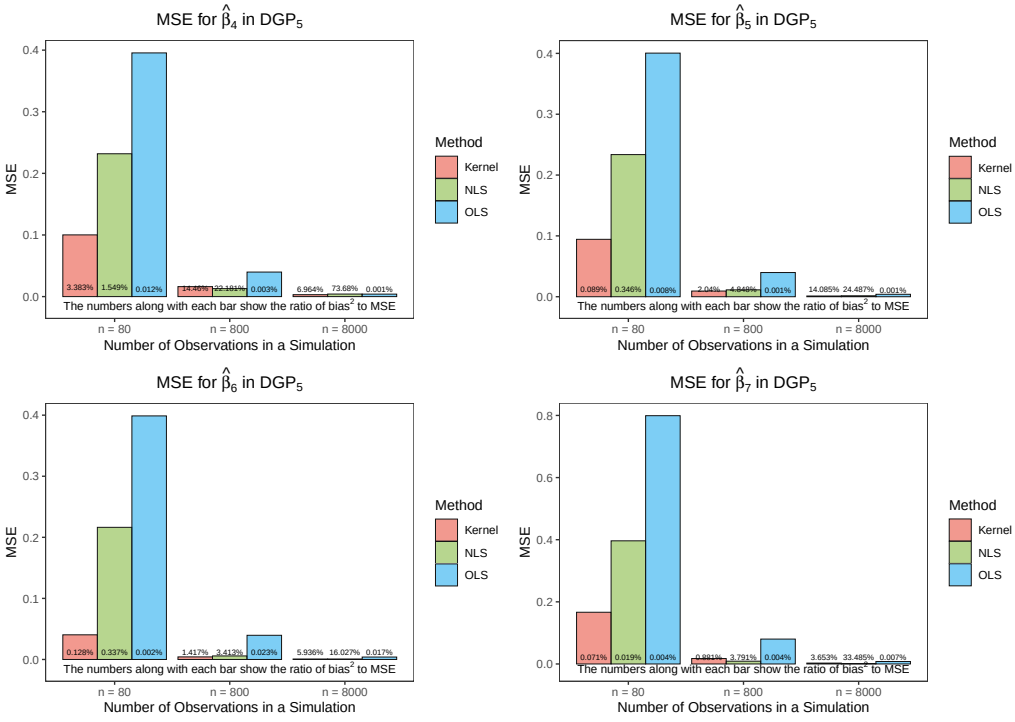


Figure C.15: MSEs under DGP 5: $\beta_4 - \beta_7$



The results are as expected. In smaller samples, the kernel-based estimator performs better in minimizing MSE due to the regularization. However, OLS performs best in minimizing bias, as it yields the smallest fractions of the MSEs corresponding to bias squared. Expectedly, as the sample size increases the differences between the methods vanishes.

It is also worth noting that often NLS outperforms OLS in minimizing MSE when the sample sizes are smaller and that it performs best amongst the three methods under DGP 1. This is unsurprising as NLS imposes restrictions between parameters that are true for this DGP. For other DGPs, however, not all such restrictions hold, and so as the sample size increases NLS's performance worsens relative to that of the other two methods.

For completeness, Figures C.16-C.25 display the MAEs for the all coefficients, DGPs, and sample sizes. Once again, results from the kernel-based estimator are shown in red, from NLS in green, and from OLS in blue.

Figure C.16: MAEs under DGP 1: $\beta_0 - \beta_3$

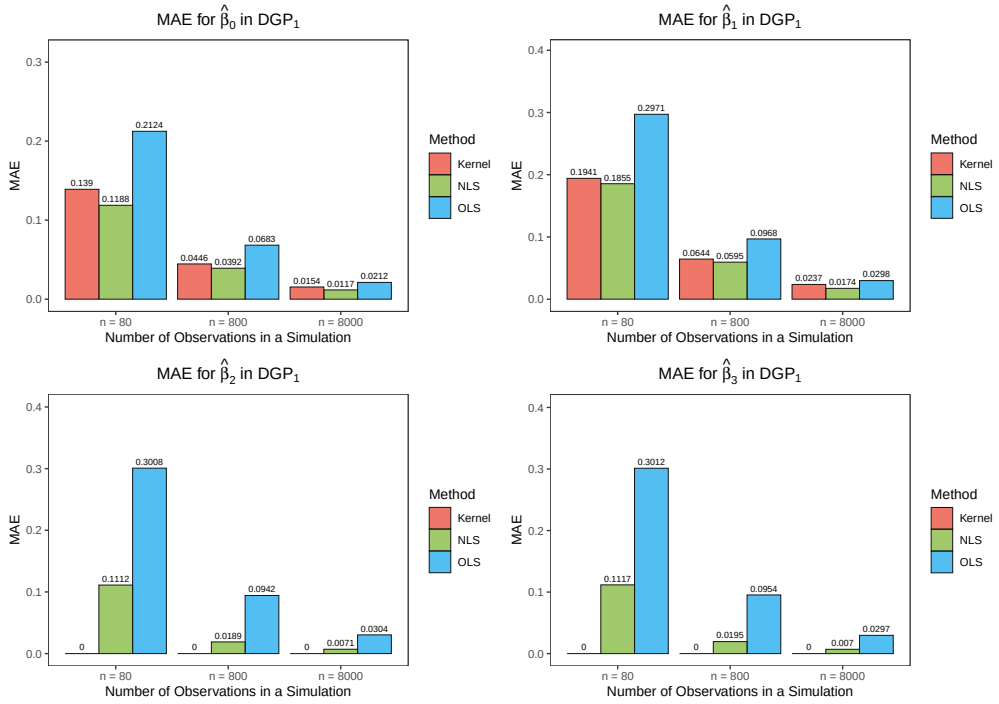


Figure C.17: MAEs under DGP 1: $\beta_4 - \beta_7$

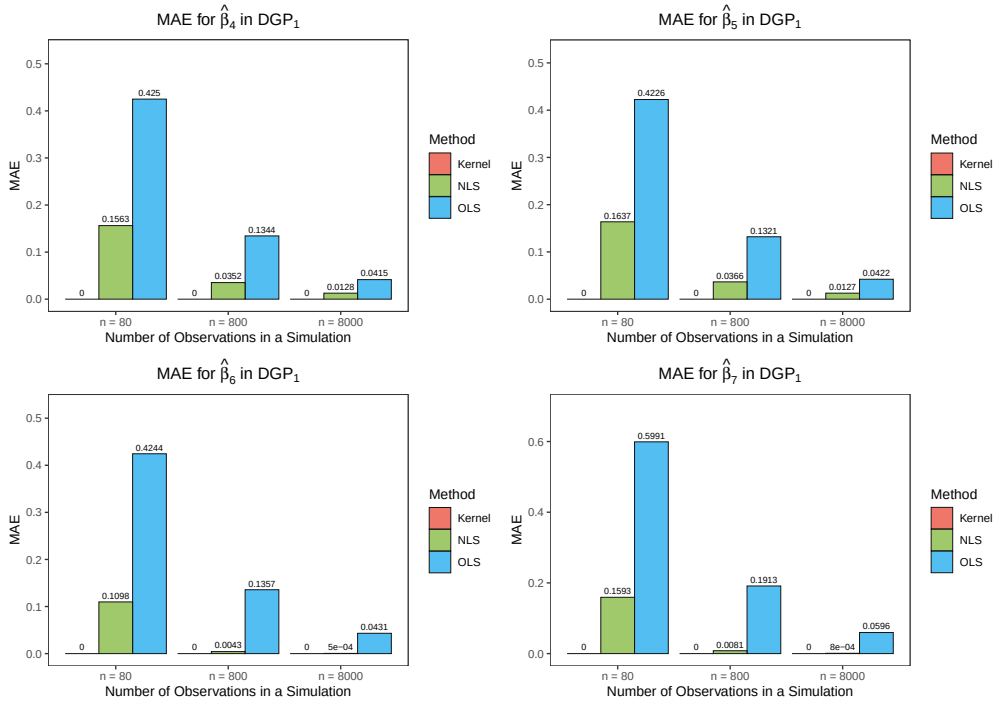


Figure C.18: MAEs under DGP 2: $\beta_0 - \beta_3$

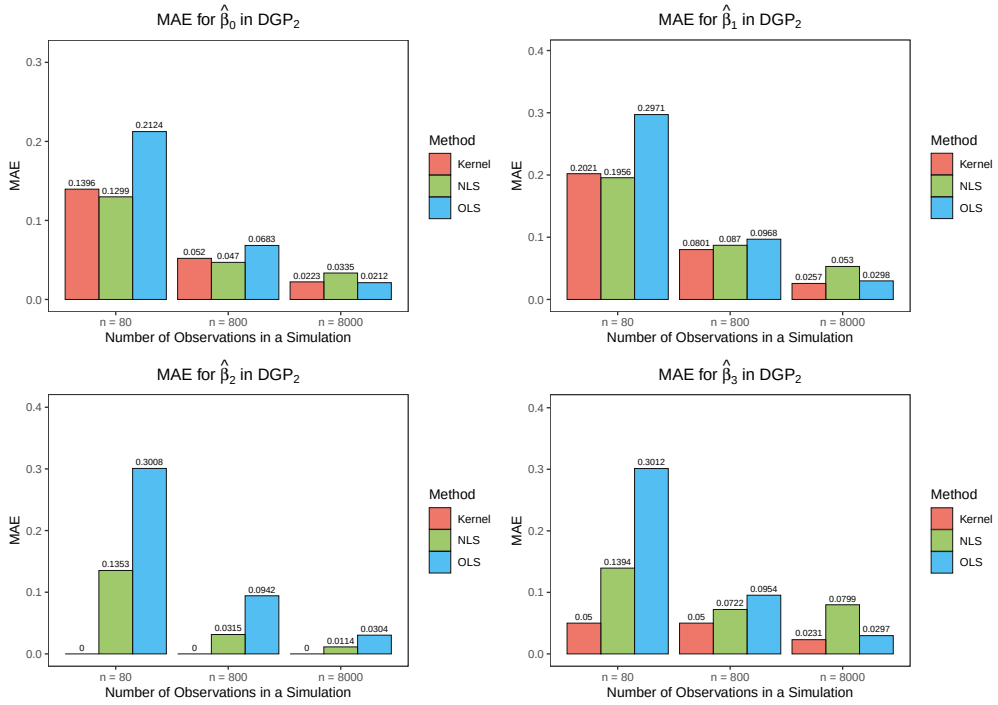


Figure C.19: MAEs under DGP 2: $\beta_4 - \beta_7$

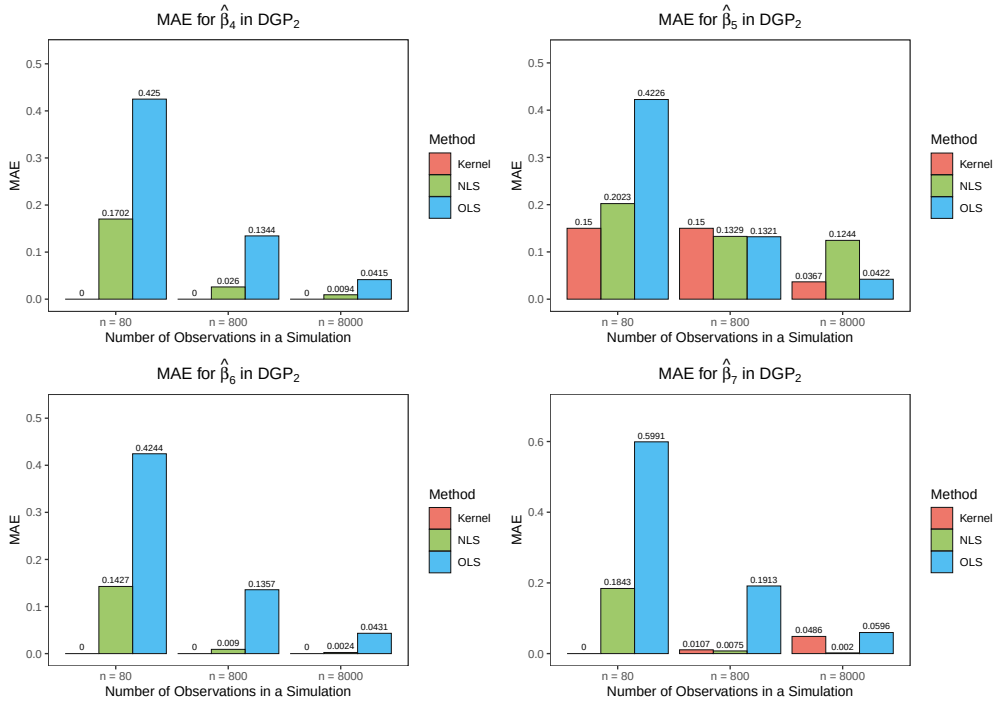


Figure C.20: MAEs under DGP 3: $\beta_0 - \beta_3$

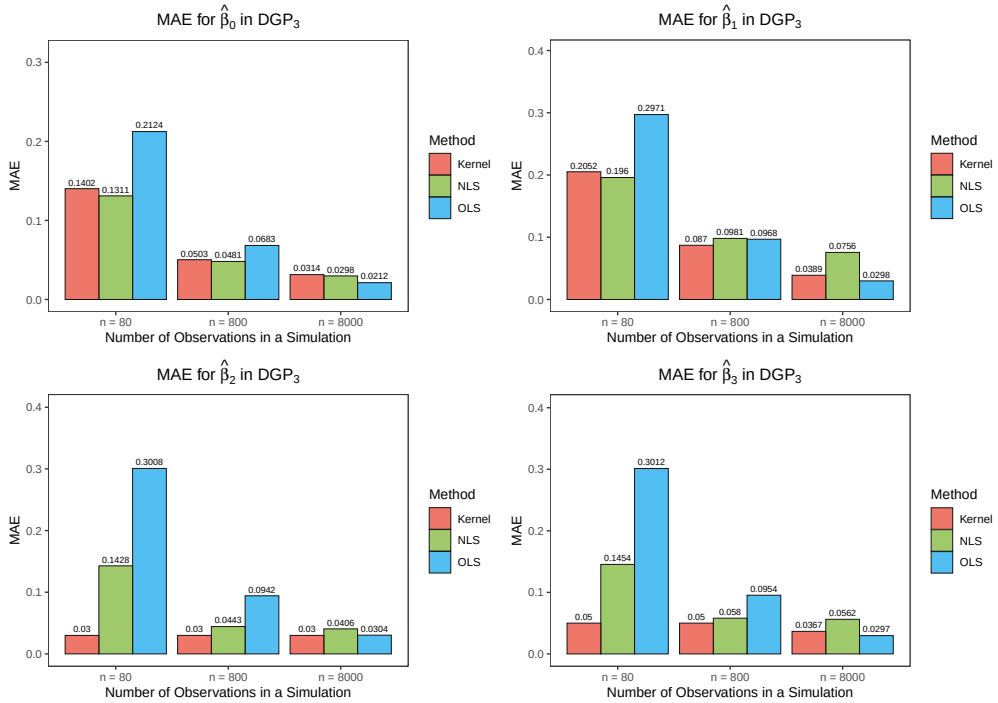


Figure C.21: MAEs under DGP 3: $\beta_4 - \beta_7$

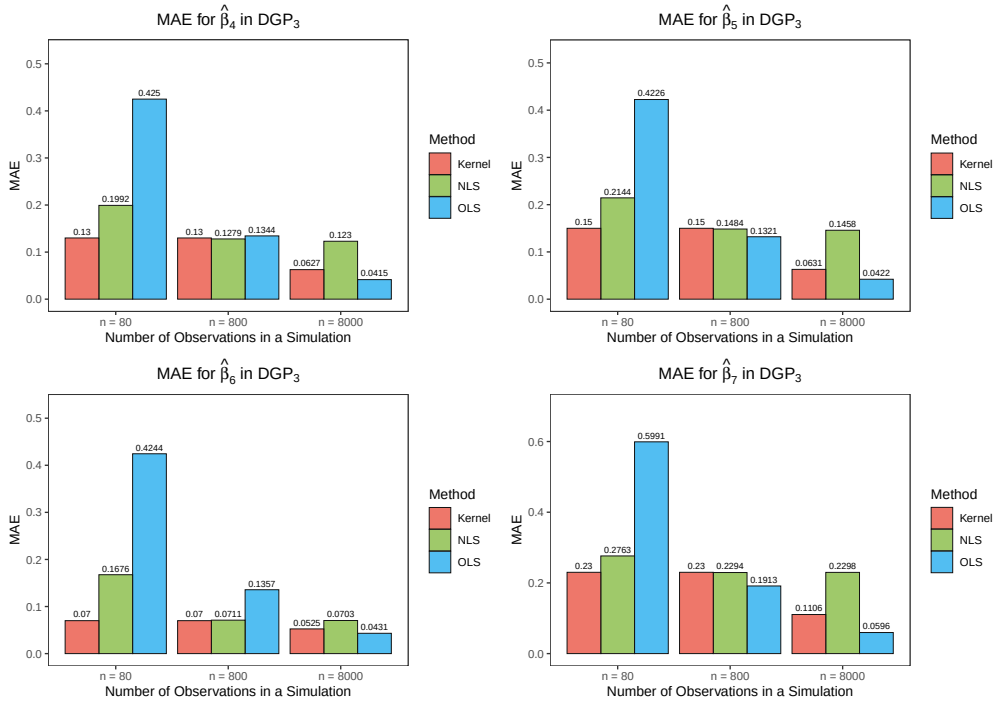


Figure C.22: MAEs under DGP 4: $\beta_0 - \beta_3$

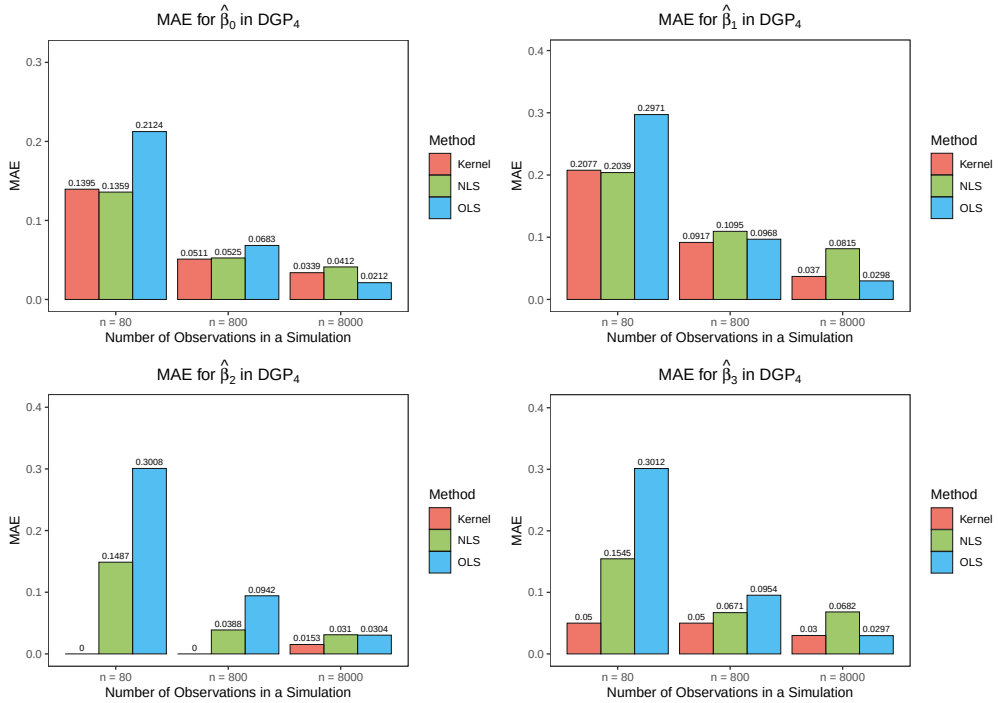


Figure C.23: MAEs under DGP 4: $\beta_4 - \beta_7$

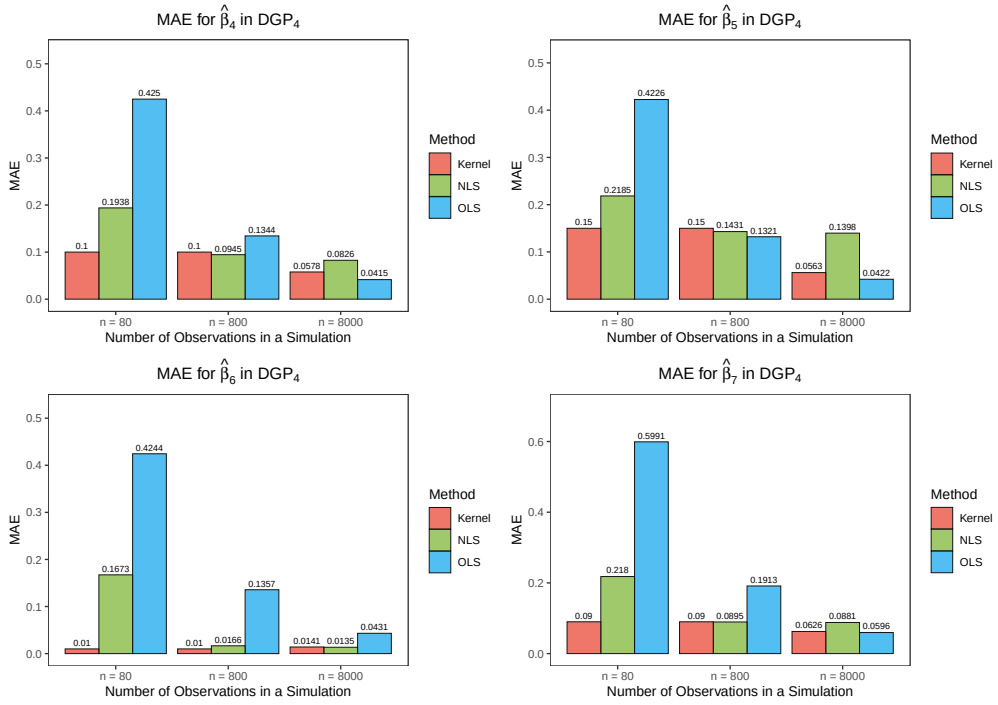


Figure C.24: MAEs under DGP 5: $\beta_0 - \beta_3$

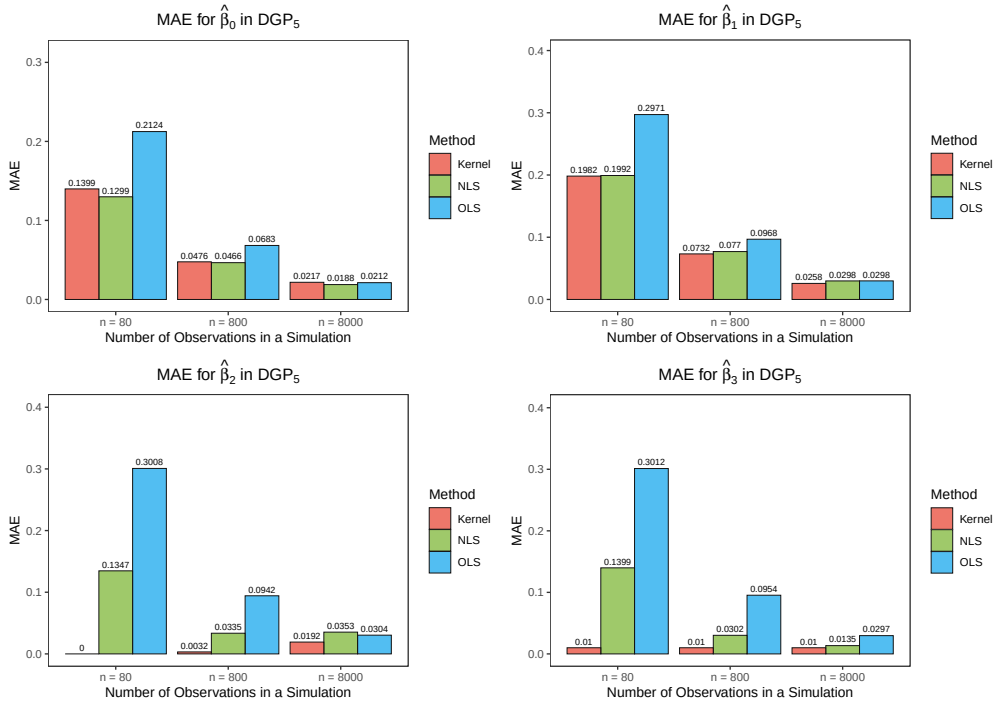
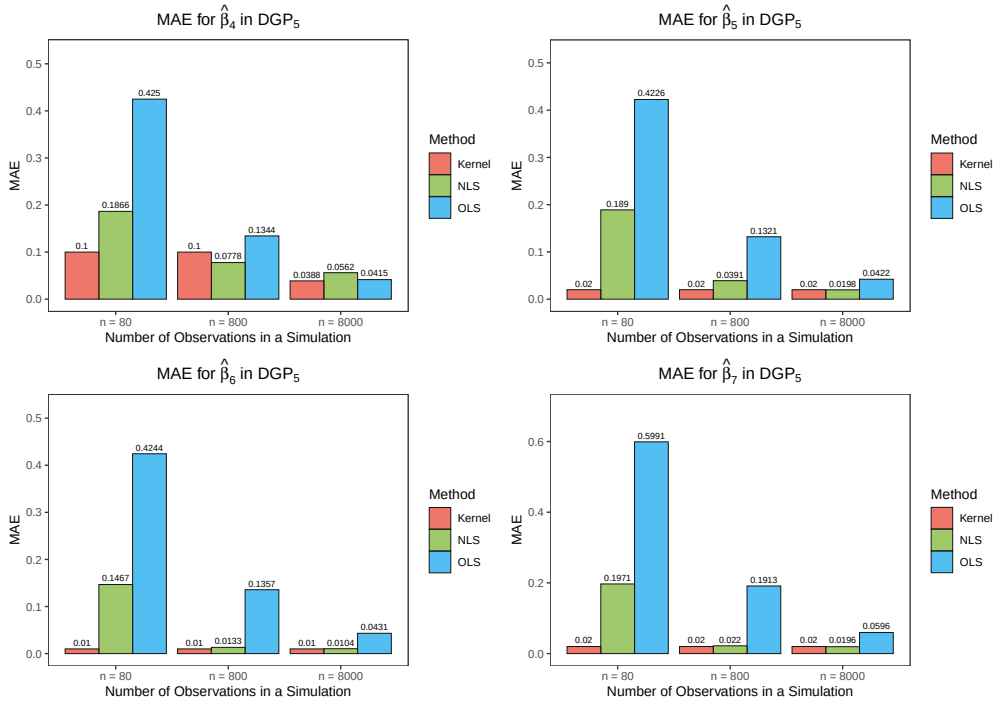


Figure C.25: MAEs under DGP 5: $\beta_4 - \beta_7$



As expected, results improve as the sample size increases. We can see that the kernel-based estimator performs very well as it often obtains an MAE of zero. This is not entirely surprising given the results shown in Figures C.1-C.5. This estimator has a propensity to set the bandwidths to one, implying that it does not detect interference between advertisers. When if fact there is no such interference, the coefficients associated with the indicators for the "non-intefering" advertisers are zero, which the kernel-based estimator will impose, finally yielding no error at all. When the coefficients are non-zero, there is no chance any estimator can often obtain no error. In such cases, the performance of the different methods varies.

References

Athey, S., Imbens, G. W., and Wager, S. (2018). Approximate residual balancing: Debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B*, 80(4):597–623.

Bradic, J., Wager, S., and Zhu, Y. (2019). Sparsity double robust inference of average treatment effects. *arXiv:1905.00744*.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *Annals of Applied Statistics*, 7(1):443–470.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects: A review. *Review of Economics and Statistics*, 86(1):4–29.

Li, Q., Ouyang, D., and Racine, J. S. (2013). Categorical semiparametric varying-coefficient models. *Journal of Applied Econometrics*, 28(4):551–579.

Li, Q., Racine, J. S., and Wooldridge, J. M. (2009). Efficient estimation of average treatment effects with mixed categorical and continuous data. *Journal of Business & Economics Statistics*, 27(2):206–223.

Ning, Y., Peng, S., and Imai, K. (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika*, 107(3):533–554.

Racine, J. S. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119(1):99–130.

Tan, Z. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Annals of Statistics*, 48(2):811–837.

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532.

Ye, Z., Zhang, Z., Zhang, D., Zhang, H., and Zhang, R. (2024). Deep learning based causal inference for large-scale combinatorial experiments: Theory and empirical evidence. *SSRN Working Paper No. 4375327*.