Statistica Sinica: Supplement

Effective use of multiple error-prone covariate measurements in capture-recapture models

Kun Xu and Yanyuan Ma

Texas A&M University

Supplementary Material

In addition to simulations 1, 2 and 3 in the main paper, we now illustrate more simulation results to cover different settings and scenarios. Simulation 4 and 5 are for low capture probabilities with $N = 300$ and 700 respectively. We use them to further investigate the finite sample performance of the five methods, namely conditional score (CS), two types of GMM (GMM1, GMM2), the semiparametric estimator using normal $f_X(x)$ (Semi-Nor) and using uniform $f_X(x)$ (Semi-Uni), in comparison with simulation 2 where $N = 500$. Simulations 6, 7 and 8 contain two covariates and low capture probabilities at sample size $N = 300$, 500 and 700. The last two simulations, 9 and 10, are for a relatively high capture probabilities when we have two covariates. All the simulation experiments are based on 1000 data sets. The details are discussed in the following sections.

S1 Simulation 4

In simulation 4, we set population size to be $N = 300$. We generate the true covariate X_i from a standard normal distribution and set the measurement error standard deviation $\sigma_u = 0.6$. We generate the observations $(Y_{ij}, W_{ij}Y_{ij})$, $j = 1, 2, 3$ from the model with true parameter values $\alpha = -1.0$ and $\beta = 1.0$. It yields an average of 179 first time captures and 75 second time captures. The estimated $\widehat{\Sigma}$ has bias -0.0054 and variance 0.0034.

We summarize the results in Table 1 where we report the mean, the sample standard error of 1000 estimates, the average of 1000 estimated standard errors, 95% coverage rate and mean squared error. We see that there are small biases in the estimation of both model parameters α, β and population size, except for the two GMM methods (GMM1, GMM2) where they show a certain amount of bias. The sample standard error of 1000 estimates and the average of 1000 estimated standard errors are close to each other for methods CS, Semi-Nor and Semi-Uni, indicating a satisfactory performance of asymptotic results for even $N = 300$. All coverage rates are close to the nominal level. Mean squared error drops 17%, 23% and 34.5% for α, β and N respectively from the CS to Semi-Nor method. Based on these observations, we obtain similar conclusions as simulation 2. GMM cannot improve upon CS in estimation efficiency because of low capture probability and a small sample size $N = 300$. The two semiparametric methods have superior performance in reducing the estimation variability.

		α	β	N
	true	-1.0	1.0	300
CS	estimate	-1.0269	1.0331	319.65
	emp se	0.2128	0.2458	82.83
	mse	0.0919	0.1165	18415
	est se	0.2108	0.2280	64.47
	95% cov	96.3%	94.7%	92.9%
GMM1	estimate	-1.1246	1.1264	366.63
	emp se	0.2654	0.3499	201.48
	mse	0.1634	0.2606	220240
	est se	0.2398	0.2735	131.49
	95% cov	94.2%	94.8%	96.3%
GMM2	estimate	-1.1189	1.1208	360.91
	emp se	0.2696	0.3485	164.79
	mse	0.1450	0.2256	129535
	est se	0.2304	0.2613	113.11
	95% cov	94.6%	95.3%	96.2%
Semi-Nor	estimate	-1.0241	1.0217	315.12
	emp se	0.1963	0.2149	72.57
	mse	0.0763	0.0897	12070
	est se	0.1912	0.2043	55.31
	95% cov	95.1%	94.9%	92.4%
Semi-Uni	estimate	-1.0244	1.0223	315.21
	emp se	0.1968	0.2158	72.66
	mse	0.0767	0.0903	12044
	est se	0.1916	0.2050	55.35
	95% cov	95.4%	94.9%	92.3%

Table 1: Simulation 4. Performance of five methods based on conditional score (CS), two types of GMM (GMM1, GMM2), in comparison with the semiparametric estimator using normal $f_X(x)$ (Semi-Nor) and using uniform $f_X(x)$ (Semi-Uni). The mean of the estimates (estimate), empirical standard error (emp se), mean squared error (mse), average of estimated standard error (est se) and the sample coverage rate of the 95% confidence interval (95% cov) are reported.

S2 Simulation 5

We set $N = 700$ and use exactly the same data generation procedure of simulation 4. The averaged first and second time captures are 417 and 175. The estimated $\widehat{\Sigma}$ has bias -0.0013 and variance 0.0016. We summarize the results in Table 2. We report the mean, the sample standard error of 1000 estimates, the average of 1000 estimated standard errors, 95% coverage rate and mean squared error. Estimates of α, β and N for all five methods have small biases. The sample and average of 1000 standard errors are fairly close to each other for CS, Semi-Nor and Semi-Uni. The coverage rates are around 95% nominal level. The mse drops 11%, 21% and 18% respectively for α, β and N when we compare Semi-Nor to CS. The analysis of the increasement of sample size from $N = 300$ to 500 and then to 700 shows two points. Firstly the semiparametric methods, either Semi-Nor or Semi-Uni, are better than CS all the time. Secondly, GMM method becomes more trustworth when we have a larger sample.

		α	β	N
	true	-1.0	1.0	700
CS	estimate	-1.0146	1.0157	715.71
	emp se	0.1292	0.1486	81.22
	mse	0.0353	0.0438	14393
	est se	0.1349	0.1453	79.70
	95% cov	95.6%	95.3%	94.7%
GMM1	estimate	-1.0423	1.0398	731.12
	emp se	0.1377	0.1579	92.19
	mse	0.0392	0.0476	18316
	est se	0.1349	0.1437	84.41
	95% cov	95.0%	94.1%	95.8%
GMM2	estimate	-1.0415	1.0398	731.32
	emp se	0.1411	0.1593	95.80
	mse	0.0404	0.0487	19337
	est se	0.1362	0.1459	85.45
	95% cov	95.0%	94.1%	95.9%
Semi-Nor	estimate	-1.0070	1.0079	711.05
	emp se	0.1266	0.1312	74.12
	mse	0.0315	0.0347	11795
	est se	0.1236	0.1310	73.07
	95% cov	94.6%	93.9%	92.7%
Semi-Uni	estimate	-1.0072	1.0084	711.23
	emp se	0.1267	0.1316	74.18
	mse	0.0316	0.0349	11831
	est se	0.1238	0.1315	73.24
	95% cov	94.4%	94.5%	93.1%

Table 2: Simulation 5. Performance of five methods based on conditional score (CS), two types of GMM (GMM1, GMM2), in comparison with the semiparametric estimator using normal $f_X(x)$ (Semi-Nor) and using uniform $f_X(x)$ (Semi-Uni). The mean of the estimates (estimate), empirical standard error (emp se), mean squared error (mse), average of estimated standard error (est se) and the sample coverage rate of the 95% confidence interval (95% cov) are reported.

S3 Simulation 6

In simulation 6, we set population size to be $N = 300$. We consider a bivariate covariate $\mathbf{X}_i = (X_{i1}, X_{i2})^{\mathrm{T}}$, where X_{i1} and X_{i2} are generated from a standard normal and Bernoulli distribution respectively. We then set the measurement error standard deviation $\sigma_u = 0.6$ for X_{i1}. We generate the observations $(Y_{ij}, \mathbf{W}_{ij}Y_{ij})$, $j = 1, 2, 3$ from the model with true parameter values $\alpha = -1.0$ and $\boldsymbol{\beta} = (1.0, 0.3)^{\mathrm{T}}$. It yields an average of 233 first time captures and 151 second time captures. The estimated $\widehat{\Sigma}$ has bias -0.0032 and variance 0.0010.

We summarize the results in Table 3 where we report the mean, the sample standard error of 1000 estimates, the average of 1000 estimated standard errors, 95% coverage rate and mean squared error. The means of α, β_1, β_2 and N have small biases. The sample standard error of 1000 estimates and the average of 1000 estimated standard errors are close to each other, indicating a satisfactory performance of asymptotic results. All coverage rates are close to the nominal level. Mean squared error drops 17%, 35%, 15% and 37% for α, β_1, β_2 and N respectively when we compare CS to Semi-Nor method. Based on these observations, we conclude that GMM cannot improve upon CS in estimation efficiency but the two semiparametric methods outperform CS.

		α	β_1	β_2	N
	true	-1.0	1.0	0.3	300
CS	estimate	-1.0187	1.0226	0.3119	306.45
	emp se	0.1884	0.1422	0.1974	27.34
	mse	0.0699	0.0402	0.0775	1604
	est se	0.1834	0.1377	0.1951	25.12
	95% cov	94.9%	93.9%	95.3%	94.1%
GMM1	estimate	-1.0469	1.0493	0.3200	311.41
	emp se	0.1963	0.1573	0.1987	31.99
	mse	0.0758	0.0495	0.0767	2497.2
	est se	0.1823	0.1354	0.1909	27.57
	95% cov	92.6%	93.2%	94.2%	95.1%
GMM2	estimate	-1.0424	1.0448	0.3203	310.30
	emp se	0.1951	0.1557	0.2006	30.70
	mse	0.0737	0.0446	0.0784	1968.2
	est se	0.1826	0.1330	0.1934	26.32
	95% cov	93.2%	93.0%	94.7%	95.1%
Semi-Nor	estimate	-1.0083	1.0084	0.3090	303.34
	emp se	0.1722	0.1121	0.1813	21.73
	mse	0.0582	0.0263	0.0659	1014.6
	est se	0.1678	0.1159	0.1809	21.59
	95% cov	94.7%	95.6%	95.3%	95.0%
Semi-Uni	estimate	-1.0085	1.0088	0.3092	303.42
	emp se	0.1721	0.1123	0.1811	21.77
	mse	0.0582	0.0264	0.0658	1019.2
	est se	0.1679	0.1163	0.1810	21.64
	95% cov	94.8%	95.9%	95.5%	94.9%

Table 3: Simulation 6 (two covariates). Performance of five methods based on conditional score (CS), two types of GMM (GMM1, GMM2), in comparison with the semiparametric estimator using normal $f_X(x)$ (Semi-Nor) and using uniform $f_X(x)$ (Semi-Uni). The mean of the estimates (estimate), empirical standard error (emp se), mean squared error (mse), average of estimated standard error (est se) and the sample coverage rate of the 95% confidence interval (95% cov) are reported.

S4 Simulation 7

Simulation 7 uses the same data generation procedure as simulation 6 with the population size being $N = 500$. The averaged first and second time captures are 388 and 253. The estimated $\widehat{\Sigma}$ has bias -0.0032 and variance 0.0010. We summarize the results in Table 4. We report the mean, the sample standard error of 1000 estimates, the average of 1000 estimated standard errors, 95% coverage rate and mean squared error. The means of the estimates of α, β_1, β_2 and N for five methods have small biases. The sample and average of 1000 standard errors are close to each other. The coverage rates are around 95% nominal level. The mse drops 13%, 26%, 11% and 20% respectively for α, β_1, β_2 and N when we compare Semi-Nor to CS. The conclusion based on Table 4 is very similar to that of Simulation 6.

		α	β_1	β_2	N
	true	-1.0	1.0	0.3	500
CS	estimate	-1.0092	1.0121	0.3058	505.89
	emp se	0.1422	0.1023	0.1493	30.42
	mse	0.0403	0.0219	0.0449	1988.9
	est se	0.1410	0.1054	0.1501	30.29
	95% cov	95.2%	96.2%	95.4%	95.2%
GMM1	estimate	-1.0240	1.0262	0.3098	509.68
	emp se	0.1443	0.1049	0.1482	33.61
	mse	0.0412	0.0230	0.0437	3134.5
	est se	0.1387	0.1011	0.1466	31.69
	95% cov	94.2%	94.6%	94.8%	96.8%
GMM2	estimate	-1.0211	1.0239	0.3097	508.81
	emp se	0.1446	0.1032	0.1506	31.66
	mse	0.0410	0.0215	0.0449	2135.8
	est se	0.1396	0.1007	0.1485	30.64
	95% cov	94.1%	94.0%	94.9%	96.0%
Semi-Nor	estimate	-1.0029	1.0028	0.3040	502.70
	emp se	0.1353	0.0894	0.1425	27.94
	mse	0.0352	0.0161	0.0399	1600.5
	est se	0.1297	0.0897	0.1398	27.27
	95% cov	93.6%	94.7%	94.4%	94.8%
Semi-Uni	estimate	-1.0031	1.0031	0.3041	502.76
	emp se	0.1352	0.0893	0.1422	27.96
	mse	0.0352	0.0161	0.0399	1603.6
	est se	0.1298	0.0899	0.1399	27.30
	95% cov	93.7%	94.8%	94.4%	94.9%

Table 4: Simulation 7 (two covariates). Performance of five methods based on conditional score (CS), two types of GMM (GMM1, GMM2), in comparison with the semiparametric estimator using normal $f_X(x)$ (Semi-Nor) and using uniform $f_X(x)$ (Semi-Uni). The mean of the estimates (estimate), empirical standard error (emp se), mean squared error (mse), average of estimated standard error (est se) and the sample coverage rate of the 95% confidence interval (95% cov) are reported.

S5 Simulation 8

Simulation 8 uses the same data generation procedure as simulation 6 with the population size being $N = 700$. The averaged first and second time captures are 544 and 354. The estimated $\widehat{\Sigma}$ has bias -0.0026 and variance 0.0007. We summarize the results in Table 5. We report the mean, the sample standard error of 1000 estimates, the average of 1000 estimated standard errors, 95% coverage rate and mean squared error. The means of the estimates of α, β_1, β_2 and N for five methods have small biases. The sample and average of 1000 standard errors are close to each other. The coverage rates are around 95% nominal level. The mse drops 14%, 28%, 15% and 19% respectively for α, β_1, β_2 and N when we compare Semi-Nor to CS. Both GMM1 and GMM2 perform equivalently well as the CS method. The two semiparametric methods outperform the other three methods.

		α	β_1	β_2	N
	true	-1.0	1.0	0.3	700
CS	estimate	-1.0043	1.0096	0.3012	705.54
	emp se	0.1191	0.0887	0.1289	34.93
	mse	0.0285	0.0159	0.0327	2571.3
	est se	0.1190	0.0889	0.1267	34.97
	95% cov	95.6%	94.9%	94.3%	94.3%
GMM1	estimate	-1.0132	1.0179	0.3035	708.32
	emp se	0.1189	0.0856	0.1280	35.03
	mse	0.0279	0.0147	0.0317	2594.8
	est se	0.1161	0.0839	0.1236	34.79
	95% cov	95.6%	94.6%	93.8%	95.8%
GMM2	estimate	-1.0117	1.0170	0.3033	707.97
	emp se	0.1197	0.0867	0.1295	35.26
	mse	0.0284	0.0150	0.0325	2617.68
	est se	0.1176	0.0847	0.1254	34.92
	95% cov	95.2%	94.6%	94.1%	95.6%
Semi-Nor	estimate	-1.0019	1.0009	0.3040	702.22
	emp se	0.1118	0.0753	0.1181	31.72
	mse	0.0245	0.0114	0.0279	2074.8
	est se	0.1094	0.0757	0.1179	31.75
	95% cov	94.8%	95.0%	95.3%	94.2%
Semi-Uni	estimate	-1.0021	1.0012	0.3041	702.29
	emp se	0.1117	0.0754	0.1181	31.74
	mse	0.0245	0.0115	0.0279	2078.7
	est se	0.1094	0.0758	0.1180	31.78
	95% cov	94.7%	95.2%	95.3%	94.2%

Table 5: Simulation 8 (two covariates). Performance of five methods based on conditional score (CS), two types of GMM (GMM1, GMM2), in comparison with the semiparametric estimator using normal $f_X(x)$ (Semi-Nor) and using uniform $f_X(x)$ (Semi-Uni). The mean of the estimates (estimate), empirical standard error (emp se), mean squared error (mse), average of estimated standard error (est se) and the sample coverage rate of the 95% confidence interval (95% cov) are reported.

S6 Simulation 9

In simulation 9, we set population size to be $N = 500$. We consider a bivariate covariate $\mathbf{X}_i = (X_{i1}, X_{i2})^{\mathrm{T}}$, where X_{i1} and X_{i2} are generated from a standard normal and Bernoulli distribution respectively. We then set the measurement error standard deviation $\sigma_u = 0.6$ for X_{i1}. We generate the observations $(Y_{ij}, \mathbf{W}_{ij}Y_{ij})$, $j = 1, 2, 3$ from the model with true parameter values $\alpha = 0.2$ and $\boldsymbol{\beta} = (1.0, 0.5)^{\mathrm{T}}$. It yields an average of 476 first time captures and 420 second time captures. The estimated $\widehat{\Sigma}$ has bias -0.0004 and variance 0.0006.

We summarize the results in Table 6 where we report the mean, the sample standard error of 1000 estimates, the average of 1000 estimated standard errors, 95% coverage rate and mean squared error. The means of α, β_1, β_2 and N have negligible biases. The sample standard error of 1000 estimates and the average of 1000 estimated standard errors are close to each other, indicating a satisfactory performance of asymptotic results. All coverage rates are close to the nominal level. Mean squared error drops 7%, 10%, 8% and 1% for α, β_1, β_2 and N respectively when we compare CS to GMM1 method. Those drops are 18%, 25%, 16% and 11% when comparing CS and Semi-Nor. Based on these observations, we conclude that GMM improve upon CS in estimation efficiency in the settings of high capture probability and moderate sample size. But the improvement is not as large as that of the two semiparametric methods.

		α	β_1	β_2	N
	true	0.2	1.0	0.5	500
CS	estimate	0.2006	1.0037	0.5012	500.74
	emp se	0.1069	0.0843	0.1260	8.5421
	mse	0.0224	0.0144	0.0316	146.64
	est se	0.1045	0.0852	0.1251	8.1810
	95% cov	95.0%	95.0%	95.1%	94.0%
GMM1	estimate	0.1989	1.0110	0.5043	501.05
	emp se	0.1036	0.0806	0.1209	8.4823
	mse	0.0209	0.0130	0.0292	144.97
	est se	0.1004	0.0798	0.1203	8.1402
	95% cov	94.6%	95.1%	95.2%	94.1%
GMM2	estimate	0.2012	1.0096	0.5026	500.94
	emp se	0.1061	0.0812	0.1236	8.5100
	mse	0.0218	0.0132	0.0304	145.24
	est se	0.1024	0.0802	0.1227	8.1358
	95% cov	94.6%	95.1%	95.4%	94.1%
Semi-Nor	estimate	0.2048	1.0021	0.4967	500.38
	emp se	0.0951	0.0750	0.1153	8.1415
	mse	0.0183	0.0108	0.0265	131.05
	est se	0.0958	0.0716	0.1148	7.7479
	95% cov	94.1%	93.2%	94.3%	93.8%
Semi-Uni	estimate	0.2049	1.0021	0.4966	500.37
	emp se	0.0951	0.0751	0.1154	8.1398
	mse	0.0183	0.0108	0.0266	131
	est se	0.0959	0.0717	0.1149	7.7466
	95% cov	94.4%	93.5%	94.3%	93.6%

Table 6: Simulation 9 (two covariates). Performance of five methods based on conditional score (CS), two types of GMM (GMM1, GMM2), in comparison with the semiparametric estimator using normal $f_X(x)$ (Semi-Nor) and using uniform $f_X(x)$ (Semi-Uni). The mean of the estimates (estimate), empirical standard error (emp se), mean squared error (mse), average of estimated standard error (est se) and the sample coverage rate of the 95% confidence interval (95% cov) are reported.

S7 Simulation 10

Simulation 10 uses the same data generation procedure as simulation 9 with the population size being $N = 700$. The averaged first and second time captures are 667 and 587. The estimated $\widehat{\Sigma}$ has bias -0.0007 and variance 0.0004. We summarize the results in Table 7. We report the mean, the sample standard error of 1000 estimates, the average of 1000 estimated standard errors, 95% coverage rate and mean squared error. The biases of the mean of the estimates are negligible. The sample and average of 1000 standard errors are close to each other. The coverage rates are around 95% nominal level. The mse drops 6%, 10%, 7% and 0.4% respectively for α, β_1, β_2 and N when we compare GMM1 to CS. The drops are 14%, 27%, 16% and 10% for Semi-Nor. We see that both GMM and semiparametric methods improve on the CS method. We will conclude similarly as Simulation 9.

		α	β_1	β_2	N
	true	0.2	1.0	0.5	700
CS	estimate	0.2020	1.0033	0.4993	700.68
	emp se	0.0886	0.0731	0.1066	10.0591
	mse	0.0156	0.0106	0.0225	202.27
	est se	0.0880	0.0718	0.1054	9.6515
	95% cov	94.2%	94.9%	94.7%	94.7%
GMM1	estimate	0.1998	1.0081	0.5021	701.02
	emp se	0.0865	0.0702	0.1034	10.0805
	mse	0.0147	0.0095	0.0210	201.44
	est se	0.0846	0.0672	0.1014	9.5871
	95% cov	95.1%	94.3%	94.3%	95.2%
GMM2	estimate	0.2017	1.0071	0.5006	700.89
	emp se	0.0882	0.0707	0.1053	10.0558
	mse	0.0153	0.0096	0.0218	200.42
	est se	0.0863	0.0676	0.1034	9.5828
	95% cov	94.5%	94.3%	94.3%	95.3%
Semi-Nor	estimate	0.2003	1.0010	0.5013	700.41
	emp se	0.0825	0.0632	0.0976	9.6491
	mse	0.0134	0.0077	0.0189	182.61
	est se	0.0810	0.0606	0.0970	9.1938
	95% cov	94.3%	93.5%	95.4%	94.5%
Semi-Uni	estimate	0.2004	1.0010	0.5011	700.41
	emp se	0.0825	0.0632	0.0977	9.6547
	mse	0.0134	0.0077	0.0190	182.7
	est se	0.0810	0.0606	0.0971	9.1922
	95% cov	94.3%	93.5%	95.4%	94.5%

Table 7: Simulation 10 (two covariates). Performance of five methods based on conditional score (CS), two types of GMM (GMM1, GMM2), in comparison with the semiparametric estimator using normal $f_X(x)$ (Semi-Nor) and using uniform $f_X(x)$ (Semi-Uni). The mean of the estimates (estimate), empirical standard error (emp se), mean squared error (mse), average of estimated standard error (est se) and the sample coverage rate of the 95% confidence interval (95% cov) are reported.