Supplementary Materials for Chapter 8: Small Sample Methods of The Handbook of Statistical Methods for Case-Control Studies

Jinko Graham, Brad McNeney and Robert Platt

Here we collect supplementary material referenced in the chapter "Small Sample Methods" of the Handbook of Statistical Methods for Case Control Studies (Graham *et al.*, 2018). In Section 1, we describe how non-interval confidence regions can result from inverting a two-sided exact conditional hypothesis test. In Section 2, we give details about using `elrm` for approximate exact inference. Finally, in Section 3, we discuss previous simulation studies and our own limited simulations to compare the different methods considered in the chapter.

Contents

1 Exact inference: non-interval confidence regions

As noted in the chapter, one strategy for constructing a confidence interval based on exact conditional distributions is to invert a two-sided exact conditional hypothesis test. However, as we now illustrate, this strategy can lead to a non-interval confidence region.

A 95% confidence region for β_1 based on inverting the conditional probabilities test, and accounting for matched sets, would be the set of all β_1^0 values such that the conditional probabilities test of $H_0 : \beta_1 = \beta_1^0$ does not reject the null hypothesis at the 5% level. This set is equivalent to all β_1^0 values that give a p-value of 0.05 or more. Figure 1 shows the $\log_{10}$ p-values from the test of the DES effect accounting for matched sets. Values of β_1^0 greater than or equal to 1.462 have p-values of 0.05 or more, suggesting a 95% confidence interval of $(1.462, \infty)$. However, as shown

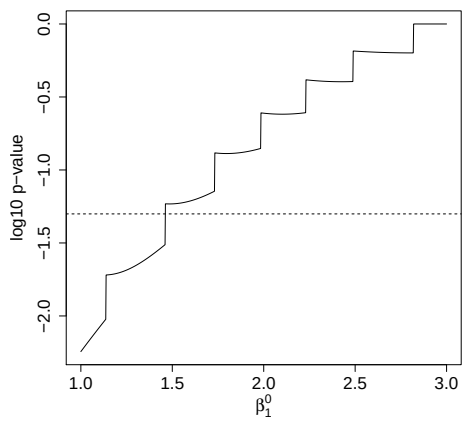

Figure 1: Log_{10} p-values from the exact conditional probabilities test of a DES effect. The dashed horizontal line is at $p = 0.05$ and the dashed vertical line is at $\beta_1^0 = 1.46$.

in Figure 2, the p-value curve need not increase, which allows for the possibility of a non-interval confidence region. For example, with a p-value threshold of 0.0585, marked by the dashed line in Figure 2, the confidence region is $(1.462, 1.474) \cup (1.494, \infty)$ and has a gap from 1.474 to 1.494.

2 Approximate exact inference: running `elrm`

The MCMC method of Zamar *et al.* (2007), has been implemented in the `elrm` R package. Analysis of unmatched or matched data is as follows.

2.1 Unmatched data

For unmatched data, users should:

1. Identify the unique covariate vector values $\mathbf{x}_{(1)}, \ldots, \mathbf{x}_{(q)}$ and stack these vectors as rows of a matrix, $\mathbf{x}$.

2. Count the number of subjects, m_j, with covariate vector value $\mathbf{x}_{(j)}$, for $j = 1, \ldots, q$, and combine $m_1, \ldots, m_q$ into a vector $\mathbf{m}$.

3. Count the number of *cases*, y_j, with covariate vector value $\mathbf{x}_{(j)}$, for $j = 1, \ldots, q$, and combine $y_1, \ldots, y_q$ into a vector $\mathbf{y}$.

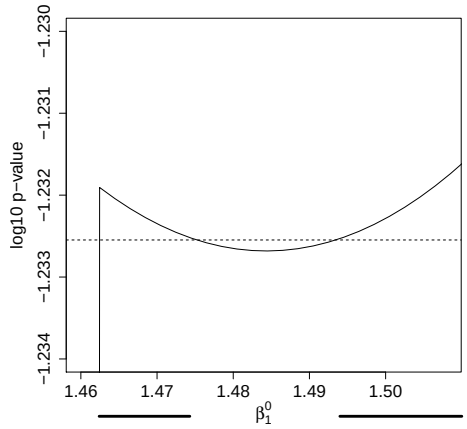

Figure 2: Zoom of $\log_{10}$ p-values for β_1^0 near 1.5. The dashed line corresponds to a p-value of 0.0585. The intervals marked below the horizontal axis comprise the 94.15% confidence region $(1.462, 1.474) \cup (1.494, \infty)$.

The software requires as input a data frame that includes **m**, **y** and **x**. For the DES data, the data frame is as follows (the order of the columns is not important).

```
 m y DES matern.smoke
 6 6   1            1
22 1   0            1
11 0   0            0
 1 1   1            0
```

The first row indicates that there are $m_1 = 6$ subjects who were exposed to DES (DES=1) and maternal smoking (matern.smoke=1) and that all $y_1 = 6$ are cases. The second row indicates that there are $m_2 = 22$ subjects who were not DES exposed (DES=0) but were exposed to maternal smoking (matern.smoke=1) and that only $y_2 = 1$ is a case. The final two rows have similar interpretations.

Users should next specify a logistic model for **y** successes out of **m** trials, given covariates **x**. Users also specify the model terms of interest, with all others being considered nuisance parameters. Details on how the logistic regression model and the model terms of interest are specified are given in the `elrm` documentation. By default, `elrm` includes an intercept term in the logistic regression model, so users do not need to include one explicitly for case-control data.

2.2 Matched data

For matched data, `elrm` users should create a stratum variable, as an R factor, and include this variable in their logistic regression model specification. Lastly, `elrm` users need to take the matched design into account when preparing the data, by determining unique covariate vector values separately for each matched set. For example, the prepared DES data on the first and fifth matched sets are given below, followed by an explanation of each row.

```
strata m y DES matern.smoke
     1 1 1   1            1
     1 2 0   0            1
```

```
1 2 0   0           0
...
5 4 1   0           1
5 1 0   0           0
...
```

The first three rows are for the three unique combinations of the `DES` and `matern.smoke` variables within the first matched set. The first of these rows indicates that there is one subject in the matched set who was exposed to DES and maternal smoking, and this subject is the case. The second row indicates that there are two subjects in the first matched set who were exposed to maternal smoking but not DES, and neither is a case, etc. The last two displayed rows are for the two unique combinations of `DES` and `matern.smoke` within the fifth matched set. The fourth row indicates that there are four individuals in the fifth matched set who were exposed to maternal smoking but not DES and one is the case, etc.

3 Simulations

In this section, we review the results of simulation studies to compare the small-sample methods discussed in the chapter. We first summarize two previous simulation studies and then describe limited simulations of our own to include methods that were not considered in previous comparisons. The methods we consider are summarized in Table 1. We do not consider the Monte Carlo approximate exact conditional methods because exact analysis was always computationally feasible for the sample sizes and numbers of covariates in our investigation. We also do not consider McCullagh's approximate exact method. While McCullagh's method has an extension to regression models (McCullagh, 1984), we are not aware of a readily available implementation.

	Unmatched data	Matched data
maximum likelihood	LR	CLR
exact	ELR	ECLR
penalized	NA	FCLR
	NA	HOA
	FLR	FLR − S
	ZLR	NA
	log-$F(m,m)$	NA

LR, logistic regression; CLR, conditional logistic regression; ELR, exact logistic regression; ECLR, exact conditional logistic regression; FCLR, Firth conditional logistic regression; HOA, HOA-penalized logistic regression; FLR, Firth logistic regression; FLR-S, stratified Firth logistic regression; ZLR, Zhang logistic regression; log-$F(m,m)$: log-$F(m,m)$-penalized logistic regression; NA, not available

Table 1: Methods considered in simulation studies. Rows indicate correspondences between methods for unmatched and matched data. NA means no corresponding method.

Before comparing the methods, it is helpful to consider the information in the data about the parameter of interest, β_1. The information about β_1 increases with sample size and decreases with the number of nuisance covariates. What is perhaps less clear is that, for a fixed sample size and number of nuisance covariates, information about β_1 decreases as the data-generating value of β_1 increases in magnitude.

We illustrate this phenomenon with a simple example in which the covariate of interest is positively associated with case status. Consider an unmatched case-control study with 25 cases and controls each and a single binary covariate X_1. Suppose that the distribution, g, of X_1 in controls is known, so that the only model parameter is the log-odds-ratio β_1 for X_1. One can show that the likelihood for β_1 depends on the data only through n_{11}, the number of cases having $X_1 = 1$. Log-likelihood curves (up to an additive constant) for β_1 at different values of n_{11} are shown in Figure 3. The likelihood is flatter for larger values of n_{11}, which implies less information in the data about the corresponding values of β_1. The question then is what values of β_1 tend to give larger n_{11} values? Starting from the model for unstratified case-control data:

$$\begin{aligned} \mathrm{P}(\mathbf{x}|D=0) &= g(\mathbf{x}) \\ \mathrm{P}(\mathbf{x}|D=1) &= c(\boldsymbol{\beta}, g)\exp(\mathbf{x}^T\boldsymbol{\beta})g(\mathbf{x}), \end{aligned} \tag{1}$$

where $c(\boldsymbol{\beta}, g)$ is a normalizing constant, one can show that $\mathrm{P}(X_1 = 1|D = 1) \propto \exp(\beta_1)$. Hence, larger values of β_1 tend to give larger values of n_{11}, and therefore provide less information about the parameter of interest than smaller values about β_1. In the simulation results, we will see that decreasing information about β_1 affects both the precision of inference (e.g., power) and bias of point estimators.

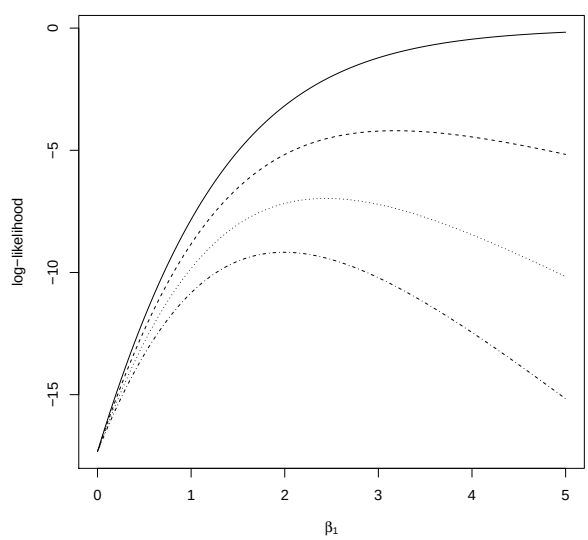

Figure 3: Log-likelihood curves for the simple example described in the text under different numbers, n_{11}, of cases having $X_1 = 1$. Solid line, $n_{11} = 25$, dashed line $n_{11} = 24$, dotted line $n_{11} = 23$ and dot-dash line $n_{11} = 22$.

3.1 Heinze and Schemper (2002)

This study compared the LR, FLR and ELR point estimators for analysis of unmatched case-control data. The ELR estimator was taken to be the CMLE if it existed or the MUE if not due to separation, as is done in LogXact. The authors investigated the performance of estimators for different sample sizes (30 or 100), number of binary covariates (3 or 10), odds-ratios (1, 2, 4 or 16) and case:control ratios (1:1 or 1:4). They obtained similar results for 1:1 and 1:4 case:control

ratios. Their results highlight failures of LR due to separation. Even when the LR estimator exists, it is notably biased. The bias increases as the information in the data about β_1 decreases; i.e., as the sample size decreases, the number of nuisance covariates increases and/or β_1 increases. ELR with LogXact was often not possible, either because of computational limits for the larger sample size or because of degenerate distributions in the case of the smaller sample size and larger number of covariates. The FLR estimator never failed and had the lowest bias among the estimators considered, though there was a negative bias in low information settings, because with a relatively flat likelihood the penalty term begins to dominate. The authors also compared Wald-based and penalized likelihood-based interval estimators for FLR. The penalized likelihood-based confidence intervals achieved the nominal coverage level in all simulation configurations while Wald-based intervals did not. Heinze and Schemper's overall recommendation was therefore to use Firth logistic regression with penalized likelihood-based interval estimators.

3.2 Heinze and Puhr (2010)

This study compared the CLR, FCLR, FLR-S and ECLR methods for the analysis of matched case-control data. ECLR point estimates were taken to be the CMLE if it existed or the MUE if not. For maximum likelihood and maximum penalized likelihood methods, confidence intervals and tests were based on likelihood-ratio statistics or penalized-likelihood-ratio statistics. Exact confidence intervals and tests were based on either conventional p-values (ECLR-conv) or mid-p-values (ECLR-mid-p). The simulation configurations included case:control ratios of 1:1 and 1:4 and odds ratios for two binary covariates of $\beta_1 = \boldsymbol{\beta}_2 = 2$ or $\beta_1 = 4$, $\boldsymbol{\beta}_2 = 1.5$. The sample size was 100. The frequency of data sets in which separation occurred ranged from 0 to 13.8 percent across the simulation setting. Degenerate distributions for exact inference were rare, occurring in less than one percent of simulated data sets. When separation did not occur, CLR performed quite well, but not as well as FCLR, FLR-S and ECLR. The bias of FCLR tended to be the lowest of the methods considered across all simulation scenarios. Coverage of FCLR and ECLR-mid-p confidence intervals for FCLR achieved the nominal level, while coverage of FLR-S and ECLR-conv intervals exceeded the nominal level. Power of FCLR and ECLR-mid-p was higher than FLR-S and ECLR-conv. Heinze and Puhr's overall recommendation was therefore to use Firth conditional logistic regression.

3.3 Additional simulations

To augment the simulations of Heinze and co-workers and to enable comparison of methods that they did not consider, we performed additional simulations. For unmatched data, we added ZLR and log-$F(m, m)$ with $m = 1$ and 2. For matched data, we added HOA.

3.3.1 Design of the simulation study

We simulated both unmatched and matched data, so that there are essentially two simulation studies, with the same configurations of simulation parameters within each study. We simulated a 1:4 ratio of cases:controls throughout. We considered a range of smaller sample sizes than Heinze and co-workers ($n = 25$, 50, or 75), and considered the effect of varying both the size of a regression parameter of interest ($\beta_1 = 0$, 1 or 2) and the number of nuisance covariates ($p = 0$, 2 or 4). We fixed the effect of the p nuisance covariates to be 1 throughout.

We simulated continous, rather than binary covariates. In particular, we set the distribution of covariates in controls to be a standard normal. Once can then show that for effect β_1 the distribution of covariates in cases is normal with mean β_1 and sd 1. In simulation configurations

with multiple covariates the covariates were simulated independently. For exact analyses, binning of covariates is necessary to avoid degenerate exact conditional distributions; we chose to bin by rounding covariate values to the nearest number of SDs from the mean (Brazzale and Davison, 2008).

For each simulation configuration we simulated 1000 datasets. Exact conditional inference with LogXact was difficult to run many times for the simulation study and we ended up analyzing only 250 of the 1000 datasets from each simulation configuration. For each method we estimated the bias and mean squared error (MSE) of point estimators, coverage of 95% confidence intervals and type I error/power of tests of $H_0 : \beta_1 = 0$ *versus* $H_a : \beta_1 \neq 0$. For maximum likelihood and maximum penalized likelihood methods, confidence intervals and tests were based on likelihood-ratio statistics or penalized-likelihood-ratio statistics. Exact confidence intervals and tests were based on mid-p-values from the conditional probabilities test.

3.3.2 Overview of results

Unless specifically noted, all methods had confidence intervals that achieved the nominal coverage level and tests that controlled type I error. Our results concerning LR and CLR are in agreement with those of Heinze and Schemper (2002) and Heinze and Puhr (2010), respectively, which have been discussed previously.

For unmatched data, we found that log-$F(2,2)$ performed best of all methods considered in terms of absolute bias and mean squared error (MSE) of the point estimator, as well as power. FLR and ZLR were very similar, so much so that improving on our crude implementation of ZLR seems unecessary. Of the penalized methods for unmatched data, log-$F(2,2)$ penalized the most, followed by FLR/ZLR, while log-$F(1,1)$ penalized the least. ELR inference frequently failed due to degenerate exact conditional distributions when $p = 4$. ELR had lower power than the penalized methods, which may be partly explained by the need to discretize the covariates.

For matched data, our simulation results suggested that penalized methods are preferable to exact methods, but within the class of penalized methods our results were inconclusive. There was a suggestion that, for $n = 25$ or 50, FLR-S may be the best approach while for $n = 75$, either HOA or FCLR may be best. The preference for FLR-S at very small sample sizes reflects the poor performance of the other approaches: HOA appeared to have poor coverage for confidence intervals and inflated type I error, FCLR did not always converge in the default number of iterations when there were nuisance covariates, and ECLR suffered from degenerate exact conditional distributions when there were nuisance covariates. At $n = 75$, all approaches appeared to give adequate coverage of confidence intervals and to control type I error. The penalized approaches all had comparable power, which was superior to ECLR. The FCLR point estimator appeared to be the best in terms of absolute bias, while the HOA point estimator appeared to have the best MSE. The inconclusive results suggest the need for a larger-scale simulation study to better evaluate existing methods, as well as the potential for developing alternate penalized methods based on log-$F(m,m)$ priors.

Further details of the simulation results are given in the next sections.

3.3.3 Results for unmatched data

We show a series of tables for the methods log-$F(2,2)$, log-$F(1,1)$, FLR, ZLR, ELR and LR defined in Table 1 of this document.

Table 2: log-$F(2,2)$-penalized logistic regression for unmatched data

n	β_1	p	p.fail	bias	SD	MSE	coverage	power
25	0	0	0.000	-0.008	0.523	0.273	0.964	0.036
25	0	2	0.000	0.027	0.574	0.331	0.979	0.021
25	0	4	0.000	0.019	0.591	0.349	0.992	0.008
25	1	0	0.000	-0.036	0.585	0.343	0.957	0.457
25	1	2	0.000	-0.043	0.578	0.336	0.982	0.302
25	1	4	0.000	-0.123	0.563	0.332	0.986	0.147
25	2	0	0.000	-0.178	0.540	0.323	0.969	0.973
25	2	2	0.000	-0.335	0.494	0.356	0.987	0.821
25	2	4	0.000	-0.514	0.471	0.486	0.987	0.541
50	0	0	0.000	0.006	0.369	0.136	0.949	0.051
50	0	2	0.000	0.026	0.439	0.193	0.952	0.048
50	0	4	0.000	0.008	0.510	0.261	0.961	0.039
50	1	0	0.000	0.011	0.425	0.181	0.962	0.781
50	1	2	0.000	0.017	0.489	0.239	0.955	0.631
50	1	4	0.000	0.010	0.523	0.274	0.979	0.454
50	2	0	0.000	-0.070	0.527	0.283	0.965	1.000
50	2	2	0.000	-0.125	0.533	0.300	0.968	0.989
50	2	4	0.000	-0.234	0.507	0.312	0.972	0.917
75	0	0	0.000	0.012	0.285	0.081	0.955	0.045
75	0	2	0.000	-0.004	0.353	0.125	0.957	0.043
75	0	4	0.000	0.009	0.425	0.180	0.952	0.048
75	1	0	0.000	0.006	0.349	0.122	0.952	0.932
75	1	2	0.000	0.002	0.419	0.176	0.952	0.783
75	1	4	0.000	0.007	0.472	0.222	0.970	0.622
75	2	0	0.000	-0.021	0.498	0.248	0.963	1.000
75	2	2	0.000	-0.017	0.528	0.279	0.969	1.000
75	2	4	0.000	-0.063	0.507	0.261	0.987	0.996

n, sample size; β_1, simulated regression effect of interest; p, number of nuisance covariates; p.fail, proportion of simulated data sets for which the method failed; bias, estimated bias of estimator; SD, sample standard deviation of estimates; MSE, estimated mean-squared error of estimator; coverage, estimated coverage of 95% confidence intervals (CIs); power, estimated power to test $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$. Results are based on 1000 simulated datasets.

- When $\beta_1 = 2$ and $p = 4$, the log-$F(2,2)$ estimator is biased towards zero and this bias increases as the sample size decreases. This suggests that the estimator can shrink too much when the likelihood is relatively flat.
- The log-$F(2,2)$ CIs achieve nominal coverage and the test controls type I error, to within the simulation error of about 0.014.

Table 3: log-$F(1,1)$-penalized logistic regression for unmatched data

n	β_1	p	p.fail	bias	SD	MSE	coverage	power
25	0	0	0.000	-0.008	0.580	0.336	0.960	0.040
25	0	2	0.000	0.032	0.737	0.545	0.955	0.045
25	0	4	0.000	0.021	0.817	0.668	0.982	0.018
25	1	0	0.000	0.079	0.716	0.520	0.944	0.480
25	1	2	0.000	0.175	0.795	0.663	0.973	0.348
25	1	4	0.000	0.134	0.809	0.672	0.984	0.181
25	2	0	0.000	0.184	0.823	0.712	0.982	0.975
25	2	2	0.000	0.111	0.754	0.580	0.995	0.825
25	2	4	0.000	-0.094	0.686	0.479	0.992	0.547
50	0	0	0.000	0.007	0.384	0.147	0.944	0.056
50	0	2	0.000	0.027	0.496	0.246	0.941	0.059
50	0	4	0.000	0.011	0.640	0.409	0.948	0.052
50	1	0	0.000	0.063	0.468	0.223	0.951	0.793
50	1	2	0.000	0.132	0.598	0.374	0.954	0.652
50	1	4	0.000	0.211	0.710	0.548	0.968	0.490
50	2	0	0.000	0.122	0.707	0.515	0.961	1.000
50	2	2	0.000	0.204	0.782	0.653	0.978	0.987
50	2	4	0.000	0.194	0.759	0.614	0.985	0.911
75	0	0	0.000	0.013	0.292	0.085	0.955	0.045
75	0	2	0.000	-0.004	0.378	0.143	0.949	0.051
75	0	4	0.000	0.011	0.492	0.242	0.945	0.055
75	1	0	0.000	0.037	0.369	0.138	0.947	0.936
75	1	2	0.000	0.067	0.478	0.233	0.941	0.787
75	1	4	0.000	0.132	0.588	0.363	0.956	0.632
75	2	0	0.000	0.109	0.624	0.401	0.950	1.000
75	2	2	0.000	0.232	0.733	0.591	0.956	1.000
75	2	4	0.000	0.323	0.753	0.670	0.985	0.994

n, sample size; β_1, simulated regression effect of interest; p, number of nuisance covariates; p.fail, proportion of simulated data sets for which the method failed; bias, estimated bias of estimator; SD, sample standard deviation of estimates; MSE, estimated mean-squared error of estimator; coverage, estimated coverage of 95% confidence intervals (CIs); power, estimated power to test $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$. Results are based on 1000 simulated datasets.

- The log-$F(1,1)$ estimator is biased away from zero when $\beta_1>0$, with the exception of the configuration $n = 25$, $\beta_1 = 2$, $p = 4$.
- The log-$F(1,1)$ estimator has higher MSE than the log-$F(2,2)$ estimator.
- The log-$F(1,1)$ CIs achieve nominal coverage and the test controls type I error, to within simulation error.
- The log-$F(1,1)$ test has lower power than the log-$F(2,2)$ approach.
- We conclude that log-$F(1,1)$ is inferior to log-$F(2,2)$ in these simulations.

Table 4: Firth-penalized logistic regression (FLR) for unmatched data

n	β_1	p	p.fail	bias	SD	MSE	coverage	power
25	0	0	0.000	-0.007	0.537	0.288	0.968	0.032
25	0	2	0.000	0.030	0.702	0.494	0.977	0.023
25	0	4	0.001	0.031	0.727	0.529	0.992	0.008
25	1	0	0.000	-0.016	0.717	0.515	0.948	0.444
25	1	2	0.000	-0.039	0.798	0.639	0.975	0.245
25	1	4	0.002	-0.266	0.716	0.583	0.978	0.074
25	2	0	0.000	-0.049	0.905	0.822	0.958	0.968
25	2	2	0.000	-0.486	0.670	0.685	0.967	0.695
25	2	4	0.003	-0.966	0.535	1.219	0.968	0.293
50	0	0	0.000	0.006	0.365	0.133	0.950	0.050
50	0	2	0.000	0.022	0.474	0.225	0.953	0.047
50	0	4	0.000	0.021	0.653	0.427	0.961	0.039
50	1	0	0.000	0.007	0.451	0.204	0.956	0.779
50	1	2	0.000	0.021	0.597	0.357	0.948	0.598
50	1	4	0.002	-0.007	0.732	0.535	0.966	0.362
50	2	0	0.000	0.014	0.792	0.628	0.950	1.000
50	2	2	0.001	-0.070	0.851	0.729	0.960	0.969
50	2	4	0.003	-0.414	0.707	0.671	0.961	0.825
75	0	0	0.000	0.012	0.282	0.080	0.959	0.041
75	0	2	0.000	-0.004	0.358	0.128	0.959	0.041
75	0	4	0.000	0.012	0.465	0.216	0.952	0.048
75	1	0	0.000	-0.001	0.358	0.128	0.951	0.924
75	1	2	0.000	-0.007	0.507	0.257	0.944	0.763
75	1	4	0.000	-0.005	0.582	0.339	0.956	0.575
75	2	0	0.000	0.034	0.699	0.490	0.946	1.000
75	2	2	0.000	0.054	0.823	0.681	0.956	0.999
75	2	4	0.002	-0.061	0.835	0.701	0.970	0.975

n, sample size; β_1, simulated regression effect of interest; p, number of nuisance covariates; p.fail, proportion of simulated data sets for which the method failed; bias, estimated bias of estimator; SD, sample standard deviation of estimates; MSE, estimated mean-squared error of estimator; coverage, estimated coverage of 95% confidence intervals (CIs); power, estimated power to test $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$. Results are based on 1000 simulated datasets.

- When $\beta_1 = 2$ and $p = 4$, the FLR estimator is biased towards zero and this bias increases as the sample size decreases. This suggests that the estimator can shrink too much when the likelihood is relatively flat. This behavior is in common with the log-$F(2, 2)$ estimator.
- FLR has larger MSE than the log-$F(m, m)$ methods.
- The FLR CIs achieve nominal coverage and the test controls type I error, to within simulation error.
- The power of the FLR test is lower than the power of the log-$F(2, 2)$ test.
- We conclude that FLR is inferior to log-$F(2, 2)$ in these simulations.

Table 5: Zhang-penalized logistic regression (ZLR) for unmatched data

n	β_1	p	p.fail	bias	SD	MSE	coverage	power
25	0	0	0.000	-0.007	0.539	0.291	0.967	0.033
25	0	2	0.000	0.031	0.707	0.500		
25	0	4	0.004	0.032	0.730	0.534		
25	1	0	0.000	-0.013	0.719	0.516	0.948	0.444
25	1	2	0.001	-0.043	0.763	0.584		
25	1	4	0.001	-0.276	0.656	0.507		
25	2	0	0.000	-0.047	0.906	0.823	0.957	0.968
25	2	2	0.002	-0.493	0.632	0.644		
25	2	4	0.003	-0.972	0.498	1.192		
50	0	0	0.000	0.006	0.366	0.134	0.950	0.050
50	0	2	0.000	0.022	0.475	0.226		
50	0	4	0.001	0.025	0.647	0.419		
50	1	0	0.000	0.008	0.451	0.204	0.956	0.779
50	1	2	0.000	0.023	0.598	0.358		
50	1	4	0.001	-0.009	0.721	0.520		
50	2	0	0.000	0.014	0.793	0.629	0.950	1.000
50	2	2	0.000	-0.069	0.851	0.728		
50	2	4	0.000	-0.417	0.701	0.665		
75	0	0	0.000	0.012	0.282	0.080	0.959	0.041
75	0	2	0.000	-0.004	0.358	0.128		
75	0	4	0.000	0.012	0.465	0.217		
75	1	0	0.000	-0.001	0.358	0.128	0.951	0.924
75	1	2	0.000	-0.007	0.508	0.258		
75	1	4	0.000	-0.004	0.583	0.340		
75	2	0	0.000	0.034	0.699	0.490	0.946	1.000
75	2	2	0.000	0.055	0.824	0.682		
75	2	4	0.000	-0.071	0.803	0.650		

n, sample size; β_1, simulated regression effect of interest; p, number of nuisance covariates; p.fail, proportion of simulated data sets for which the method failed; bias, estimated bias of estimator; SD, sample standard deviation of estimates; MSE, estimated mean-squared error of estimator; coverage, estimated coverage of 95% confidence intervals (CIs); power, estimated power to test $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$. Results are based on 1000 simulated datasets.

- The ZLR estimator, CI and test are nearly identical to those of FLR.

Table 6: Exact logistic regression (ELR) for unmatched data.

n	β_1	p	p.CMLE	p.MUE	p.fail	bias	SD	MSE	coverage	power
25	0	0	0.996	0.004	0.000	-0.029	0.553	0.307	0.960	0.028
25	0	2	0.808	0.180	0.012	0.059	0.690	0.480	0.980	0.020
25	0	4	0.112	0.272	0.616	-0.033	0.392	0.154	1.000	0.000
25	1	0	0.960	0.040	0.000	0.045	0.689	0.477	0.940	0.356
25	1	2	0.644	0.336	0.020	-0.149	0.594	0.375	0.971	0.118
25	1	4	0.056	0.392	0.552	-0.790	0.315	0.723	1.000	0.000
25	2	0	0.636	0.364	0.000	0.168	0.564	0.347	0.988	0.948
25	2	2	0.204	0.772	0.024	-0.767	0.539	0.880	0.984	0.504
25	2	4	0.000	0.472	0.528	-1.794	0.271	3.291	1.000	0.008
50	0	0	1.000	0.000	0.000	-0.033	0.358	0.129	0.948	0.040
50	0	2	1.000	0.000	0.000	-0.022	0.510	0.261	0.936	0.056
50	0	4	0.768	0.164	0.068	-0.041	0.608	0.372	0.957	0.021
50	1	0	1.000	0.000	0.000	0.103	0.521	0.282	0.932	0.744
50	1	2	0.972	0.028	0.000	0.039	0.571	0.328	0.964	0.504
50	1	4	0.556	0.396	0.048	-0.154	0.526	0.301	0.979	0.223
50	2	0	0.892	0.108	0.000	0.396	0.704	0.652	0.968	1.000
50	2	2	0.684	0.316	0.000	0.178	0.598	0.389	0.984	0.964
50	2	4	0.144	0.800	0.056	-0.966	0.611	1.305	0.992	0.547
75	0	0	1.000	0.000	0.000	0.035	0.279	0.079	0.964	0.032
75	0	2	1.000	0.000	0.000	0.014	0.390	0.153	0.936	0.056
75	0	4	0.976	0.024	0.000	0.020	0.458	0.210	0.968	0.032
75	1	0	1.000	0.000	0.000	0.015	0.427	0.182	0.932	0.880
75	1	2	0.996	0.004	0.000	0.115	0.557	0.323	0.944	0.744
75	1	4	0.876	0.120	0.004	0.080	0.550	0.309	0.968	0.498
75	2	0	0.968	0.032	0.000	0.384	0.621	0.534	0.964	1.000
75	2	2	0.864	0.136	0.000	0.458	0.696	0.694	0.976	0.996
75	2	4	0.424	0.576	0.000	-0.235	0.551	0.359	0.984	0.928

n, sample size; β_1, simulated regression effect of interest; p, number of nuisance covariates; p.CMLE, proportion of simulated data sets for which the ELR estimator is the CMLE; p.MUE, proportion of simulated data sets for which the ELR estimator is the MUE; p.fail, proportion of simulated data sets for which the ELR estimator is not defined and therefore ELR fails; bias, estimated bias of estimator; SD, sample standard deviation of estimates; MSE, estimated mean-squared error of estimator; coverage, estimated coverage of 95% confidence intervals; power, estimated power to test $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$. Results are based on 250 simulation replicates.

- The ELR estimator is the conditional MLE (CMLE) when it exists and the median unbiased estimate (MUE) otherwise, so long as the exact conditional distribution is not degenerate. In the case of a degenerate exact conditional distribution the ELR estimator is not defined and the method fails.
- ELR fails frequently when $p = 4$ and $n = 25$ at any value of β_1, but otherwise rarely fails.
- In simulation configurations where the ELR estimator tends to be the CMLE (large p.CMLE) and $\beta_1 > 0$ (e.g., $n = 75$, $\beta_1 = 2$, $p = 2$) the bias of the ELR estimator is away from zero, suggesting that the CMLE is biased away from zero when $\beta_1 > 0$.
- In simulation configurations where the ELR estimator tends to be the MUE (large p.MUE) and $\beta_1 > 0$ (e.g., $n = 25$, $\beta_1 = 2$, $p = 2$) the ELR estimator is biased towards zero, suggesting that the MUE is biased towards zero when $\beta_1 > 0$.
- The MSE of the ELR estimator is higher than the MSE of the log-$F(2, 2)$ estimator.
- The ELR CIs achieve nominal level and the test controls type I error, to within simulation error of about 0.028.
- The power of ELR is lower than that of the penalized methods.
- We conclude that ELR is inferior to log-$F(2, 2)$ in these simulations.

Table 7: Logistic regression (LR) for unmatched data

n	β_1	p	p.fail	bias	SD	MSE	coverage	power
25	0	0	0.001	-0.013	0.637	0.407	0.950	0.050
25	0	2	0.161	0.066	1.028	1.061	0.934	0.066
25	0	4	0.654	0.023	1.502	2.257	0.953	0.046
25	1	0	0.008	0.240	0.972	1.002	0.930	0.503
25	1	2	0.258	0.460	1.243	1.756	0.963	0.368
25	1	4	0.767	0.256	1.339	1.859	0.955	0.163
25	2	0	0.150	0.555	1.285	1.958	0.983	0.973
25	2	2	0.588	0.340	1.173	1.491	0.995	0.791
25	2	4	0.923	0.045	1.494	2.233	0.971	0.377
50	0	0	0.000	0.007	0.400	0.160	0.942	0.058
50	0	2	0.006	0.034	0.576	0.333	0.935	0.065
50	0	4	0.170	0.006	0.894	0.798	0.917	0.083
50	1	0	0.000	0.123	0.526	0.292	0.938	0.805
50	1	2	0.030	0.290	0.790	0.708	0.929	0.666
50	1	4	0.305	0.368	0.994	1.123	0.944	0.504
50	2	0	0.023	0.330	0.888	0.898	0.962	1.000
50	2	2	0.208	0.434	0.993	1.175	0.981	0.982
50	2	4	0.649	0.286	0.969	1.020	0.977	0.886
75	0	0	0.000	0.013	0.299	0.090	0.951	0.049
75	0	2	0.000	-0.004	0.412	0.170	0.938	0.062
75	0	4	0.050	0.015	0.601	0.361	0.926	0.074
75	1	0	0.000	0.071	0.394	0.160	0.945	0.938
75	1	2	0.003	0.151	0.556	0.332	0.924	0.792
75	1	4	0.068	0.289	0.763	0.665	0.926	0.638
75	2	0	0.007	0.247	0.733	0.599	0.942	1.000
75	2	2	0.108	0.420	0.829	0.864	0.954	0.999
75	2	4	0.449	0.445	0.841	0.906	0.984	0.991

n, sample size; β_1, simulated regression effect of interest; p, number of nuisance covariates; p.fail, proportion of simulated data sets for which the method failed; bias, estimated bias of estimator; SD, sample standard deviation of estimates; MSE, estimated mean-squared error of estimator; coverage, estimated coverage of 95% confidence intervals (CIs); power, estimated power to test $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$. Results are based on 1000 simulated datasets.

- When $n = 25$ and there are nuisance covariates ($p > 0$) LR has a substantial failure rate due to separation, regardless of the value of β_1.
- LR tends not to fail for $n = 75$, except when $p \geq 2$ and $\beta_1 = 2$.
- The LR estimator is biased away from zero when $\beta_1 > 0$, and the magniture of the bias increases as β_1 increases, p increases and n decreases.
- The LR CIs have poor coverage when the LR estimator is biased away from zero.
- The LR test sometimes has inflated type I error.

3.3.4 Results for matched data

We show a series of tables for the methods FCLR, FLR-S, HOA, ECLR and CLR defined in Table 1 of this document.

Table 8: Firth conditional logistic regression (FCLR) for matched data

n	β_1	p	p.fail	bias	SD	MSE	coverage	power
25	0	0	0.000	0.029	0.599	0.359	0.961	0.039
25	0	2	0.018	-0.002	0.650	0.422	0.982	0.018
25	0	4	0.162	0.008	0.496	0.246	0.999	0.001
25	1	0	0.001	0.003	0.701	0.491	0.973	0.448
25	1	2	0.017	-0.290	0.610	0.456	0.976	0.123
25	1	4	0.180	-0.596	0.454	0.561	0.983	0.016
25	2	0	0.001	-0.426	0.563	0.498	0.950	0.966
25	2	2	0.019	-1.041	0.452	1.288	0.959	0.426
25	2	4	0.159	-1.437	0.289	2.148	0.973	0.056
50	0	0	0.000	0.031	0.379	0.144	0.947	0.053
50	0	2	0.002	0.028	0.540	0.292	0.958	0.042
50	0	4	0.025	0.008	0.558	0.312	0.989	0.011
50	1	0	0.000	0.020	0.534	0.285	0.951	0.762
50	1	2	0.007	-0.092	0.605	0.374	0.976	0.439
50	1	4	0.031	-0.346	0.516	0.386	0.971	0.133
50	2	0	0.000	-0.070	0.788	0.626	0.959	1.000
50	2	2	0.016	-0.630	0.530	0.678	0.946	0.885
50	2	4	0.038	-1.119	0.376	1.394	0.973	0.414
75	0	0	0.000	0.005	0.292	0.085	0.954	0.046
75	0	2	0.000	-0.011	0.413	0.170	0.953	0.047
75	0	4	0.020	0.002	0.536	0.287	0.976	0.026
75	1	0	0.000	0.009	0.368	0.136	0.955	0.932
75	1	2	0.007	0.023	0.591	0.350	0.964	0.676
75	1	4	0.037	-0.176	0.542	0.325	0.971	0.322
75	2	0	0.001	-0.029	0.746	0.558	0.973	1.000
75	2	2	0.017	-0.387	0.552	0.454	0.951	0.974
75	2	4	0.051	-0.910	0.395	0.984	0.963	0.695

n, sample size; β_1, simulated regression effect of interest; p, number of nuisance covariates; p.fail, proportion of simulated data sets for which the method failed; bias, estimated bias of estimator; SD, sample standard deviation of estimates; MSE, estimated mean-squared error of estimator; coverage, estimated coverage of 95% confidence intervals (CIs); power, estimated power to test $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$. Results are based on 1000 simulated datasets.

- In simulation configurations with $p > 0$, and regardless of n and β_1 (e.g., $n = 25$, $\beta_1 = 1$, $p = 4$), FCLR was prone to non-trivial failure rates. Failures were declared when the iterative maximization procedure did not converge in the default number of iterations (100).
- When $\beta_1 = 2$ and $p = 4$, the FCLR estimator is biased towards zero and this bias increases as the sample size decreases. This suggests that the estimator can shrink too much when the likelihood is relatively flat.

- The FCLR CIs achieve nominal coverage and the test controls type I error, to within the simulation error of about 0.014.
- When the FCLR estimator is highly biased towards zero (e.g., $n = 25$, $\beta_1 = 2$, $p = 4$), the power of the FCLR test is very low.

Table 9: Firth logistic regression (FLR-S) for matched data

n	β_1	p	p.fail	bias	SD	MSE	coverage	power
25	0	0	0.000	0.026	0.579	0.336	0.960	0.040
25	0	2	0.003	-0.003	0.693	0.480	0.973	0.027
25	0	4	0.004	-0.007	0.585	0.343	0.997	0.003
25	1	0	0.000	0.051	0.710	0.506	0.974	0.461
25	1	2	0.001	-0.201	0.656	0.471	0.980	0.163
25	1	4	0.002	-0.465	0.530	0.497	0.982	0.041
25	2	0	0.000	-0.372	0.549	0.439	0.954	0.974
25	2	2	0.001	-0.878	0.459	0.981	0.958	0.549
25	2	4	0.001	-1.245	0.326	1.656	0.951	0.134
50	0	0	0.000	0.031	0.386	0.150	0.945	0.055
50	0	2	0.003	0.019	0.506	0.257	0.953	0.047
50	0	4	0.003	-0.000	0.524	0.274	0.988	0.012
50	1	0	0.000	0.049	0.520	0.272	0.948	0.768
50	1	2	0.004	-0.080	0.541	0.299	0.976	0.486
50	1	4	0.007	-0.294	0.475	0.313	0.970	0.191
50	2	0	0.002	-0.199	0.549	0.340	0.961	1.000
50	2	2	0.003	-0.634	0.430	0.587	0.945	0.926
50	2	4	0.002	-0.996	0.355	1.118	0.945	0.578
75	0	0	0.000	0.005	0.296	0.087	0.953	0.047
75	0	2	0.000	-0.011	0.394	0.155	0.950	0.050
75	0	4	0.006	0.000	0.464	0.215	0.979	0.021
75	1	0	0.000	0.041	0.372	0.140	0.952	0.934
75	1	2	0.003	-0.008	0.489	0.239	0.968	0.704
75	1	4	0.003	-0.202	0.427	0.224	0.969	0.364
75	2	0	0.002	-0.185	0.456	0.242	0.971	1.000
75	2	2	0.007	-0.522	0.394	0.428	0.933	0.987
75	2	4	0.002	-0.880	0.304	0.867	0.923	0.842

n, sample size; β_1, simulated regression effect of interest; p, number of nuisance covariates; p.fail, proportion of simulated data sets for which the method failed; bias, estimated bias of estimator; SD, sample standard deviation of estimates; MSE, estimated mean-squared error of estimator; coverage, estimated coverage of 95% confidence intervals (CIs); power, estimated power to test $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$. Results are based on 1000 simulated datasets.

- The FLR-S approach rarely failed to converge.
- When $\beta_1 = 2$ and $p = 4$, the FLR-S estimator is biased towards zero and this bias increases as the sample size decreases. This suggests that the estimator can shrink too much when the likelihood is relatively flat.
- The magnitude of the FLR-S bias tends to be less than that of FCLR when $n = 25$ or 50, but tends to be greater than that of FCLR when $n = 75$.
- The MSE of the FLR-S estimator tends to be lower than that of FCLR.

- The FLR-S CIs achieve nominal coverage and the test controls type I error, to within the simulation error.
- The power of the FLR-S test tends to be higher than that of FCLR when $n = 25$ or 50, but the power comparison is inconclusive when $n = 75$.
- We cannot declare either FCLR or FLR-S superior based on these simulations, though there is a suggestion that FLR-S performs better for $n = 25$ or 50, while FCLR performs better for $n = 75$.

Table 10: HOA-penalized likelihood (HOA) for matched data. NB: confidence intervals and tests only implemented for $p = 0$

n	β_1	p	p.fail	bias	SD	MSE	coverage	power
25	0	0	0.000	0.033	0.738	0.546	0.933	0.067
25	0	2	0.005	-0.013	1.370	1.877		
25	0	4	0.095	0.023	2.113	4.466		
25	1	0	0.000	0.309	0.927	0.955	0.947	0.534
25	1	2	0.004	0.329	1.326	1.866		
25	1	4	0.073	0.291	1.886	3.643		
25	2	0	0.000	-0.081	0.639	0.415	0.981	0.981
25	2	2	0.003	-0.260	1.022	1.112		
25	2	4	0.027	-0.575	1.269	1.939		
50	0	0	0.000	0.036	0.438	0.193	0.926	0.074
50	0	2	0.000	0.025	0.613	0.377		
50	0	4	0.000	-0.009	0.781	0.610		
50	1	0	0.000	0.173	0.559	0.342	0.935	0.799
50	1	2	0.000	0.076	0.624	0.396		
50	1	4	0.000	-0.076	0.722	0.527		
50	2	0	0.000	-0.172	0.417	0.204	0.976	1.000
50	2	2	0.000	-0.493	0.445	0.442		
50	2	4	0.000	-0.733	0.518	0.805		
75	0	0	0.000	0.006	0.324	0.105	0.946	0.054
75	0	2	0.001	-0.014	0.443	0.196		
75	0	4	0.000	-0.004	0.513	0.263		
75	1	0	0.000	0.134	0.398	0.177	0.943	0.942
75	1	2	0.000	0.057	0.448	0.204		
75	1	4	0.001	-0.128	0.483	0.249		
75	2	0	0.000	-0.218	0.323	0.152	0.972	1.000
75	2	2	0.000	-0.542	0.343	0.411		
75	2	4	0.000	-0.800	0.347	0.759		

n, sample size; β_1, simulated regression effect of interest; p, number of nuisance covariates; p.fail, proportion of simulated data sets for which the method failed; bias, estimated bias of estimator; SD, sample standard deviation of estimates; MSE, estimated mean-squared error of estimator; coverage, estimated coverage of 95% confidence intervals (CIs); power, estimated power to test $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$. Results are based on 1000 simulated datasets.

- In simulation configurations with $n = 25$, any value of β_1 and $p = 4$ HOA had non-trivial rates of failure due to lack of convergence of the iterative maximizer.
- The bias of the HOA estimator can be either away from zero or towards zero when $\beta_1 > 0$. Bias away from zero is more common when $\beta_1 = 1$ (e.g., $n = 25$, $\beta_1 = 1$, $p = 4$) and bias towards zero is more common when $\beta_1 = 2$ (e.g., $n = 25$, $\beta_1 = 2$, $p = 4$).
- Our conclusions about coverage of CIs and power and level of the test are somewhat limited by the fact that we did not implemented the CIs and tests for $p > 0$. However, from the

simulations with $p = 0$ we see that coverage of the CIs is sometimes less than nominal and the type I error is sometimes greater than nominal (e.g., $n = 25$, $\beta_1 = 0$, $p = 0$).

- We do not recommend HOA because of the possible anti-conservative nature of the CIs and tests.

Table 11: Exact conditional logistic regression (ECLR) for matched data.

n	$beta_1$	p	p.CMLE	p.MUE	p.fail	bias	SD	MSE	coverage	power
25	0	0	0.972	0.028	0.000	-0.034	0.887	0.787	0.968	0.020
25	0	2	0.484	0.372	0.144	0.009	0.568	0.323	0.986	0.009
25	0	4	0.008	0.100	0.892	0.026	0.135	0.019	1.000	0.000
25	1	0	0.836	0.164	0.000	0.367	0.695	0.617	0.968	0.220
25	1	2	0.300	0.560	0.140	-0.447	0.469	0.420	0.991	0.033
25	1	4	0.000	0.080	0.920	-0.927	0.208	0.903	1.000	0.000
25	2	0	0.316	0.684	0.000	0.226	0.394	0.206	1.000	0.876
25	2	2	0.068	0.852	0.080	-1.421	0.403	2.181	0.987	0.083
25	2	4	0.000	0.108	0.892	-1.950	0.101	3.814	1.000	0.000
50	0	0	0.992	0.008	0.000	0.040	0.841	0.710	0.968	0.020
50	0	2	0.928	0.072	0.000	0.018	0.693	0.481	0.964	0.032
50	0	4	0.176	0.300	0.524	0.019	0.458	0.210	1.000	0.000
50	1	0	0.812	0.188	0.000	0.802	0.773	1.242	0.908	0.488
50	1	2	0.740	0.256	0.004	0.221	0.654	0.476	0.984	0.402
50	1	4	0.062	0.434	0.504	-0.706	0.378	0.641	0.992	0.024
50	2	0	0.364	0.636	0.000	0.893	0.446	0.995	0.992	0.988
50	2	2	0.257	0.739	0.004	-0.281	0.502	0.331	0.996	0.887
50	2	4	0.000	0.600	0.400	-1.729	0.275	3.066	1.000	0.067
75	0	0	0.984	0.016	0.000	0.076	0.887	0.792	0.956	0.032
75	0	2	0.992	0.008	0.000	-0.034	0.714	0.511	0.952	0.044
75	0	4	0.600	0.292	0.108	-0.028	0.644	0.415	0.955	0.040
75	1	0	0.856	0.144	0.000	0.795	0.823	1.310	0.892	0.492
75	1	2	0.900	0.100	0.000	0.505	0.748	0.815	0.936	0.564
75	1	4	0.336	0.508	0.156	-0.243	0.547	0.358	0.986	0.175
75	2	0	0.452	0.548	0.000	0.966	0.556	1.242	0.924	0.964
75	2	2	0.460	0.540	0.000	0.397	0.518	0.426	0.988	0.996
75	2	4	0.048	0.820	0.132	-1.289	0.447	1.861	0.991	0.401

n, sample size; β_1, simulated regression effect of interest; p, number of nuisance covariates; p.CMLE, proportion of simulated data sets for which the ECLR estimator is the CMLE; p.MUE, proportion of simulated data sets for which the ECLR estimator is the MUE; p.fail, proportion of simulated data sets for which the ECLR estimator is not defined and therefore ECLR fails; bias, estimated bias of estimator; SD, sample standard deviation of estimates; MSE, estimated mean-squared error of estimator; coverage, estimated coverage of 95% confidence intervals; power, estimated power to test $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$. Results are based on 250 simulation replicates.

- The ECLR estimator is the conditional MLE (CMLE) when it exists and the median unbiased estimate (MUE) otherwise, so long as the exact conditional distribution is not degenerate. In the case of a degenerate exact conditional distribution the ECLR estimator is not defined and the method fails.
- ECLR fails frequently when $n = 25$, any value of β_1 and $p = 2$ or 4. ECLR also has a substantial failure rate when $n = 50$ or 75, any value of β_1 and $p = 4$.
- In simulation configurations where the ECLR estimator tends to be the CMLE (large p.CMLE) and $\beta_1 > 0$ (e.g., $n = 25$, $\beta_1 = 1$, $p = 0$) the bias of the ECLR estimator is away from zero, suggesting that the CMLE is biased away from zero when $\beta_1 > 0$.
- In simulation configurations where the ECLR estimator tends to be the MUE (large p.MUE) and $\beta_1 > 0$ (e.g., $n = 25$, $\beta_1 = 2$, $p = 2$) the ELCR estimator is biased towards zero, suggesting that the MUE is biased towards zero when $\beta_1 > 0$.
- The ECLR CIs achieve nominal level and the test controls type I error, to within simulation error of about 0.028, except for the simulation configuration with $n = 75$, $\beta_1 = 1$ and $p = 0$.
- We do not recommend ECLR for the sample sizes we considered because of the high failure rates when there are nuisance covariates.

Table 12: Conditional logistic regression (CLR) for matched data

n	$beta_1$	p	p.fail	bias	SD	MSE	coverage	power
25	0	0	0.001	0.026	0.695	0.483	0.997	0.056
25	0	2	0.391	0.110	1.333	1.790	1.000	0.049
25	0	4	0.911	0.010	1.332	1.774	1.000	0.034
25	1	0	0.033	0.307	1.140	1.393	0.975	0.499
25	1	2	0.560	0.209	1.301	1.736	0.991	0.207
25	1	4	0.947	0.163	1.473	2.196	0.981	0.057
25	2	0	0.332	0.332	1.499	2.358	0.933	0.969
25	2	2	0.845	0.026	1.558	2.429	0.903	0.555
25	2	4	0.993	0.538	1.633	2.955	1.000	0.286
50	0	0	0.000	0.035	0.417	0.175	0.960	0.065
50	0	2	0.073	0.011	0.732	0.535	0.994	0.067
50	0	4	0.583	0.018	1.375	1.890	0.998	0.058
50	1	0	0.002	0.167	0.713	0.536	0.974	0.790
50	1	2	0.171	0.337	1.134	1.401	0.979	0.520
50	1	4	0.737	0.315	1.270	1.712	0.989	0.304
50	2	0	0.118	0.479	1.345	2.037	0.957	1.000
50	2	2	0.573	0.324	1.290	1.768	0.934	0.948
50	2	4	0.953	-0.084	1.228	1.514	0.936	0.638
75	0	0	0.000	0.005	0.309	0.095	0.963	0.051
75	0	2	0.006	-0.018	0.519	0.270	0.969	0.064
75	0	4	0.313	0.055	0.961	0.927	0.993	0.071
75	1	0	0.000	0.093	0.421	0.186	0.968	0.937
75	1	2	0.052	0.309	0.829	0.783	0.981	0.726
75	1	4	0.538	0.515	1.465	2.413	0.976	0.485
75	2	0	0.038	0.353	1.069	1.268	0.972	1.000
75	2	2	0.416	0.542	1.437	2.358	0.959	0.988
75	2	4	0.869	0.277	1.370	1.953	0.939	0.878

n, sample size; β_1, simulated regression effect of interest; p, number of nuisance covariates; p.fail, proportion of simulated data sets for which the method failed; bias, estimated bias of estimator; SD, sample standard deviation of estimates; MSE, estimated mean-squared error of estimator; coverage, estimated coverage of 95% confidence intervals (CIs); power, estimated power to test $H_0 : \beta_1 = 0$ *vs.* $H_1 : \beta_1 \neq 0$. Results are based on 1000 simulated datasets.

- The failure rate of CLR was substantial in most of our simulation configurations, suggesting it is not useful for the sample sizes we considered.

3.3.5 Software implementations

The implementations of the methods are as follows.

- LR is implemented in the `glm()` function from the R package `stats`.
- FLR is implemented in the `logistf()` function f the logistf R package.
- ZLR is our own implementation. We were able to implement the estimator and standard errors generally, but the penalized profile likelihood-based confidence intervals and tests only for the case of one covariate (i.e., no nuisance covariates, or $p = 0$).
- log-$F(m, m)$ was implemented in R by data augmentation, as described by Greenland and Mansournia (2015).
- ELR is implemented in LogXact. The point estimate returned by LogXact is the CMLE if it exists and the MUE if not.
- CLR is implemented in the `clogit()` function of the R package survival.
- FCLR is a function we wrote ourselves based on instructions from G. Heinze (personal communication), based on the `coxphf()` function for Firth-penalized Cox regression.
- ECLR is implemented in LogXact.
- FLR-S uses the `logistf()` function from the logistf R package.
- HOA was our own implementation. We were able to implement the estimator and standard errors generally, but the penalized profile likelihood-based confidence intervals and tests only for the case of one covariate (i.e., no nuisance covariates, or $p = 0$).

All R code is available online at the website of the handbook:
`www.mn.uio.no/math/english/research/groups/statistics-biostatistics/handbook-of-case-control-studies/`.

References

Brazzale, A. R. and Davison, A. C. (2008). Accurate parametric inference for small samples. *Statistical Science*, **23**, 465–484.

Graham, J., McNeney, B., and Platt, R. W. (2018). Small sample methods. In Ø. Borgan, N. Breslow, N. Chatterjee, M. H. Gail, A. Scott, and C. J. Wild, editors, *Handbook of Statistical Methods for Case-Control Studies*, chapter 8, pages 133–162. Chapman and Hall/CRC Press, Boca Raton, Florida.

Greenland, S. and Mansournia, M. A. (2015). Penalization, bias reduction, and default priors in logistic and related categorical and survival regressions. *Statistics in Medicine*, **34**, 3133–3143.

Heinze, G. and Puhr, R. (2010). Bias-reduced and separation-proof conditional logistic regression with small or sparse data sets. *Statistics in Medicine*, **29**, 770–777.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, **21**, 2409–2419.

McCullagh, P. (1984). On the elimination of nuisance parameters in the proportional odds model. *Journal of the Royal Statistical Society Series B*, **46**, 250–256.

Zamar, D., McNeney, B., and Graham, J. (2007). elrm: Software implementing exact-like inference for logistic regression models. *Journal of Statistical Software*, **21**(3).