# A bias-corrected parameter estimation approach in Logistic Errors-in-variable Regression

Pei Geng and Huyen Nguyen

*Department of Mathematical and Statistics, University of New Hampshire, Durham, NH 03824*
*Department of Statistics, University of Connecticut, Storrs, CT 06269*

## Motivation

- Measurement error is widely observed during the data collection process such as daily calories intake, self-reported survey data, exposure dose to radiation.
- In Logistic regression with binary responses, ignoring measurement error in covariates causes parameter estimation bias (Stefanski and Carroll, 1985).
- In case-control studies, when the covariate is error free, Geng and Sakhanenko (2016) developed an integrated square distance (ISD) estimation approach which shows superior performance in severely imbalanced cases.
- When measurement error is present in covariate, we aim to investigate the bias in the naive ISD estimation and propose a bias-corrected estimator using the deconvolution kernel density estimation.

## Model & Data Structure

- Logistic regression:

$$P(Y=1|X=x)=\frac{exp(\alpha^*+\beta x)}{1+exp(\alpha^*+\beta x)}$$

- Measurement error model:

$$Z=X+U$$

- Let $f_0(x)$ and $f_1(x)$ be the covariate density of the control and case groups. In the case control framework:

$$\mathbf{ln}\left\{\frac{\boldsymbol{f_1(x)}}{\boldsymbol{f_0(x)}}\right\}=\boldsymbol{\alpha}+\boldsymbol{\beta x}$$

- Data structure:
  Case group ($Y=1$): $\{z_1^1, z_2^1, \ldots, z_{n_1}^1\}$
  Control group ($Y=0$): $\{z_1^0, z_2^0, \ldots, z_{n_0}^0\}$

## Methodology

- Integrated Square Distance:

$$T_n(s,t)=\int_a^b\left\{\ln\left\{\frac{\hat{f}_1(x)}{\hat{f}_0(x)}\right\}-s-tx\right\}^2dx$$

- $\hat{f}_i(x)$ are the deconvolution kernel density estimators

$$\hat{f}_i(x)=(n_ih_i)^{-1}\sum_{j=1}^{n_i}K_{h_i}^*\left(\frac{z_j^i-x}{h_i}\right)$$

$$K_h^*(y)=\frac{1}{2\pi h}\int\left(\exp(-ity)\frac{\phi_K(t)}{\phi_U(t/h)}\right)dt$$

- The bias-corrected estimator is defined as

$$(\hat{\alpha},\hat{\beta})=\underset{s,t}{argmin}\,T_n(s,t)$$

## Bias corrected ISD estimators

$$\hat{\beta}=\frac{12}{(b-a)^3}\int_a^b ln\left\{\frac{\hat{f}_1(x)}{\hat{f}_0(x)}\right\}\left(x-\frac{a+b}{2}\right)dx;\quad \hat{\alpha}=\frac{1}{b-a}\int_a^b\left[ln\left\{\frac{\hat{f}_1(x)}{\hat{f}_0(x)}\right\}-\hat{\beta}x\right]dx$$

Theorem 1. $\left(0<\rho=\lim\left(\frac{n_1}{n_1+n_0}\right)<1\right)$ For ordinary smooth measurement errors in covariate ($\phi_U(t)=O(t^{-\tau})$), with proper choices of the Kernel density $K$ and bandwidths $h_0, h_1$, we have $(\hat{\alpha},\hat{\beta})$ are consistent estimators of $(\alpha,\beta)$. Moreover,

$$\sqrt{n_0+n_1}\begin{pmatrix}\hat{\alpha}-\alpha\\ \hat{\beta}-\beta\end{pmatrix}\rightarrow_D N(0,\Sigma),\Sigma=\rho^{-1}\Sigma_0+(1-\rho)^{-1}\Sigma_1.$$

Theorem 2. ($\rho=0$ *or* $1$) $\sqrt{n_1}\begin{pmatrix}\hat{\alpha}-\alpha\\ \hat{\beta}-\beta\end{pmatrix}\rightarrow_D N(0,\Sigma_1)$ for $\rho=0$; $\sqrt{n_0}\begin{pmatrix}\hat{\alpha}-\alpha\\ \hat{\beta}-\beta\end{pmatrix}\rightarrow_D N(0,\Sigma_0)$ for $\rho=1$.

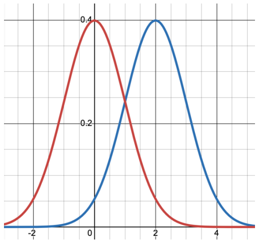
## Simulation Study

Covariate $X$: Gaussian or Laplace
Measurement error U: Double exponential

Estimation comparison:

$\hat{\boldsymbol{\beta}}$: the proposed bias-corrected ISD estimator
$\hat{\boldsymbol{\beta}}_{\boldsymbol{ISD}}$: the naïve ISD estimator
$\hat{\boldsymbol{\beta}}_{\boldsymbol{MLE}}$: the naïve maximum likelihood estimator
$\hat{\boldsymbol{\beta}}_{\boldsymbol{BC}}$: the bias-corrected est. by Stefanski & Carroll (1985)

Gaussian covariates



Laplace covariates

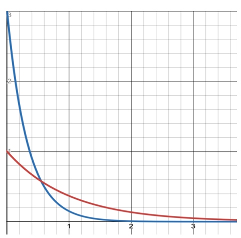


Table 1: Bias and RMSE comparison of estimators for imbalanced sample sizes with Gaussian covariates when $\sigma_U^2=0.5^2$ and $\sigma_U^2=1$.

| Estimator | | $n_0=50$<br>$n_1=150$ | $n_0=50$<br>$n_1=500$ | $n_0=100$<br>$n_1=500$ | $n_0=100$<br>$n_1=1000$ | $n_0=500$<br>$n_1=100$ | $n_0=1000$<br>$n_1=100$ | $n_0=50$<br>$n_1=150$ | $n_0=50$<br>$n_1=500$ | $n_0=100$<br>$n_1=500$ | $n_0=100$<br>$n_1=1000$ | $n_0=500$<br>$n_1=100$ | $n_0=1000$<br>$n_1=100$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{\beta}$ | \|Bias\| | 0.1011 | 0.1375 | 0.0770 | 0.0823 | 0.0958 | 0.0862 | 0.1947 | 0.1316 | 0.1065 | 0.0853 | 0.0986 | 0.1282 |
| | RMSE | 0.5897 | 0.5500 | 0.4891 | 0.4434 | 0.4271 | 0.4799 | 0.8299 | 0.8309 | 0.7714 | 0.7575 | 0.7935 | 0.7414 |
| $\hat{\beta}_{ISD}$ | \|Bias\| | 0.3201 | 0.3599 | 0.3457 | 0.3484 | 0.3576 | 0.3638 | 0.7619 | 0.7670 | 0.7979 | 0.7899 | 0.7951 | 0.7964 |
| | RMSE | 0.5379 | 0.5307 | 0.4578 | 0.4458 | 0.4455 | 0.4618 | 0.8576 | 0.8575 | 0.8458 | 0.8332 | 0.8369 | 0.8368 |
| $\hat{\beta}_{MLE}$ | \|Bias\| | 0.3398 | 0.3871 | 0.3827 | 0.3881 | 0.3873 | 0.4014 | 0.9514 | 1.0065 | 0.9881 | 1.0103 | 0.9913 | 1.0183 |
| | RMSE | 0.4355 | 0.4408 | 0.4151 | 0.4123 | 0.4207 | 0.4215 | 0.9735 | 1.0171 | 0.9946 | 1.015 | 0.9979 | 1.0227 |
| $\hat{\beta}_{BC}$ | \|Bias\| | 0.1234 | 0.2569 | 0.2227 | 0.2606 | 0.0642 | 0.0796 | 0.5988 | 0.7602 | 0.7068 | 0.7695 | 0.3415 | 0.3438 |
| | RMSE | 0.3875 | 0.3570 | 0.2982 | 0.3072 | 0.2589 | 0.2145 | 0.6941 | 0.7882 | 0.7269 | 0.7810 | 0.4181 | 0.3984 |

Table 2: Bias and RMSE comparison of estimators for imbalanced sample sizes with Laplace covariates when $\sigma_U^2=0.1^2$ and $\sigma_U^2=0.2^2$

| Estimator | | $n_0=100$<br>$n_1=500$ | $n_0=100$<br>$n_1=1000$ | $n_0=200$<br>$n_1=1000$ | $n_0=1000$<br>$n_1=200$ | $n_0=1000$<br>$n_1=100$ | $n_0=500$<br>$n_1=100$ | $n_0=100$<br>$n_1=500$ | $n_0=100$<br>$n_1=1000$ | $n_0=200$<br>$n_1=1000$ | $n_0=1000$<br>$n_1=200$ | $n_0=1000$<br>$n_1=100$ | $n_0=500$<br>$n_1=100$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{\beta}$ | \|Bias\| | 0.0633 | 0.0155 | 0.0046 | 0.0030 | 0.0150 | 0.0441 | 0.0597 | 0.1013 | 0.0278 | 0.0433 | 0.0836 | 0.0801 |
| | RMSE | 0.8707 | 0.8493 | 0.6688 | 0.5213 | 0.6663 | 0.7321 | 0.9167 | 0.9546 | 0.7709 | 0.6153 | 0.6939 | 0.8012 |
| $\hat{\beta}_{ISD}$ | \|Bias\| | 0.3089 | 0.3573 | 0.1711 | 0.0034 | 0.0447 | 0.0514 | 0.5171 | 0.5619 | 0.3945 | 0.1594 | 0.0870 | 0.1279 |
| | RMSE | 0.5626 | 0.5715 | 0.4541 | 0.3975 | 0.4540 | 0.4955 | 0.6998 | 0.7143 | 0.5654 | 0.4332 | 0.4534 | 0.5003 |
| $\hat{\beta}_{MLE}$ | \|Bias\| | 0.0955 | 0.1258 | 0.1114 | 0.0530 | 0.0388 | 0.0416 | 0.3850 | 0.4276 | 0.3989 | 0.2221 | 0.1918 | 0.2193 |
| | RMSE | 0.3115 | 0.3213 | 0.2337 | 0.1799 | 0.2284 | 0.2436 | 0.4537 | 0.4855 | 0.4336 | 0.2743 | 0.2869 | 0.3124 |
| $\hat{\beta}_{BC}$ | \|Bias\| | 0.0507 | 0.0864 | 0.0682 | 0.0500 | 0.0766 | 0.0634 | 0.2635 | 0.3221 | 0.2808 | 0.0946 | 0.1666 | 0.1006 |
| | RMSE | 0.3157 | 0.3208 | 0.2260 | 0.1942 | 0.2587 | 0.2701 | 0.3825 | 0.4146 | 0.3422 | 0.2293 | 0.3267 | 0.3071 |

## Results

- The proposed bias-corrected ISD estimator shows the smallest bias for both small and large measurement errors in most of the imbalanced cases.
- The bias corrected estimator by Stefanski & Carroll (1985) works fairly comparable with the proposed estimator when the control sample size is much larger than the case sample size such as $n_0=500, n_1=100$.
- The MSE of the proposed estimation is larger compared to the bias corrected estimator by Stefanski & Carroll (1985) due to the slow convergence rate of the Deconvolution kernel density estimator.

## Real Data Application

- We applied our method to the Framingham Heart Study to explore the relation between the systolic blood pressure (two repeated measurements $Z_1$ and $Z_2$) and consequences of cardiovascular disease.
- There are 128 individuals with cardiovascular disease and 1487 individuals without the disease. To fit the case-control framework, we generated a nested case-control dataset by matching 5 controls for each case with cardiovascular disease according to their age and smoking status.
- The four parameter estimators are shown in the table below. The proposed estimator seems to capture higher effect size of the blood pressure.
- In the proposed estimator, the measurement error was assumed to follow the double exponential distribution with estimated variance $\hat{\sigma}_U^2=5.54^2$.

| | $Z_1$ | $Z_2$ |
|---|---|---|
| $\hat{\beta}$ | 0.022709 | 0.023068 |
| $\hat{\beta}_{ISD}$ | 0.025736 | 0.027160 |
| $\hat{\beta}_{MLE}$ | 0.014023 | 0.012568 |
| $\hat{\beta}_{BC}$ | 0.014032 | 0.012578 |

## Discussion

- Although the estimation approach was proposed for the ordinary smooth errors, the method can be applied to super smooth errors such as Gaussian error, however, the theoretical results in Theorems 1-2 need to be re-established.
- Because of superior bias reduction for imbalanced case control cases of the ISD method, it is desirable to further generalize the approach to weighted ISD for more flexibility and possibly reduced MSE.

## References

Geng, P., & Nguyen, H. (2024). Parameter estimation for Logistic errors-in-variables regression under case–control studies. *Statistical Methods & Applications*, 33(2), 661-684.

Geng, P., and Sakhanenko, L. (2016). Parameter estimation for the logistic regression model under case-control study. *Statistics & Probability Letters*, 109, 168-177.

Stefanski, L. A., & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *The annals of statistics*, 1335-1351.

Stefanski, L. A. and Carroll, R. J. (1990). Deconvolving kernel density estimators. *Statistics*, 21(2), 169-184.