Constrained Maximum Likelihood Estimation for Model Calibration Using Summary-level Information from External Big Data Sources

Yi-Hau Chen
Institute of Statistical Science, Academia Sinica

Joint with Nilanjan Chatterjee, Paige Maas, Raymond Carroll

National University of Singapore
June 2016
Background

- increasingly, there are very **large public or private data sources** that provide **summary** or **crude (error-prone)** information, although individual data in such external data may not be accessible.

- to analyze data from specific **internal studies that collect more detailed and precise data** while utilizing **crude or summary information from external big data sources**
A Motivation Example

- A study based on the **Health Interview Survey** data is to examine the relationship of **Herpes Zoster (HZ)** with **chronic obstructive pulmonary disease (COPD)**, adjusting for comorbidity (hypertension, diabetes, coronary artery disease, cancer, ...), smoking and drinking.

- Data on all the variables except for smoking and drinking are available in the large **Health Insurance Database**.

- The external data provide information on the **reduced model** for the relationship of HZ with COPD adjusting for comorbidity but not for smoking and drinking.
such information on the reduced model from the large Health Insurance Database may be utilized and incorporated to the analysis of the internal study based on the smaller Health Interview Survey data
Models: External Data

- $Y, X$: outcome and (crude, error-prone) covariates; we may not have individual data on them, but summary information is available.

- $g_\theta(y|x)$: model that has been built based on external data; may be mis-specified.

- $\theta$: parameters in the external reduced model, whose estimates $\hat{\theta}$ are available.
Models: Internal Data

- **$Y, X$:** outcome and (crude, error-prone) covariates

- **$Z$:** more accurate covariate information available in internal study

- Individual data on $Y, X, Z$ are available

- $f_\beta(y|x, z)$: model for internal data, assumed to be correctly specified

- $\beta$: parameters in the internal model; aims of inference
Relationship Between Internal and External Models

- $\hat{\theta}, U(Y|X, \theta)$: the external estimate and the estimating function $U(Y|X, \hat{\theta}) = 0$

- the limiting value $\theta^*$ of $\hat{\theta}$ satisfies

$$E\{U(Y|X, \theta^*)\} = \int U(y|x, \theta^*)pr(y|x)pr(x)dydx = 0$$

or

$$\int_x \int_z \left\{ \int_y U(y|x, \theta^*)f_{\beta_0}(y|x, z)dy \right\} dF(x, z) = 0$$

$\beta_0$ is true value of $\beta$
\[ \int u_{\beta_0}(x, z; \theta^*) dF(x, z) = 0 \] where

\[ u_\beta(X, Z; \theta^*) = \int_y U(y|X, \theta^*) f_\beta(y|X, Z) dy \]
semiparametric constrained maximum likelihood

- likelihood based on internal data \((Y_i, X_i, Z_i)\) for \(i = 1, \ldots, N:\)

\[
L_{\beta,F} = \prod_{i=1}^{N} f_{\beta}(Y_i|X_i, Z_i) dF(X_i, Z_i)
\]

- semiparametric constrained likelihood:

\[
l_{\beta,\lambda} = \log L_{\beta,F} + \lambda^T \int u_{\beta}(X, Z; \theta) dF(X, Z)
\]

- \(\lambda\): Lagrange multipliers
- \(\theta\) is fixed at \(\theta = \theta^* \approx \hat{\theta}\) when external data is very large
- \(F(X, Z)\) is common and treated nonparametrically
Empirical (Profile) Likelihood

• \((\delta_j)_{j=1}^m\): masses of \(F(X, Z)\) at \(m\) unique values in \((X_i, Z_i)_{i=1}^N\)

• by Lagrange multipliers, we maximize over \((\beta, \lambda, \gamma, \delta_1, \ldots, \delta_m)\)

\[
\sum_{i=1}^{N} \log f_\beta(Y_i | X_i, Z_i) + \sum_{j=1}^{m} n_j \log \delta_j + \lambda^T \sum_{j=1}^{m} u_\beta(X_j, Z_j; \theta) \delta_j + \gamma \left( \sum_{j=1}^{m} \delta_j - 1 \right)
\]

- loglikelihood of internal data
- constraint from external data
- constraint for \(F\)

• profiling out \(\delta_1, \ldots, \delta_m\) first leads to the log pseudo-likelihood:

\[
l_{\beta, \lambda}^* = \sum_{i=1}^{N} \log \left\{ \frac{f_\beta(Y_i | X_i, Z_i)}{1 - \lambda^T u_\beta(X_i, Z_i; \theta)} \right\}
\]
The Proposed Estimator for $\beta$

- let $\eta = (\beta^T, \lambda^T)^T$

- $\hat{\eta} = (\hat{\beta}^T, \hat{\lambda}^T)^T$ is the solution to $\partial l^*_{\beta,\lambda}/\partial \eta = 0$

- the proposed constrained maximum likelihood (CML) estimator
**Computation**

- $\hat{\eta}$ is obtained by solving for the **stationary point**, indeed the saddle point, over the expanded parameter $\eta = (\beta^T, \lambda^T)^T$ for the **log pseudo-likelihood function**

- the conventional **Newton-Raphson method** works well when initial value of $\lambda$ is set to zero

- it is easy to calculate the score and the Hessian, and then do the maximization
the constrained maximum likelihood can be derived in the same manner under a variety of sampling designs for the internal study, including simple random, case-control and stratified case-control sampling designs.

owing to the use of external information, parameters unidentifiable in the internal sample under a biased sampling design, such as the intercept parameter of logistic regression model in the case-control sample, can still be identifiable in the constrained maximum likelihood analysis.
Case-Control Design in Internal Study

- $Y$ is binary

- $N_1$ and $N_0$ the numbers of cases and controls sampled in internal study

- $p_1 = 1 - p_0 = \int f_\beta(Y = 1|x, z) dF(x, z)$ the marginal disease probability for a given value of $\beta$
Constrained Likelihood under Case-Control Design

- the likelihood for the internal case-control sample:

\[
L_{\beta,F}^{cc} = \left\{ \prod_{i=1}^{N_1+N_0} f_\beta(Y_i|X_i,Z_i) dF(X_i,Z_i) \right\} \times p_1^{-N_1} p_0^{-N_0}
\]

- constrained likelihood:

\[
l^{cc}_\lambda = \log \left( L_{\beta,F}^{cc} \right) + \lambda^T \int u_\beta(X,Z;\theta) dF(X,Z)
\]

- profiling out the masses of \( F(X,Z) \) leads to the log pseudo-likelihood

\[
l_{\beta,\lambda,\mu_1}^{*,cc} = \sum_{i=1}^{N} \log \left\{ \frac{f_\beta(Y_i|X_i,Z_i)}{\sum_y f_\beta(y|X_i,Z_i) \mu_y - \lambda^T u_\beta(X_i,Z_i;\theta)} \right\} + \sum_y N_y \log \mu_y
\]
\[ \mu_1 = \frac{N_1}{p_1}, \quad \mu_0 = \frac{N_0}{p_0} \]
Asymptotic Theory (Qin and Lawless (1994 Ann Stat))

- \( \hat{\eta} = (\hat{\beta}, \hat{\lambda}) \longrightarrow_p \eta_0 = (\beta_0^T, 0)^T \)
  - \( \beta_0 \): true value of \( \beta \); \( 0 \): zero vector with same dimension as \( \lambda \)

- as \( N \rightarrow \infty \), \( N^{1/2}(\hat{\eta} - \eta_0) \sim \mathcal{N}(0, \Omega) \)

\[
\Omega = \begin{bmatrix}
(B + CL^{-1}C^T)^{-1} & O \\
O & (L + C^T B^{-1} C)^{-1}
\end{bmatrix}
\]

\[ B = E \left\{ -\frac{\partial^2 \log f_{\beta}(Y|X,Z)}{\partial \beta \partial \beta^T} \right\} \]
\[ L = E \left\{ u_{\beta}(X,Z) u_{\beta}^T(X,Z) \right\} \]

\[ C = E \left\{ \int_y \frac{\partial \log f_{\beta}(y|X,Z)}{\partial \beta} U^T(y|X,\theta)f_{\beta}(y|X,Z)dy \right\} \]
Asymptotic Properties

- The CML estimator $\hat{\beta}$ is asymptotically more efficient than that based on the internal data only

  $$\text{var} \, \hat{\beta} = (B + C L^{-1} C^T)^{-1} \preceq B^{-1} = \text{var} \, \hat{\beta}^I$$

- $\hat{\beta}$ is asymptotically independent of $\hat{\lambda}$

- Asymptotic variance $\Omega$ can be consistently estimated by substituting the corresponding sample means for the expected quantities in the expression
Simulations: Missing Covariate

- binary $Y$ and full covariate $(X, Z)$ available in **internal study**
- $(Y, X)$ available in **external study**

**internal study model:**

$$\logit P(Y = 1|X, Z) = \beta_0 + X\beta_X + Z\beta_Z + XZ\beta_{XZ}$$

**external study:** information on the **reduced model**

$$\logit P(Y = 1|X) = \theta_0 + X\theta_X$$
Simulation Setups: Missing Covariate

- \((X, Z)\) bivariate standard normal with correlation 0.3

- \(Y\):

\[
\text{logit } P(Y = 1|X, Z) = \beta_0 + X\beta_X + Z\beta_Z + XZ\beta_{XZ}
\]

relative risks for the main effects \(\sim 1.50\) and for the interaction \(\sim 1.25\), population disease prevalence \(\sim 20\%\)

- internal sample size \(N = 1000\) (in case-control sample, 500 cases and 500 controls)
Comparisons with Alternative Methods

- **internal data-only estimate**

  \[ \hat{\beta}^I : \text{the solution to } 0 = \sum_{i=1}^{N} \frac{\partial}{\partial \beta} \log f_\beta(Y_i|X_i, Z_i) \]

- not using external information

- consistent but **losing efficiency when external information is available**
• **generalized regression (GR)** (Chen and Chen, 2000 JRSSB)

\[
\hat{\beta}^{GR} = \hat{\beta}^I + H_1^{-1}C_{12}C_{22}^{-1}H_2(\hat{\theta} - \hat{\theta}^I)
\]

\[
H_1 = E\left\{ \frac{\partial^2 \log f_\beta(Y|X, Z)}{\partial \beta \partial \beta^T} \right\}, \quad H_2 = E_I\left\{ \frac{\partial}{\partial \theta^T} U(Y|X, \theta) \right\}
\]

\[
C_{22} = E_I\left\{ U(Y|X, \theta)U^T(Y|X, \theta) \right\}, \quad C_{12} = E_I\left\{ \frac{\partial}{\partial \beta} \log f_\beta(Y|X, Z)U^T(Y|X, \theta) \right\}
\]

• originally developed for internal study under simple random sampling

• ad-hoc modifications required for general sampling designs
### Results (multiplied by $10^3$; coverage probability (CP) reported by %)

<table>
<thead>
<tr>
<th></th>
<th>Int</th>
<th>$\beta_0$</th>
<th>CML</th>
<th>Int</th>
<th>$\beta_X$</th>
<th>CML</th>
<th>Int</th>
<th>$\beta_Z$</th>
<th>CML</th>
<th>Int</th>
<th>$\beta_{XZ}$</th>
<th>CML</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>simple random; $N = 1000$</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-8.94</td>
<td>2.67</td>
<td><strong>2.84</strong></td>
<td>2.42</td>
<td>3.30</td>
<td><strong>3.37</strong></td>
<td>1.29</td>
<td>1.50</td>
<td><strong>0.95</strong></td>
<td>1.33</td>
<td>1.27</td>
<td><strong>2.42</strong></td>
</tr>
<tr>
<td>SE</td>
<td>91.4</td>
<td>32.5</td>
<td><strong>32.4</strong></td>
<td>96.8</td>
<td>39.0</td>
<td><strong>38.9</strong></td>
<td>94.3</td>
<td>94.4</td>
<td><strong>94.3</strong></td>
<td>89.4</td>
<td>89.4</td>
<td><strong>89.5</strong></td>
</tr>
<tr>
<td>ESE</td>
<td>91.8</td>
<td>32.1</td>
<td><strong>32.3</strong></td>
<td>92.3</td>
<td>38.8</td>
<td><strong>38.9</strong></td>
<td>92.4</td>
<td>92.3</td>
<td><strong>92.5</strong></td>
<td>85.8</td>
<td>85.6</td>
<td><strong>86.9</strong></td>
</tr>
<tr>
<td>MSE</td>
<td>8.42</td>
<td>1.06</td>
<td><strong>1.06</strong></td>
<td>9.38</td>
<td>1.53</td>
<td><strong>1.53</strong></td>
<td>8.89</td>
<td>8.91</td>
<td><strong>8.89</strong></td>
<td>7.98</td>
<td>7.99</td>
<td><strong>8.01</strong></td>
</tr>
<tr>
<td>CP</td>
<td>95.4</td>
<td>94.7</td>
<td><strong>95.3</strong></td>
<td>94.3</td>
<td>93.4</td>
<td><strong>94.0</strong></td>
<td>94.6</td>
<td>94.5</td>
<td><strong>95.1</strong></td>
<td>93.6</td>
<td>93.7</td>
<td><strong>93.8</strong></td>
</tr>
<tr>
<td><strong>case-control; $N = 1000$</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-</td>
<td>-</td>
<td><strong>2.59</strong></td>
<td>2.40</td>
<td>14.8</td>
<td><strong>0.88</strong></td>
<td>5.06</td>
<td>5.01</td>
<td><strong>5.11</strong></td>
<td>-1.51</td>
<td>-1.53</td>
<td><strong>-1.57</strong></td>
</tr>
<tr>
<td>SE</td>
<td>-</td>
<td>-</td>
<td><strong>22.7</strong></td>
<td>75.7</td>
<td>25.1</td>
<td><strong>26.8</strong></td>
<td>72.2</td>
<td>72.3</td>
<td><strong>72.2</strong></td>
<td>72.9</td>
<td>72.9</td>
<td><strong>72.8</strong></td>
</tr>
<tr>
<td>ESE</td>
<td>-</td>
<td>-</td>
<td><strong>22.8</strong></td>
<td>73.3</td>
<td>26.1</td>
<td><strong>27.9</strong></td>
<td>73.1</td>
<td>73.2</td>
<td><strong>73.2</strong></td>
<td>71.4</td>
<td>71.4</td>
<td><strong>71.6</strong></td>
</tr>
<tr>
<td>MSE</td>
<td>-</td>
<td>-</td>
<td><strong>0.52</strong></td>
<td>5.73</td>
<td>0.85</td>
<td><strong>0.72</strong></td>
<td>5.24</td>
<td>5.24</td>
<td><strong>5.24</strong></td>
<td>5.31</td>
<td>5.31</td>
<td><strong>5.30</strong></td>
</tr>
<tr>
<td>CP</td>
<td>-</td>
<td>-</td>
<td><strong>94.7</strong></td>
<td>94.2</td>
<td>91.3</td>
<td><strong>96.2</strong></td>
<td>95.4</td>
<td>95.6</td>
<td><strong>95.4</strong></td>
<td>94.7</td>
<td>94.4</td>
<td><strong>94.5</strong></td>
</tr>
</tbody>
</table>

SE: standard error; ESE, estimated standard error
MSE: mean squared error
Simulations: Mismeasured Covariate

- binary $Y$, crude covariate $X$ and accurate covariate $Z$ collected in internal study

- only $Y$ and $X$ are observed in external study

- internal study model:

  $$\text{logit } P(Y = 1|X, Z) = \beta_0 + Z\beta_Z$$

  $Y$ is independent of $X$ given $Z$

  (non-differential measurement error)
• external study: information on the reduced model

\[ \text{logit } P(Y = 1|X) = \theta_0 + X\theta_X \]
Simulation Setups: Mismeasured Covariate

- \((X, Z)\) bivariate standard normal with correlation 0.3

- \(Y\): \[\text{logit } P(Y = 1|Z) = \beta_0 + Z\beta_Z\]

  relative risk for the main effect \(\sim 1.50\), population disease prevalence \(\sim 20\%\)

- internal sample size \(N = 1000\) (in case-control sample, 500 cases and 500 controls)
### Results (multiplied by $10^3$; coverage probability (CP) reported by %)

<table>
<thead>
<tr>
<th></th>
<th>$\beta_0$</th>
<th></th>
<th></th>
<th>$\beta_Z$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Int</td>
<td>GR</td>
<td>CML</td>
<td>Int</td>
<td>GR</td>
<td>CML</td>
</tr>
<tr>
<td><strong>simple random; $N = 1000$</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-2.12</td>
<td>-3.73</td>
<td>0.20</td>
<td>0.80</td>
<td>1.23</td>
<td>1.13</td>
</tr>
<tr>
<td>SE</td>
<td>87.7</td>
<td>25.1</td>
<td>15.1</td>
<td>89.6</td>
<td>84.7</td>
<td>40.1</td>
</tr>
<tr>
<td>ESE</td>
<td>87.1</td>
<td>23.9</td>
<td>15.2</td>
<td>86.3</td>
<td>82.5</td>
<td>38.7</td>
</tr>
<tr>
<td>MSE</td>
<td>7.69</td>
<td>0.64</td>
<td>0.23</td>
<td>8.02</td>
<td>7.17</td>
<td>1.61</td>
</tr>
<tr>
<td>CP</td>
<td>95.9</td>
<td>92.6</td>
<td>94.1</td>
<td>94.2</td>
<td>94.0</td>
<td>94.1</td>
</tr>
<tr>
<td><strong>case-control; $N = 1000$</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bias</td>
<td>-</td>
<td>-</td>
<td>0.99</td>
<td>2.85</td>
<td>2.91</td>
<td>1.74</td>
</tr>
<tr>
<td>SE</td>
<td>-</td>
<td>-</td>
<td>12.8</td>
<td>66.0</td>
<td>62.5</td>
<td>37.6</td>
</tr>
<tr>
<td>ESE</td>
<td>-</td>
<td>-</td>
<td>12.9</td>
<td>66.6</td>
<td>63.8</td>
<td>36.3</td>
</tr>
<tr>
<td>MSE</td>
<td>-</td>
<td>-</td>
<td>0.16</td>
<td>4.36</td>
<td>3.63</td>
<td>1.42</td>
</tr>
<tr>
<td>CP</td>
<td>-</td>
<td>-</td>
<td>95.6</td>
<td>95.7</td>
<td>96.1</td>
<td>94.6</td>
</tr>
</tbody>
</table>

SE: standard error; ESE: estimated standard error
MSE: mean squared error
Analysis of Relationship Between HZ and COPD Based on the LHID and HIS databases

- **Internal Data**: Health Interview Survey 2005 (HIS) by Health Research Institute and Bureau of Health Promotion in Taiwan, with data on medical claims, health behaviors, and quality of life for 26,658 Taiwan residents in 2005

- the internal sample consists of **244 COPD patients** (diagnosed before January 2004) and **904 age- and gender-matched non-COPD subjects** from the HIS, all of them had no diagnosis of **Herpes Zoster (HZ)** before 2004
• outcome $Y$ is the development of **Herpes Zoster (HZ)** by December 31 2006

• covariate data $(X, Z)$ are **COPD status, comorbidity (diabetes mellitus, hypertension, coronary artery disease, chronic liver disease, autoimmune disease, and cancer)** and **cumulative smoking and alcohol consumptions**
- **external data source**: the Longitudinal Health Insurance Database 2005 (LHID), containing all the medical claims data for one million beneficiaries, randomly sampled from 25.68 million enrollees in Taiwan

- **8,486 COPD patients** diagnosed before January 1 2004 and **33,944 age- and gender-matched non-COPD subjects** randomly selected from LHID, all of them had no diagnosis of HZ before 2004
• outcome $Y$ is the development of Herpes Zoster (HZ) by December 31 2006

• covariate variables include all those collected in the internal sample except for cumulative smoking and alcohol consumptions
Analysis Models

- the internal data analysis employs the **logistic regression model** for the development of HZ with covariates **COPD**, comorbidity, cumulative smoking and alcohol consumptions (ordinal data)

- the external data analysis is based on the **reduced logistic regression model** for the development of HZ with covariates **COPD** and comorbidity but **without cumulative smoking and alcohol consumptions**
coefficients of the external reduced logistic regression model obtained from LHID are used for the constrained maximum likelihood analysis combining the internal HIS data with the external LHID data
## Association Between HZ and COPD

<table>
<thead>
<tr>
<th>Model/Method</th>
<th>Estimate (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjusting Comorbidity (external data)</td>
<td>0.530</td>
</tr>
<tr>
<td>Adjusting Comorbidity, Smoking &amp; Drinking (internal data)</td>
<td>1.041 (0.552)</td>
</tr>
<tr>
<td>Generalized Regression</td>
<td>0.641 (0.059)*</td>
</tr>
<tr>
<td>Constrained Maximum Likelihood</td>
<td>0.620 (0.055)*</td>
</tr>
</tbody>
</table>

*: p value < 0.05
Conclusions

- we have proposed the constrained maximum likelihood (CML) to exploit summary-level information from a big external data in usual analysis for an internal study sample.

- the method is semiparametric in nature, assuming a common parametric model for the conditional distribution of the outcome given the covariates, and a common covariate distribution in both the internal and external populations, but without imposing parametric assumptions for the common covariate distribution.
• applicable to various sampling designs

• it is easy to modify the CML estimator by δ method to account for uncertainty about the external information \( \hat{\theta} \) when it cannot be ignored

• when the covariate distributions between the internal and external populations are different, we propose using a reference sample which is representative for the external population to estimate the external covariate distribution

• then applying a variant of the CML method based on the estimated external covariate distribution
Synthetic Constrained Maximum Likelihood

- \{ (X_j^*, Z_j^*), j = 1, \ldots, N_r \}: reference sample

- \delta_i \ (i = 1, \ldots, N): masses of internal covariate distribution \( F(X, Z) \)

- \delta_j^* \ (j = 1, \ldots, N_r): masses of external covariate distribution \( F^*(X^*, Z^*) \)
• synthetic constrained likelihood (Han and Lawless 2016 JASA):

\[
\sum_{i=1}^{N} \log f_{\beta}(Y_i|X_i, Z_i) + \sum_{i=1}^{N} \log \delta_i + \sum_{j=1}^{N_r} \log \delta_j^* + \\
\lambda \sum_{j=1}^{N_r} \mu_{\beta}(X_j^*, Z_j^*; \theta) \delta_j^* + \gamma \left( \sum_{i=1}^{N} \delta_i - 1 \right) + \gamma^* \left( \sum_{j=1}^{N_r} \delta_j^* - 1 \right)
\]
Thank You!!