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

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

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

- $\boldsymbol{\theta}$: parameters in the external reduced model, whose estimates $\widehat{\boldsymbol{\theta}}$ are available

Models: Internal Data

- $\boldsymbol{Y}, \boldsymbol{X}$: outcome and (crude, error-prone) covariates
- $\boldsymbol{Z}$: more accurate covariate information available in internal study
- individual data on $\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}$ are available
- $\boldsymbol{f_\beta(y|x,z)}$: model for internal data, assumed to be correctly specified
- $\boldsymbol{\beta}$: parameters in the internal model; aims of inference

Relationship Between Internal and External Models

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

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

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

or

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

$\boldsymbol{\beta_0}$ is true value of $\boldsymbol{\beta}$

- namely $\int \boldsymbol{u_{\beta_0}(x,z;\theta^*)dF(x,z)=0}$ where

$$\boldsymbol{u_\beta(X,Z;\theta^*)=\int_y U(y|X,\theta^*)f_\beta(y|X,Z)dy}$$

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**:

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

  – λ: **Lagrange multipliers**

  – θ is fixed at $\theta = \theta^* \approx \widehat{\theta}$ 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 $(\beta, \lambda, \gamma, \delta_1, \ldots, \delta_m)$

$$\underbrace{\sum_{i=1}^{N} \log f_\beta(Y_i|X_i, Z_i) + \sum_{j=1}^{m} n_j \log \delta_j}_{\text{loglikelihood of internal data}} + \underbrace{\lambda^T \sum_{j=1}^{m} u_\beta(X_j, Z_j; \theta)\delta_j}_{\text{constraint from external data}} + \underbrace{\gamma \left(\sum_{j=1}^{m} \delta_j - 1 \right)}_{\text{constraint for } F}$$

- 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 β

- let $\eta = (\beta^T, \lambda^T)^T$
- $\widehat{\eta} = (\widehat{\beta}^T, \widehat{\lambda}^T)^T$ is the solution to $\partial l^*_{\beta,\lambda}/\partial \eta = 0$
- the proposed **constrained maximum likelihood (CML) estimator**

Computation

- $\widehat{\eta}$ 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 λ is set to zero**

- it is easy to calculate the score and the Hessian, and then do the maximization

Extension to Other Sampling Designs in Internal Study

- 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
- $\boldsymbol{p_1 = 1 - p_0 = \int f_\beta(Y = 1|x, z) dF(x, z)}$ the marginal disease probability for a given value of β

Constrained Likelihood under Case-Control Design

- the likelihood for the internal case-control sample:

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

- **constrained likelihood**:

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

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

$$l^{*,cc}_{\beta,\lambda,\mu_1} = \sum_{i=1}^{N} \log\left\{ \frac{f_\beta(Y_i|X_i,Z_i)}{\sum_y f_\beta(y|X_i,Z_i)\mu_y - \lambda^T u_\beta(X_i,Z_i;\theta)} \right\} + \sum_y N_y \log \mu_y$$

$$\mu_1 = N_1/p_1,\ \mu_0 = N_0/p_0$$

Asymptotic Theory (Qin and Lawless (1994 Ann Stat))

- $\widehat{\eta} = (\widehat{\beta}, \widehat{\lambda}) \longrightarrow_p \eta_0 = (\beta_0^T, 0)^T$
  β_0: true value of β; 0: zero vector with same dimension as λ

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

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

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

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

Asymptotic Properties

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

$$\mathbf{var}\,\widehat{\beta} = (B + CL^{-1}C^T)^{-1} \preceq B^{-1} = \mathbf{var}\,\widehat{\beta}^I$$

- **$\widehat{\beta}$ is asymptotically independent of $\widehat{\lambda}$**

- **asymptotic variance Ω 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**:

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

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

$$\textbf{logit } P(Y = 1|X) = \theta_0 + X\theta_X$$

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 ~ 1.50 and for the interaction ~ 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**

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

- not using external information

- consistent but **losing efficiency when external information is available**

- **generalized regression (GR) (Chen and Chen, 2000 JRSSB)**

$$\widehat{\beta}^{GR} = \widehat{\beta}^{I} + H_1^{-1} C_{12} C_{22}^{-1} H_2 (\widehat{\theta} - \widehat{\theta}^{I})$$

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

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

- 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 %)

	β_0			β_X			β_Z			β_{XZ}		
	Int	GR	**CML**	Int	GR	**CML**	Int	GR	**CML**	Int	GR	**CML**
simple random; $N = 1000$												
Bias	-8.94	2.67	**2.84**	2.42	3.30	**3.37**	1.29	1.50	**0.95**	1.33	1.27	**2.42**
SE	91.4	32.5	**32.4**	96.8	39.0	**38.9**	94.3	94.4	**94.3**	89.4	89.4	**89.5**
ESE	91.8	32.1	**32.3**	92.3	38.8	**38.9**	92.4	92.3	**92.5**	85.8	85.6	**86.9**
MSE	8.42	1.06	**1.06**	9.38	1.53	**1.53**	8.89	8.91	**8.89**	7.98	7.99	**8.01**
CP	95.4	94.7	**95.3**	94.3	93.4	**94.0**	94.6	94.5	**95.1**	93.6	93.7	**93.8**
case-control; $N = 1000$												
Bias	-	-	**2.59**	2.40	14.8	**0.88**	5.06	5.01	**5.11**	-1.51	-1.53	**-1.57**
SE	-	-	**22.7**	75.7	25.1	**26.8**	72.2	72.3	**72.2**	72.9	72.9	**72.8**
ESE	-	-	**22.8**	73.3	26.1	**27.9**	73.1	73.2	**73.2**	71.4	71.4	**71.6**
MSE	-	-	**0.52**	5.73	0.85	**0.72**	5.24	5.24	**5.24**	5.31	5.31	**5.30**
CP	-	-	**94.7**	94.2	91.3	**96.2**	95.4	95.6	**95.4**	94.7	94.4	**94.5**

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**:

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

Y is independent of X given Z
(non-differential measurement error)

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

$$\textbf{logit } P(Y = 1|X) = \theta_0 + X\theta_X$$

Simulation Setups: Mismeasured Covariate

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

- Y:

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

  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 %)

	β_0			β_Z		
	Int	GR	**CML**	Int	GR	**CML**
simple random; $N = 1000$						
Bias	-2.12	-3.73	**0.20**	0.80	1.23	**1.13**
SE	87.7	25.1	**15.1**	89.6	84.7	**40.1**
ESE	87.1	23.9	**15.2**	86.3	82.5	**38.7**
MSE	7.69	0.64	**0.23**	8.02	7.17	**1.61**
CP	95.9	92.6	**94.1**	94.2	94.0	**94.1**
case-control; $N = 1000$						
Bias	-	-	**0.99**	2.85	2.91	**1.74**
SE	-	-	**12.8**	66.0	62.5	**37.6**
ESE	-	-	**12.9**	66.6	63.8	**36.3**
MSE	-	-	**0.16**	4.36	3.63	**1.42**
CP	-	-	**95.6**	95.7	96.1	**94.6**

SE: standard error; ESE: estimated standard error
MSE: mean squared error

Analysis of Relationship Between HZ and COPD Based on the LHID and HIS databaeses

- **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

- **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

model/method	Estimate (SE)
Adjusting Comorbidity (external data)	**0.530**
Adjusting Comorbidity, Smoking & Drinking (internal data)	**1.041 (0.552)**
Generalized Regression	**0.641 (0.059)***
Constrained Maximum Likelihood	**0.620 (0.055)***

***: 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** $\widehat{\theta}$ 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

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

- δ_j^* $(j = 1, \ldots, N_r)$: masses of **external covariate distribution** $F^*(X^*, Z^*)$

- **synthetic constrained likelihood** (Han and Lawless 2016 JASA):

$$\sum_{i=1}^{N} \log f_{\beta}(Y_i|X_i, Z_i) + \sum_{i=1}^{N} \log \delta_i + \sum_{j=1}^{N_r} \log \delta_j^* +$$

$$\lambda \sum_{j=1}^{N_r} u_{\beta}(X_j^*, Z_j^*; \theta)\delta_j^* + \gamma \left(\sum_{i=1}^{N} \delta_i - 1 \right) + \gamma^* \left(\sum_{j=1}^{N_r} \delta_j^* - 1 \right)$$

Thank You !!