Updating Established Prediction Models with New Biomarkers

Jeremy M. G. Taylor, Wenting Cheng, Tian Gu, Jason Estes, Bhramar Mukherjee

Department of Biostatistics, University of Michigan

12th April 2018

Background

- In clinical biomedicine, many well-known models are used to predict a measure of disease from patient characteristics
  - Framingham risk score model for cardiovascular disease
  - Gail model for breast cancer
  - Prostate Cancer Prevention Trial risk calculator
- Possible that including additional variables and constructing an expanded model will improve the prediction ability
- The challenge: additional variables are measured only on a small number of subjects in a new dataset

Background

- Don't have access to the original data from which the established model was built
- The external information may not come in a direct or convenient form
- The information from an existing prediction model can be available in the form of
    1. coefficient estimates (with or without standard errors)
    2. individual prediction probabilities
    3. multiple models for predicting the same outcome

Prostate Cancer Prevention Trial

- Randomized trial of 11,000 men
    - Enrolled men without prostate cancer
    - Followed men for 7 years
    - Did biopsy at the end of 7 years
- Biopsy result
    - No cancer
    - Low grade cancer
    - High grade (bad) cancer
- Other variables
    - PSA - prostate specific antigen - blood draw
    - DRE - digital rectal exam
    - age, race, previous biopsy

Predicting high grade prostate cancer

Prostate Cancer Prevention Trial risk calculator: (Thompson et al 2006)
Predict the risk of high grade prostate cancer based on age at biopsy, race, PSA level, DRE result, previous prostate biopsy.
Uses data from placebo arm of the trial

Online PCPTrc (version 1.0) is based on a logistic regression model

The estimated regression coefficients and their variance-covariance matrices are provided online

University of Michigan Prostate Cancer Data

- PCA3 is a biomarker for early detection of prostate cancer
  - Measured from urine
- Can PCA3 plus PCPT variables improve on PCPT calculator?
- UM study (Tomlins et al 2015)
- Measured age at biopsy, race, PSA level, DRE result, previous prostate biopsy + PCA3
- n=679 in training dataset
- n=1218 in validation dataset

Equations for P(high grade prostate cancer)

Equation for PCPT calculator

$$log(p_i/(1-p_i)) = \beta_0 + \beta_1^T X_i$$

$p_i = Prob$(high grade prostate cancer for subject i)
$X_i =$ the 5 PCPT covariates

Equation in Tomlins et al (2015) that uses PCA3

$$log(p_i/(1-p_i)) = \eta_0 + \eta_1 \hat{p}_i + \eta_2 PCA3_i$$

PCPT high grade calculator + PCA3

Model	PSA	age	dre	prior	AA	Brier Score	AUC
PCPThg	1.29	.03	1.00	-0.36	.96	0.558	0.707
PCPThg+PCA3	–	–	–	–	–	0.568	0.752

AUC improves with using PCA3. Could further improvement be made using better statistical approach?
Brier score does not improve with using PCA3.

Notation

- Y is the outcome variable
- X is a set of standard risk covariates
- B is a new biomarker

- Form of outcome variable Y
  - Continuous
  - Binary
  - Survival time
- Form of new biomarker B
  - Continuous, Gaussian
  - Binary
  - General distribution
  - Multiple new biomarkers

Assumptions

- E denotes external data
- I denotes internal data, size n
- assume (for now) $[Y|X,B]_E = [Y|X,B]_I$
- assume (for now) $[B|X]_E = [B|X]_I$
- hence $[Y|X]_E = [Y|X]_I$
- $[X]_E$ may be different from $[X]_I$
- assume $[Y|X]_E$ model does fit the external data, and all the X's are important
- typically $dim(X)$ is 3 - 10, and $dim(B)$ is 1 - 3

Possible statistical approaches

- Use ideas from calibration in survey research (Lumley et al 2011)
- Update predictions directly (Grill et al 2015)
- Constrained maximum likelihood, where constraint derived from score equation (Chatterjee et al 2016)
- Use empirical likelihood ideas (Han and Lawless 2017)
- Our approach - directly link parameters
  - Parameters β in model $E[Y|X, \beta]$ are known
  - Parameters γ in model $E[Y|X, B, \gamma]$ are unknown
  - From exact or approximate relationship between γ and β
- Synthetic data approach (Reiter)

Models: Y and B continuous

The model of primary interest is:

$$E(Y|X,B) = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_p X_p + \gamma_{p+1} B$$

We could also estimate $E(B|X)$ in this small dataset:

$$E(B|X) = \theta_0 + \theta_1 X_1 + \cdots + \theta_p X_p$$

The established model for Y given X:

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

The knowledge from the previous study is summary-level information: $\bar{\beta}_j$ **and** $\bar{S}_j, j = 0, \ldots, p$

Relationship Equations: Y and B continuous

$$
\begin{aligned}
E(Y|X) &= E[E(Y|X,B)|X] \\
&= \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_p X_p + \gamma_{p+1} \times E(B|X)
\end{aligned}
$$

Hence

$$\beta_j = \gamma_j + \gamma_{p+1}\theta_j, j = 0, \ldots, p$$

Thus

$$\gamma_j \neq \beta_j, j = 1, \ldots, p$$

Thus, should not use prediction from established model as an offset

Constrained Solutions: Constrained Maximum Likelihood

$$\min_{\gamma,\theta}\left\{\frac{1}{\sigma_1^2}\sum_{i=1}^{n}(Y_i-\sum_{j=0}^{p}\gamma_j X_{ij}-\gamma_{p+1}B_i)^2+\frac{1}{\sigma_2^2}\sum_{i=1}^{n}(B_i-\sum_{j=0}^{p}\theta_j X_{ij})^2\right\}$$

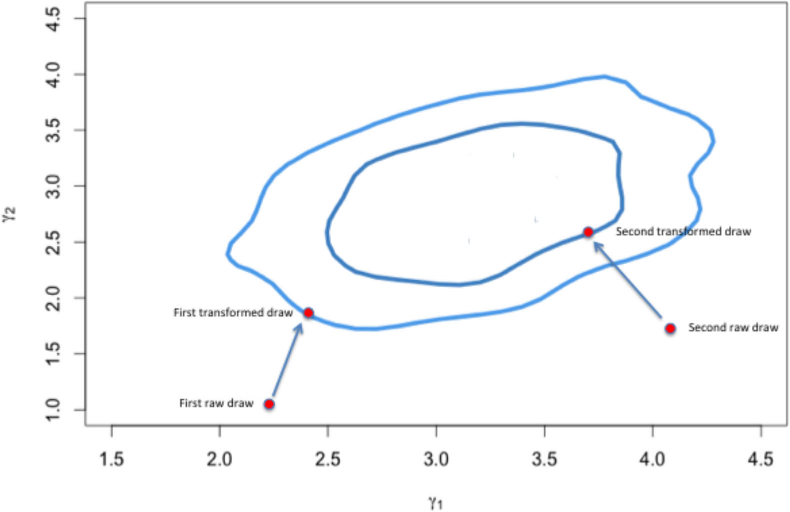
$s.t.\gamma_j+\gamma_{p+1}\theta_j\in[\bar{\beta}_j-d\bar{S}_j,\bar{\beta}_j+d\bar{S}_j],j=0,\ldots,p$

- d is a scale parameter, d=1
- An optimization problem with nonlinear inequality constraints

Constrained Solutions: Informative Full Bayes

- Bayesian approach with informative priors, using a Metropolis-Hastings sampling algorithm
- The joint likelihood function: $L(Y|X, B, \gamma, \sigma_1^2) \times L(B|X, \theta, \sigma_2^2)$
- $(\theta, \gamma, \sigma_1^2, \sigma_2^2) \to (\beta, \gamma, \sigma_1^2, \sigma_2^2)$
- Incorporate information on β directly: $\beta_j = \gamma_j + \gamma_{p+1}\theta_j \sim N(\bar{\beta}_j, \bar{S}_j^2), j = 0, \ldots, p$
- Non-informative priors for $\gamma, \sigma_1^2, \sigma_2^2$
- It is not computationally efficient

Constrained Solutions: Bayesian Transformation Approach



Constrained Solutions: Bayesian Transformation Approach

- Draws from standard Bayes: $\gamma_0, \dots, \gamma_{p+1}$, $\theta_0, \dots, \theta_p$
- OLS estimates: $s^2_{\gamma_0}, \dots, s^2_{\gamma_{p+1}}, s^2_{\theta_0}, \dots, s^2_{\theta_p}$
- Then $\gamma^\star, \theta^\star$ are obtained by solving the optimization problem:

$$\min_{\gamma^\star,\theta^\star} \left\{ \sum_{j=0}^{p+1} \frac{(\gamma_j - \gamma_j^\star)^2}{s^2_{\gamma_j}} + \sum_{k=0}^{p} \frac{(\theta_k - \theta_k^\star)^2}{s^2_{\theta_k}} \right\}$$
$$\text{s.t.} \gamma_j^\star + \gamma_{p+1}^\star \theta_j^\star \in [\bar{\beta}_j - d\bar{S}_j, \bar{\beta}_j + d\bar{S}_j], j = 0, \dots, p$$
where $d \sim |N(0,1)|$

- Repeat the above step for each draw until all draws are transformed

Simulation Studies

- Compared to not using external information
  - Considerable gain in efficiency for estimating $\gamma_1, \gamma_2, ..., \gamma_p$
  - Very little gain in efficiency for estimating γ_{p+1}
  - Modest gains in efficiency for prediction of future Y, as measured by R^2
  - Limited gains if sample size (n) is large

Binary outcome

The model of primary interest:

$$\mathrm{logit}(\mathrm{Pr}(\mathrm{Y}=1|\mathrm{X},\mathrm{B}))=\gamma_0+\gamma_1\mathrm{X}_1+\cdots+\gamma_\mathrm{p}\mathrm{X}_\mathrm{p}+\gamma_{\mathrm{p}+1}\mathrm{B}$$

Another model can be estimated in the small dataset:

$$f(B|X)=\eta(\theta_0+\theta_1X_1+\cdots+\theta_pX_p)$$

Established model for Y given X is:

$$\mathrm{logit}(\mathrm{Pr}(\mathrm{Y}=1|\mathrm{X}))=\beta_0+\beta_1\mathrm{X}_1+\cdots+\beta_\mathrm{p}\mathrm{X}_\mathrm{p}$$

The knowledge from the previous study is summary-level information : $\bar{\beta}_j$ and $\bar{S}_j, j=0,\ldots,p$

Logistic Regression Approximation

- A logistic model for $Pr(Y = 1|X, B)$ does not reduce to a logistic model for $Pr(Y = 1|X)$.
- Key linking relationship

$$\mathrm{Pr(Y = 1|X)} = \int Pr(Y = 1|X, B)P(B|X)dB$$

Logistic Regression Approximation

- Assume $B|X \sim N(X\theta, \sigma_2^2)$
- Need to approximate $\mathrm{logit}(\mathrm{Pr}(\mathrm{Y} = 1|\mathrm{X}, \gamma, \theta))$ and equate that to $\mathrm{logit}(\mathrm{Pr}(\mathrm{Y} = 1|\mathrm{X}, \beta)) = \mathrm{X}\beta$

$$\mathrm{Pr}(\mathrm{Y} = 1|\mathrm{X}, \gamma, \theta) = \int Pr(Y = 1|X, B, \gamma)P(B|X, \theta)dB$$

$$= \frac{\int H(\gamma_0 + X\gamma_x + B\gamma_B)e^{-\frac{(B-X\theta)^2}{2\sigma_2^2}} dB}{\sqrt{2\pi\sigma_2^2}}$$

where $H(v) = \{1 + e^{-v}\}^{-1}$ is the logistic distribution function

Logistic Regression Approximation

An approximate relationship between γ, θ and β :

$$\beta_j \approx \frac{\gamma_j + \gamma_{p+1}\theta_j}{(1+\gamma_{p+1}^2\sigma_2^2/1.7^2)^{\frac{1}{2}}}, j = 0, \ldots, p$$

Estimation methods for Gaussian Y can be adapted to handle binary Y

- Constrained MLE
- Bayes with informative priors
- Bayes transformation approach

Predicting high grade prostate cancer

Prostate Cancer Prevention Trial risk calculator:
predict the risk of high grade prostate cancer based on age at biopsy, race, family history, PSA level, DRE result, previous prostate biopsy

Online PCPTrc (version 1.0) is based on a logistic regression model

The estimated regression coefficients and their variance-covariance matrices are provided online

Application to Prostate Cancer Data

- PCA3 is a biomarker for early detection of prostate cancer
- Can PCA3 plus PCPT variables improve on PCPT calculator?
  - AUC for discrimination
  - Brier score for accuracy = MSE
  - Calibration plots
- n=679 in training dataset
- n=1218 in validation dataset

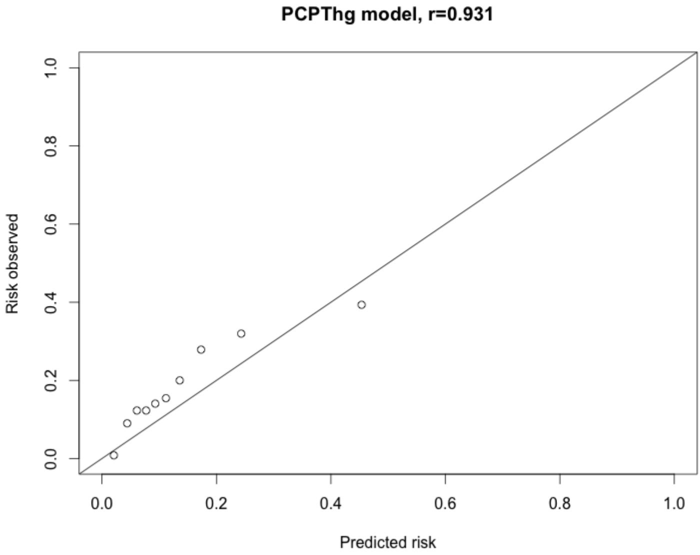
PCPT high grade calculator + PCA3

Model	PSA	age	dre	prior	AA	PCA3	Brier Score	AUC
PCPThg	1.29	.03	1.00	-0.36	.96	–	0.558	0.707
PCPThg+PCA3	–	–	–	–	–	–	0.568	0.752
Non-informative Bayes	0.98	.01	1.05	-1.27	.04	0.56	0.567	0.767
Constrained ML	1.32	.01	0.95	-0.40	.67	0.56	0.533	0.762
Informative Bayes	1.21	.01	0.97	-0.74	.28	0.59	0.566	0.767
Transformation	1.23	.01	0.96	-0.49	.41	0.55	0.530	0.764

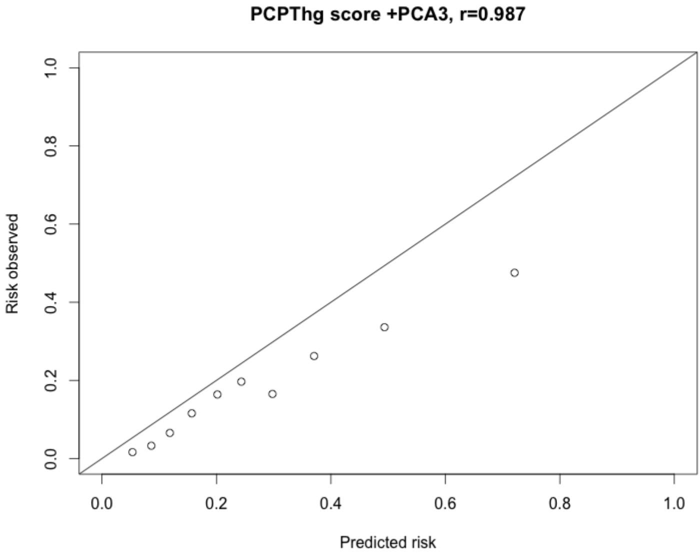
Calibration plots

- Sort the 1218 predictions $\hat{p}_i$ from smallest to largest
- Separate into 10 equal groups
- For each group calculate observed response rate
- Scatterplot of predicted p versus observed p
- Should be close to the diagonal

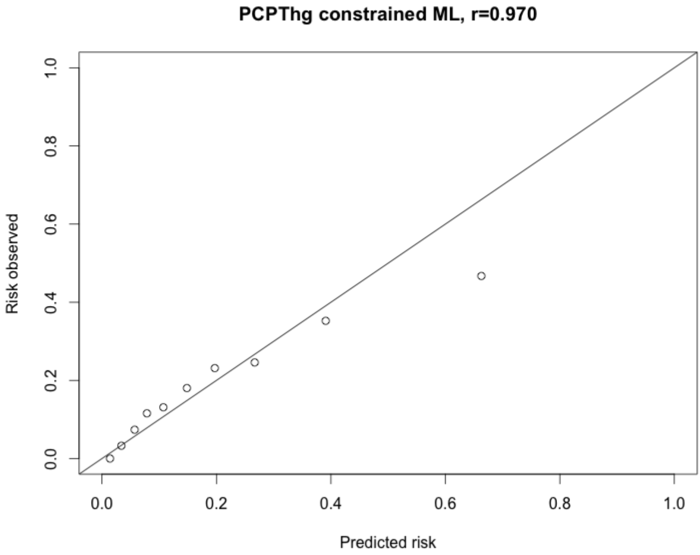
Calibration plot: PCPT calculator



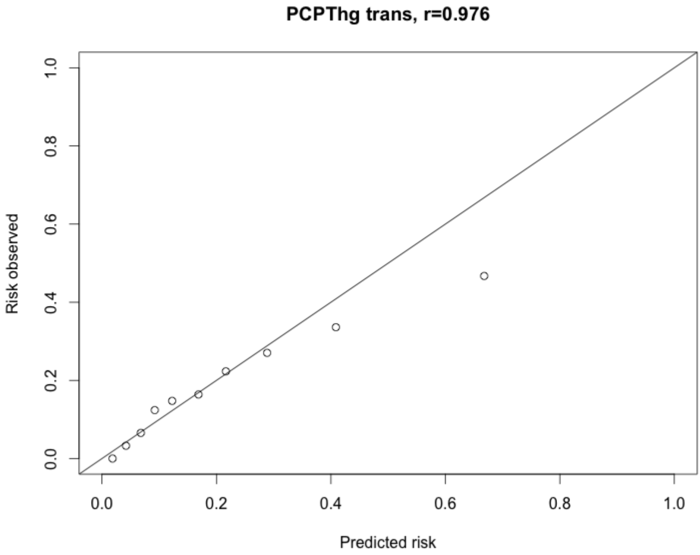
Calibration plot: Published PCPT+PCA3



Calibration plot: Constrained ML



Calibration plot: Transformation approach



Discussion of prostate cancer example

- PCA3 is a useful biomarker that adds a lot
- Sample size is quite large (n=679)
- Simple methods were all that was needed for this application
- The distribution of X differed considerably between the PCPT study, the training data and the validation data.

Discussion

- Need to model $B|X$
  - Methods that directly link parameters don't generalize well if distribution of B is complex
- How to link parameters between $Y|X$ and $Y|X, B$ has to be worked out on a case by case basis
- Want methods that work even if distribution of X differs between external data and the training data
- Do you want an equation for $\hat{Y}$ or just a way of predicting $\hat{Y}$?
- Maybe the new data comes from a case-control study
  - Trust the coefficients of X in $Y|X$ model
  - Don't trust the intercept in $Y|X$ model

Chatterjee et al 2016 approach

- Flexible, applies for any form of distribution of B
- Does not need to model $B|X$
- Does require $[X, B]_E = [X, B]_I$
- assumes that $SE(\bar{\beta}) = 0$
- Maximizes the likelihood wrt γ and $F_n(X, B)$, where $F_n(X, B)$ is the discrete distribution of (X, B), subject to constraint derived from integrated score equation using $Y|X$ model.
- denote estimator $\hat{\gamma}_C$

Estes et al, submitted

- E denotes external population
- I denotes internal population
- Maybe $[Y|X, B]_E \neq [Y|X, B]_I$
- Maybe $[Y|X]_E \neq [Y|X]_I$
- Which population do you want your prediction model to apply to?
- Chatterjee estimator $\hat{\gamma}_C$ can be writen as $\psi(\beta_E)$
- MLE $\hat{\gamma}_I = \psi(\hat{\beta}_I)$

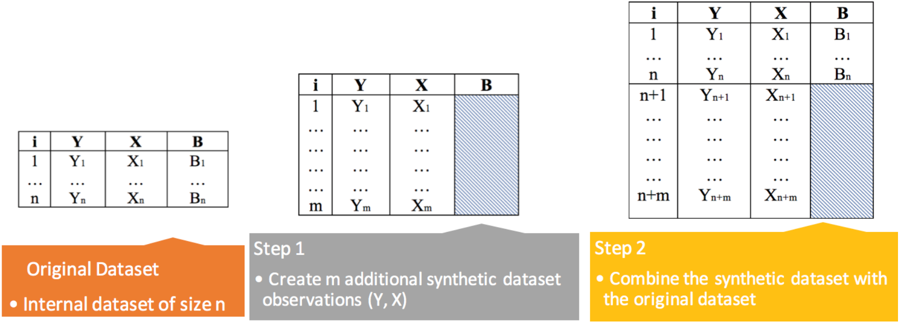
Estes et al, Empirical Bayes estimator

- Empirical Bayes approach
- idea shrink β_I towards β_E
- hence shrink $\hat{\gamma}_I$ towards $\hat{\gamma}_E$
  - $\beta \sim N(\beta_0, A)$
  - Bayes estimate $\psi(\beta) = W\psi(\hat{\beta}_I) + (I - W)\psi(\beta_0)$
  - $\hat{\gamma} = \hat{W}\psi(\hat{\beta}_I) + (I - \hat{W})\psi(\beta_E)$
  - $\hat{\gamma} = \hat{W}\hat{\gamma}_I + (I - \hat{W})\hat{\gamma}_C$

Synthetic data approach

- Original data (Y, X, B) of size n
- Append to it synthetic data (Y, X) of size m, m is large
- Analyze combined dataset of size $n + m$ using methods that handle missing data
- Construction of synthetic data
  - replicate X
  - For each row of synthetic data generate a new Y from known model for $[Y|X]$
- Analyze combined dataset using multiple imputation techniques (if exact methods are not possible)

Synthetic data construction



i	Y	X	B
1	Y_1	X_1	B_1
…	…	…	…
n	Y_n	X_n	B_n

Original Dataset

- Internal dataset of size n

i	Y	X	B
1	Y_1	X_1	▨
…	…	…	
…	…	…	
…	…	…	
…	…	…	
m	Y_m	X_m	

Step 1

- Create m additional synthetic dataset observations (Y, X)

i	Y	X	B
1	Y_1	X_1	B_1
…	…	…	…
n	Y_n	X_n	B_n
n+1	Y_{n+1}	X_{n+1}	▨
…	…	…	
…	…	…	
…	…	…	
…	…	…	
n+m	Y_{n+m}	X_{n+m}	

Step 2

- Combine the synthetic dataset with the original dataset

Figure: Data cnstruction for synthetic data method

A result when (Y,X,B) is trivariate normal

- Form of the data
  - (Y,X,B) for n observations
  - (Y,X) for m observations
  - Interested in $[Y|X, B, \gamma]$
  - $Y = \gamma_0 + \gamma_1 X + \gamma_2 B + \sigma\epsilon$
- 9 parameters
- reparametrize in different ways
  - as mean and variance of MVN(Y,X,B)
  - as parameters of $[Y|X, B]$ and $[X, B]$

A result when (Y,X,B) is trivariate normal

- METHOD 1, analyze n+m observations, 9 parameters
- There is an analytic form for MLE $\hat{\gamma}_{(1)}$, (Gourieroux and Monfort 1981)
- Get variance of $\hat{\gamma}_{(1)}$ from information matrix

A result when (Y,X,B) is trivariate normal

- METHOD 2, analyze n observations with constraints on parameters
- Constraint, $\beta_j = \gamma_j + \theta_j \gamma_2$, $j = 0, 1$
- $\theta_j = (\beta_j - \gamma_j)/\gamma_2$, β_j known
- Variance of $[Y|X]$ is also known, another constraint
- 6 free parameters
- Maximize likelihood, $L(\gamma, \theta, (Y, B|X))$ over 6 parameters
- get $\hat{\gamma}_{(2)}$, and variance of $\hat{\gamma}_{(2)}$ from information matrix

A result when (Y,X,B) is trivariate normal

- RESULT
- for infinitely large m, Asymp Var $\hat{\gamma}_{(1)}$ = Asymp Var $\hat{\gamma}_{(2)}$
- Also Asymp Var $\hat{\gamma}_{(1)}$ less than Asymp Var from $\hat{\gamma}_{(Chatterjee)}$

A result when (Y,X,B) are all binary

- The result also holds if Y, X, B are all binary, with saturated models
- For binary Asymp Var $\hat{\gamma}_{(1)}$ = Asymp Var from $\hat{\gamma}_{(Chatterjee)}$

Proof of result that constrained MLE is equivalent to synthetic data method

- The $[Y, B|X]$ model can be factored in different ways
  - $f(Y, B|X, \omega)$
  - $f(Y|B, X, \gamma)f(B|X, \theta)$
  - $f(Y|X, \beta)f(B|Y, X, \phi)$
- Assume $[Y|B, X]$ model is compatible with the known $[Y|X]$ model
- Assume a 1-to-1 relationship between parameters
  - If you know β and ϕ,
  - then ω is uniquely determined,
  - then γ and θ are uniquely determined
- assume m/n is very large

Proof of result that constrained MLE is equivalent to synthetic data method

- constrained MLE is
    - $max \prod_1^n f(Y_i, B_i|X_i, \omega)$ with β fixed
- synthetic data method has likelihood
    - $\prod_1^n f(Y_i, B_i|X_i, \omega) \prod_1^m f(Y_k|X_k, \beta)$
    - or equivalently $\prod_1^n f(Y_i, B_i|X_i, \beta, \phi) \prod_1^m f(Y_k|X_k, \beta)$
- Since $m >> n$ the 2nd term completely dominates the 1st term and hence essentially determines the β's on its own, and gives the correct value of β.
- Hence the method reduces to maximizing
- $\prod_1^n f(Y_i, B_i|X_i, \beta, \phi)$ over ϕ with β fixed,
- which is exactly the same as the constrained MLE.

How to implement the synthetic data method in practice

- Choose m
- Replicate X matrix m/n times
- Generate m new values of Y from $[Y|X]$ model
- Data structure
  - n observations with (Y, X, B)
  - m observations with (Y, X)
- It is a missing data problem
- Use multiple imputation as a device to analyze these "observed" data
  - Create V complete datasets by imputing B
  - Analyze each dataset using the desired model $[Y|X, B]$
  - Combine the V analysis results in the usual way to get estimates of γ
- Need flexible and robust methods of imputing B from $[B|X, Y]$
- Want method of imputing B to be compatible with $[Y|X, B]$ model

Notes on the synthetic data method

- You don't need to know the exact model for $[Y|X]$, just need to be able to generate data from that model
- Can handle multiple B's, using chained equations MI
- Can extend if there are multiple different known models for $[Y|X]$
  - $[Y|X^1]$, $[Y|X^2]$, ... , $[Y|X^R]$
  - Where $X^1, X^2, \ldots, X^R$ may differ from each other, but be overlapping

References

- Grill et al, 2016, J Clin Epi
- Steyerberg et al, 2000, Stat in Med
- Newcombe et al, 2012, Genetic Epi
- Gunn and Dunson, 2005, Biostatistics
- Qin et al, 2015, Biometrika
- Chatterjee et al, 2016, JASA
- Chen et al, 2015, Elect J Stat
- Tomlins et al, 2016, European Urology
- Thompson et al, 2006, JNCI
- Cheng et al, 2018, Stat in Med