Exploiting Auxiliary information to improve power for testing genetic effects with binary or time-to-event data

Xiaomin Lu

Department of Biostatistics

University of Florida

http://users.phhp.ufl.edu/xlu2

(Joint work with Jinbo Chen)

Outlines

- Motivating example: a genetic association study
- Model framework and notation
- Improved estimators when response is binary
- Improved estimators when response is time-to-event

Motivating example: a genetic association study

A randomized study on relapsed acute lymphoblastic leukemia (ALL): novel agent + backbone vs backbone alone:

- Primary outcomes: 1) Complete response (CR) at the end of Block 1; 2) event free survival (EFS)
- Objectives: if there association of cytokine receptor gene CRLF2 with the primary outcomes and if the association interact with the assigned treatment
- Commonly used method:
  - Logistic regression model: MLE
    *proc logistic; model CR = trt gen trt*gen; run;*
  - Proportional hazard regression model: MPLE
    *proc phreg; model time*status(0) = trt gen trt*gen; run;*

Still a room to improve the efficiency of MLE and MPLE or detect genetic effect with smaller sample size?

Note that:

- The standard methods only used the information on primary response, treatment assignment and genetic data.
- Additional baseline auxiliary covariates (e.g., age,gender,race,etc.) are usually collected before randomization,
  - Independent of treatment assignment
  - Some are correlated with the primary outcomes, and/or independent of the genetic data

Making use of the information of such auxiliary covariates will increase the efficiency of the MLE/MPLE and hence increase the power to detect the corresponding genetic effect.

How?

The most powerful test may be developed "*in the ideal world*", i.e. when we know for each individual the potential response for each gene expression level and each treatment assignment.

In reality, several sources of lost information are usually observed in randomized clinical trials:

- Not able to observe both the response to treatment and the response to placebo on the same individual.
- Not able to observe the response every gene expression level on the same individual.
- If primary outcome is EFS: not able to observe the actual time-to-event if censored for some individual.

Use the auxiliary covariates that correlate with primary outcome to recover lost information.

Model framework and notation

Observed data $D_i = (Y_i, Z_i, G_i, X_i)$, $i = 1, \ldots, n$.

- $Y_i = (U_i, \Delta_i)$ if it's censored survival time, where $U_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$
  - T_i denotes the underlying survival time
  - C_i denotes the potential censoring time
- Z_i - treatment indicator
- G_i - genetic variable, usually categorical
- X_i - a vector of auxiliary covariates

Improved estimators when response is binary

Assumption1: $E(Y|Z,G) = \text{expit}(\alpha + \beta_1 Z + \beta_2 G + \beta_3 ZG)$.

Assumption2: $Z \perp\!\!\!\perp (G,X)$ and $Pr(Z=1) = \pi$.

Assumption3: $G \perp\!\!\!\perp X$.

All semiparametric estimators for β_i, i=1,2,3 can be written as the solution to the estimation equations

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \Big([\vec{a}(Y_i, Z_i, G_i) - \hat{E}\{\vec{a}(Y, Z, G)|G = G_i\}] \\ &+ (Z_i - \pi)\vec{f}_1(X_i, G_i) + \{G_i - \overline{G_i}\}\{\vec{f}_2(X_i) - \overline{\vec{f}}_2\} \Big) \end{aligned}$$

for arbitrary functions $\vec{a}(Y, Z, G)$, $\vec{f}_1(X, G)$ and $\vec{f}_2(X)$.

Let $\vec{a}(Y, Z, G) = (1, Z, G, ZG)^T\{Y - E(Y|Z, G)\}$ and $\vec{f}_1(\cdot) = \vec{f}_2(\cdot) = 0 \Rightarrow$ MLE of the logistic regression parameters

Improved estimator than MLE of logistic regression parameters:

- Propose parametric models $\vec{f}_1(X, G, \mathbf{a}_1) = \mathbf{a}_1 \vec{q}_1(X, G)$, $\vec{f}_2(X, \mathbf{a}_2) = \mathbf{a}_1 \vec{q}_2(X)$, and consider the subclass of RAL estimators that solve the estimating equations

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \Big[(1, Z_i, G_i, Z_i G_i)^T \{ Y_i - \text{expit}(\alpha + \beta_1 Z_i + \beta_2 G_i + \beta_3 Z_i G_i) \} \\ &+ (Z_i - \pi) \vec{f}_1(X_i, G_i, \mathbf{a}_1) + \{ G_i - \overline{G} \} \{ \vec{f}_2(X_i, \mathbf{a}_2) - \overline{\vec{f}}_2(\mathbf{a}_2) \} \Big] \end{aligned}$$

  for all $\mathbf{a}_1 \in \mathbb{R}^{4 \times r_{a_1}}$ and $\mathbf{a}_2 \in \mathbb{R}^{4 \times r_{a_2}}$.

- The optimal value of $(\mathbf{a}_1, \mathbf{a}_2)$ that leads to the smallest variance for the estimated βs can be obtained using standard regression method, e.g. OLS.

Simulation study:

We considered one baseline covariate X and one binary genetic variable. The data were generated to satisfy:

- Z, G, X are mutually independent.
- $P(Z = 1) = 0.5$ and $P(G = 1) = 0.49$.
- $P(Y = 1|Z, G) = \text{expit}(\alpha + \beta_1 Z + \beta_2 G + \beta_3 ZG)$.
- Correlation between X and Y is $\sim$0.5.
- Posited the models $\vec{f}_1(X, G, \mathbf{a}_1) = a_{10} + a_{11}G + a_{12}GX + a_{13}X + a_{14}X^2$ and $\vec{f}_2(X, \mathbf{a}_2) = a_{21}X + a_{22}X^2$.

Table 1: Logistic regression (5000 simulation sets)

Hypothesis		**Sample Size**	**Estimator**	**Bias**	**SE**	**MCSE**	**MSE**	**Power**	**95% CI**
Null	Genetic effect	500	MLE	-0.002	0.255	0.259	0.067	0.055	0.945
			Improved	0.000	0.221	0.224	0.050	0.055	0.945
		1000	MLE	-0.004	0.180	0.181	0.033	0.053	0.947
			Improved	-0.004	0.156	0.157	0.025	0.051	0.949
	Interaction effect	500	MLE	0.004	0.360	0.368	0.136	0.057	0.943
			Improved	-0.002	0.312	0.319	0.102	0.056	0.944
		1000	MLE	0.006	0.254	0.259	0.067	0.055	0.945
			Improved	0.003	0.220	0.225	0.051	0.052	0.948
Alternative	Genetic effect	500	MLE	0.003	0.257	0.263	0.069	0.427	0.946
			Improved	0.007	0.215	0.219	0.048	0.572	0.946
		1000	MLE	0.001	0.181	0.184	0.034	0.704	0.945
			Improved	0.002	0.152	0.154	0.024	0.846	0.948
	Interaction effect	500	MLE	0.010	0.382	0.393	0.155	0.427	0.943
			Improved	0.006	0.323	0.333	0.111	0.561	0.945
		1000	MLE	0.006	0.269	0.275	0.076	0.704	0.948
			Improved	0.004	0.227	0.234	0.055	0.839	0.944

30-45% improving in efficiency

Improved estimators when response is time-to-event

Assumption1: $\lambda_{T|Z,G}(t|Z,G) = \lambda(t)\exp(\beta_1 Z + \beta_2 G + \beta_3 ZG)$.

Assumption2: $Z \perp\!\!\!\perp (G, X)$ and $Pr(Z = 1) = \pi$.

Assumption3: $G \perp\!\!\!\perp X$

Assumption4: Non-informative censoring $C \perp\!\!\!\perp (T, X)|(Z, G)$.

All semiparametric estimators for β_i, i=1,2,3 can be represented as the solution to

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \Big[\int \{\vec{a}(u, Z_i, G_i) - \bar{\vec{a}}(u, \beta)\} dN_i(u) \\ &+ (Z_i - \pi)\vec{f}_1(X_i, G_i) + \{G_i - \bar{G}\}\{\vec{f}_2(X_i) - \bar{\vec{f}}_2\} \\ &+ \int \{\vec{b}(u, Z_i, G_i, X_i) - \bar{\vec{b}}(u, Z_i, G_i)\} dN_{C_i}(u) \Big], \end{aligned}$$

for arbitrary functions $\vec{a}(u, Z, G)$, $\vec{f}_1(X, G)$, $\vec{f}_2(X)$, and $\vec{b}(u, Z, G, X)$, where $N_i(u) = I(U_i \leq u, \Delta_i = 1)$ and $N_{C_i}(u) = I(U_i \leq u, \Delta_i = 0)$ are the counting process which counts the number of observed events and censored observations, respectively, up to and including time u for patient i,

$$\bar{\vec{a}}(u) = \frac{\sum_{i=1}^{n} \vec{a}(u, Z_i, G_i) \exp(\beta_1 Z_i + \beta_2 G_i + \beta_3 Z_i G_i) Y_i(u)}{\sum_{i=1}^{n} \exp(\beta_1 Z_i + \beta_2 G_i + \beta_3 Z_i G_i) Y_i(u)},$$

$$\text{and } \bar{\vec{b}}(u, Z_i, G_i) = \frac{\sum_{j=1}^{n} \vec{b}(u, Z_j, G_j, X_j) I(Z_j = Z_i) I(G_j = G_i) Y_j(u)}{\sum_{j=1}^{n} I(Z_j = Z_i) I(G_j = G_i) Y_j(u)}.$$

Let $\vec{a}(u, Z, G) = (Z, G, ZG)^T$ and $\vec{f}_1(\cdot) = \vec{f}_2(\cdot) = \vec{b}(\cdot) = 0 \Rightarrow$ MPLE

Improved estimator than MPLE of Cox regression model:

- Propose parametric models $\vec{f}_1(X, G, \mathbf{a}_1) = \mathbf{a}_1 \vec{q}_1(X, G)$, $\vec{f}_2(X, \mathbf{a}_2) = \mathbf{a}_1 \vec{q}_2(X)$ and $\vec{b}(u, Z, G, X; \mathbf{b}) = \mathbf{b}^T \vec{w}(u, Z, G, X)$, and consider the subclass of RAL estimators which solve the estimating equations

$$\begin{aligned} 0 &= \sum_{i=1}^{n} \Big[\{(Z, G, ZG)^T - \overline{(Z, G, ZG)^T}(u, \beta)\} dN_i(u) \\ &+ (Z_i - \pi)\vec{f}_1(X_i, G_i, \mathbf{a}_1) + \{G_i - \overline{G}\}\{\vec{f}_2(X_i, \mathbf{a}_2) - \overline{\vec{f}}_2(\mathbf{a}_2)\} \\ &+ \int \{\vec{b}(u, Z_i, G_i, X_i, \mathbf{b}) - \overline{\vec{b}}(u, Z_i, G_i\mathbf{b})\} dN_{C_i}(u) \Big] \end{aligned}$$

  for all $\mathbf{a}_1 \in \mathbb{R}^{3 \times r_{a_1}}$, $\mathbf{a}_2 \in \mathbb{R}^{3 \times r_{a_2}}$ and $\mathbf{b} \in \mathbb{R}^{3 \times r_b}$.

- The optimal $(\mathbf{a}_1, \mathbf{a}_2)$ and $\mathbf{b}$ can be obtained by using standard regression method, respectively.

Simulation study:

Again, we considered one baseline covariate X and one binary genetic variable. The data were generated to satisfy:

- Z, G, X are mutually independent.
- $P(Z = 1) = 0.5$ and $P(G = 1) = 0.49$.
- $\lambda_{T|Z,G}(t|z, g) = \lambda(t) \exp(\beta_1 Z + \beta_2 G + \beta_3 ZG)$.
- Correlation between X and T is $\sim$0.6.
- Posited the models $\vec{f}_1(X, G, \mathbf{a}_1) = a_{10} + a_{11}G + a_{12}GX + a_{13}X + a_{14}X^2$, $\vec{f}_2(X, \mathbf{a}_2) = a_{21}X + a_{22}X^2$ and $\vec{b}_2(u, Z, G, X, \mathbf{b}) = b_1 G + b_2 X + b_3 XZ + b_4 GZ$.

Table 2: Cox proportional hazard regression (sample size=500, 5000 simulation sets)

Hypothesis		**Censor Prop.**	**Estimator**	**Bias**	**SE**	**MCSE**	**MSE**	**Power**	**95% CI**
Null	Genetic effect	0.25	MPLE	-0.001	0.147	0.147	0.022	0.049	0.951
			Improved	-0.003	0.112	0.114	0.013	0.058	0.942
		0.50	MPLE	0.000	0.181	0.180	0.032	0.049	0.951
			Improved	-0.004	0.139	0.142	0.020	0.055	0.945
	Interaction effect	0.25	MPLE	0.004	0.208	0.208	0.043	0.049	0.951
			Improved	0.003	0.158	0.160	0.026	0.056	0.944
		0.50	MPLE	0.003	0.256	0.258	0.067	0.052	0.948
			Improved	0.002	0.196	0.200	0.040	0.055	0.945
Alternative	Genetic effect	0.25	MPLE	-0.002	0.141	0.142	0.020	0.698	0.949
			Improved	-0.005	0.088	0.091	0.008	0.980	0.945
		0.50	MPLE	-0.001	0.173	0.174	0.030	0.521	0.948
			Improved	-0.007	0.109	0.111	0.012	0.904	0.945
	Interaction effect	0.25	MPLE	-0.002	0.210	0.211	0.045	0.672	0.947
			Improved	-0.004	0.135	0.138	0.019	0.961	0.945
		0.50	MPLE	-0.003	0.258	0.264	0.070	0.498	0.947
			Improved	-0.006	0.166	0.172	0.030	0.857	0.941

60-156% improving in efficiency