Robust methods for analyzing secondary phenotypes in case-control genetic association studies

Andrew S. Allen

Department of Biostatistics and Bioinformatics,
Duke University

DukeMedicine

Joint work with:

- Chuanhua Xing
- Janice McCarthy
- Josée Dupuis
- L. Adrienne Cupples
- James Meigs
- Xihong Lin

DukeMedicine

Outline

- Case-control study and secondary phenotypes
- Previous approaches
- Our approach
- Simulation study
- Example
- Discussion

Case-control study

- Comprised of two separate samples:
  - Cases–with disease
  - Controls–without disease
- Allows oversampling of cases (so similar number as controls)
- Minimize # of exposures that need to be assessed for a given level of statistical power
- Economical approach for assessing association between (genetic) exposures and disease

Case-control study

- Measuring exposures in genetic association studies is expensive
  - GWAS
  - Whole exome/genome sequencing
- ‘Make the most’ of considerable investment

Case-control study

- Measuring exposures in genetic association studies is expensive
  - GWAS
  - Whole exome/genome sequencing
- 'Make the most' of considerable investment → Secondary phenotypes

Secondary phenotypes

- Most studies measure phenotypes in addition to primary (case-control)
  - opportunistic
  - related to underlying disease process

- Studying genetic influences on secondary phenotype may be of interest in itself or may help understanding of underlying disease process

Problem

- Case-control study does not constitute a random sample from the general population
- If sampling isn't taken into account during analysis

Problem

- Case-control study does not constitute a random sample from the general population
- If sampling isn't taken into account during analysis

  → association between genetic variant and secondary phenotype can be BIASED

Previous Studies

- Richardson et al. (2007) [1] proposed a weighted regression model.
- Monsees et al. (2009) [2] extended the approach it be applicable to more general phenotypes and genetic exposures. Both approaches require that the sampling probabilities are known (nested case-control design).
- Lin and Zeng (2009) [3] proposed a method (SPREG) based on retrospective likelihood of the genotype and secondary phenotypes conditional on the disease status.
- Li et al. (2010) [4] proposed a rare disease model assuming binary secondary phenotype.
- Wei et al. (2013) [5] proposed a robust regression approach.

Our Approach: preview

- Based on inverse-probability-weighted estimating equations from restricted-moment-model framework
  - Flexible modeling of various types of secondary phenotypes
  - Covariates
- Computationally efficient
- Provides practical tool for genome-wide analyses
- For clarity, this presentation will focus on linear model

DukeMedicine

Notation

G – genotype information; i.e., $G = 0, 1, 2$

D – case-control status. $D = 1$ if case; $D = 0$ if control

Y – secondary phenotype

n_1 – $\#$ of cases

n_0 – $\#$ of controls

Random sampling

- If Y is a quantitative phenotype, we can model the relationship between Y and G by

$$Y = \beta_0 + \beta_1 G + \epsilon,$$

where $\beta = (\beta_0, \beta_1)^T$ are parameters and $E(\epsilon|G) = 0$

Random sampling

- If Y is a quantitative phenotype, we can model the relationship between Y and G by

$$Y = \beta_0 + \beta_1 G + \epsilon,$$

where $\beta = (\beta_0, \ \beta_1)^T$ are parameters and $E(\epsilon|G) = 0$

- If subjects $(G_i, Y_i); i = 1, ..., n$ represent a random sample from the population, we can estimate β by obtaining the root, $\hat{\beta}$, of the following estimating equations

$$U_\beta = \sum_{i=1}^{n} U_{\beta,i}(Y_i, G_i) = \left(\begin{array}{c} \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 G_i) \\ \sum_{i=1}^{n} G_i(Y_i - \beta_0 - \beta_1 G_i) \end{array} \right)$$

DukeMedicine

Random sampling

- If Y is a quantitative phenotype, we can model the relationship between Y and G by

  $$Y = \beta_0 + \beta_1 G + \epsilon,$$

  where $\beta = (\beta_0,\ \beta_1)^T$ are parameters and $E(\epsilon | G) = 0$

- If subjects $(G_i, Y_i); i = 1, ..., n$ represent a random sample from the population, we can estimate β by obtaining the root, $\hat{\beta}$, of the following estimating equations

  $$U_\beta = \sum_{i=1}^{n} U_{\beta,i}(Y_i, G_i) = \begin{pmatrix} \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 G_i) \\ \sum_{i=1}^{n} G_i(Y_i - \beta_0 - \beta_1 G_i) \end{pmatrix}$$

- $\hat{\beta}$ is a consistent estimator of the true population β

DukeMedicine

Case-control sampling

- However, when D is related to G and/or Y, the distribution of (G, Y) obtained under case-control sampling can be distorted from the population distribution

Case-control sampling

- However, when D is related to G and/or Y, the distribution of (G, Y) obtained under case-control sampling can be distorted from the population distribution

- Thus, if $(G_i, Y_i); i = 1, ..., n_0 + n_1$ represents a combined case-control sample, the root, $\hat{\beta}_{naive}$, of

$$U_\beta = \sum_{i=1}^{n_0+n_1} U_{\beta,i}(Y_i, G_i) = \left(\begin{array}{c} \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 G_i) \\ \sum_{i=1}^{n} G_i(Y_i - \beta_0 - \beta_1 G_i) \end{array} \right)$$

is not always a consistent estimator of the true population β

Case-control sampling

- However, when D is related to G and/or Y, the distribution of (G, Y) obtained under case-control sampling can be distorted from the population distribution

- Thus, if $(G_i, Y_i); i = 1, ..., n_0 + n_1$ represents a combined case-control sample, the root, $\hat{\beta}_{naive}$, of

$$U_\beta = \sum_{i=1}^{n_0+n_1} U_{\beta,i}(Y_i, G_i) = \left(\begin{array}{c} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 G_i) \\ \sum_{i=1}^{n} G_i(Y_i - \beta_0 - \beta_1 G_i) \end{array} \right)$$

is not always a consistent estimator of the true population β

- We call $\hat{\beta}_{naive}$ the *naive* estimator of β

DukeMedicine

Our approach

We can prove the following:

$$E\left[\frac{U_\beta(G,Y)}{1-p_{GY}}\middle| D=0\right]=0 \iff E_*[U_\beta(G,Y)]=0$$

and

$$E\left[\frac{U_\beta(G,Y)}{p_{GY}}\middle| D=1\right]=0 \iff E_*[U_\beta(G,Y)]=0,$$

- '$*$' indicates that this expectation is taken with respect to the *true* distribution that generated G and Y in the *population*
- p_{GY} denotes the conditional probability of being a case in the *population*

DukeMedicine

Our approach

Thus, if we define two new estimating equations as

$$\widetilde{U}_\beta^0 = \sum_{i=1}^{n_0} \frac{U_{\beta,i}(Y_i, G_i)}{1 - p_{G_i Y_i}}$$

and

$$\widetilde{U}_\beta^1 = \sum_{i=n_0+1}^{n_1+n_0} \frac{U_{\beta,i}(Y_i, G_i)}{p_{G_i Y_i}},$$

the roots, $\hat{\beta}^0$ of $\widetilde{U}_\beta^0$ and $\hat{\beta}^1$ of $\widetilde{U}_\beta^1$, will each be consistent estimators of the population β

Estimating p_{GY} with case-control data

If we model p_{GY} as

$$p_{GY} \equiv Pr(D = 1|G, Y) = \frac{e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}{1 + e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}$$

Estimating p_{GY} with case-control data

If we model p_{GY} as

$$p_{GY} \equiv Pr(D = 1|G, Y) = \frac{e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}{1 + e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}$$

- γ_1 and γ_2 can be reliably estimated from case-control data

Estimating p_{GY} with case-control data

If we model p_{GY} as

$$p_{GY} \equiv Pr(D = 1|G, Y) = \frac{e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}{1 + e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}$$

- γ_1 and γ_2 can be reliably estimated from case-control data
- In general, the population intercept is not identifiable from case-control data

Estimating p_{GY} with case-control data

If we model p_{GY} as

$$p_{GY} \equiv Pr(D = 1|G, Y) = \frac{e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}{1 + e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}$$

- γ_1 and γ_2 can be reliably estimated from case-control data
- In general, the population intercept is not identifiable from case-control data
- However, we can still estimate p_{GY} in two complementary cases:

Estimating p_{GY} with case-control data

If we model p_{GY} as

$$p_{GY} \equiv Pr(D = 1|G, Y) = \frac{e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}{1 + e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}$$

- γ_1 and γ_2 can be reliably estimated from case-control data
- In general, the population intercept is not identifiable from case-control data
- However, we can still estimate p_{GY} in two complementary cases:
  1. When the population prevalence is known

DukeMedicine

Estimating p_{GY} with case-control data

If we model p_{GY} as

$$p_{GY} \equiv Pr(D = 1|G, Y) = \frac{e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}{1 + e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}$$

- γ_1 and γ_2 can be reliably estimated from case-control data
- In general, the population intercept is not identifiable from case-control data
- However, we can still estimate p_{GY} in two complementary cases:
  1. When the population prevalence is known
  2. When the disease is rare in the population

Estimating p_{GY} with known population prevalence

- Let γ_0^* be the intercept implied by applying the logistic regression model to case-control data. Let λ denote the true population disease prevalence, then

$$\gamma_0 = \gamma_0^* + \log\left(\frac{n_0}{n_1}\right) + \log\left(\frac{\lambda}{1-\lambda}\right)$$

Estimating p_{GY} with known population prevalence

- Let γ_0^* be the intercept implied by applying the logistic regression model to case-control data. Let λ denote the true population disease prevalence, then

$$\gamma_0 = \gamma_0^* + \log\left(\frac{n_0}{n_1}\right) + \log\left(\frac{\lambda}{1-\lambda}\right)$$

- Thus, we can estimate p_{GY} by

$$\widehat{p}_{GY} = \frac{e^{\widehat{\gamma}_0^* + \log\left(\frac{n_0}{n_1}\right) + \log\left(\frac{\lambda}{1-\lambda}\right) + \widehat{\gamma}_1 G + \widehat{\gamma}_2 Y}}{1 + e^{\widehat{\gamma}_0^* + \log\left(\frac{n_0}{n_1}\right) + \log\left(\frac{\lambda}{1-\lambda}\right) + \widehat{\gamma}_1 G + \widehat{\gamma}_2 Y}}, \qquad (1)$$

where $\widehat{\gamma}_0^*, \widehat{\gamma}_1, \widehat{\gamma}_2$ are the parameter estimates obtained by applying logistic regression to the case-control sample

Estimating p_{GY} under a rare disease assumption

- When the disease is rare in the population, we have that

$$\begin{aligned} p_{GY} &= Pr(D=1|G,Y) \approx e^{\gamma_0+\gamma_1 G+\gamma_2 Y} \\ 1-p_{GY} &= Pr(D=0|G,Y) \approx 1 \end{aligned}$$

DukeMedicine

Estimating p_{GY} under a rare disease assumption

- When the disease is rare in the population, we have that

$$\begin{aligned} p_{GY} &= Pr(D=1|G,Y) \approx e^{\gamma_0+\gamma_1 G+\gamma_2 Y} \\ 1-p_{GY} &= Pr(D=0|G,Y) \approx 1 \end{aligned}$$

- In this case

$$\begin{aligned} \widetilde{U}_\beta^0 &= \sum_{i=1}^{n_0} \frac{U_\beta(Y_i,G_i)}{1-p_{G_iY_i}} \approx \sum_{i=1}^{n_0} U_\beta(Y_i,G_i) \\ \widetilde{U}_\beta^1 &= \sum_{i=n_0+1}^{n} \frac{U_\beta(Y_i,G_i)}{p_{G_iY_i}} \approx e^{-\gamma_0} \sum_{i=n_0+1}^{n} \frac{U_\beta(Y_i,G_i)}{e^{\gamma_1 G_i+\gamma_2 Y_i}} \end{aligned}$$

DukeMedicine

Estimating p_{GY} under a rare disease assumption

- When the disease is rare in the population, we have that

$$\begin{aligned} p_{GY} &= Pr(D=1|G,Y) \approx e^{\gamma_0+\gamma_1 G+\gamma_2 Y} \\ 1-p_{GY} &= Pr(D=0|G,Y) \approx 1 \end{aligned}$$

- In this case

$$\begin{aligned} \widetilde{U}_\beta^0 &= \sum_{i=1}^{n_0} \frac{U_\beta(Y_i,G_i)}{1-p_{G_iY_i}} \approx \sum_{i=1}^{n_0} U_\beta(Y_i,G_i) \\ \widetilde{U}_\beta^1 &= \sum_{i=n_0+1}^{n} \frac{U_\beta(Y_i,G_i)}{p_{G_iY_i}} \approx e^{-\gamma_0} \sum_{i=n_0+1}^{n} \frac{U_\beta(Y_i,G_i)}{e^{\gamma_1 G_i+\gamma_2 Y_i}} \end{aligned}$$

- Thus γ_0 does not affect estimation of β

Combining estimators or estimating equations

Recall that we are interested in estimating β_1 in the linear model

$$Y = \beta_0 + \beta_1 G + \epsilon$$

and we have shown how we can estimate β_1 from cases ($\hat{\beta}_1^1$) and controls ($\hat{\beta}_1^0$)

Combining estimators or estimating equations

Recall that we are interested in estimating β_1 in the linear model

$$Y = \beta_0 + \beta_1 G + \epsilon$$

and we have shown how we can estimate β_1 from cases ($\hat{\beta}_1^1$) and controls ($\hat{\beta}_1^0$)

How should we combine $\hat{\beta}_1^1$ and $\hat{\beta}_1^0$?

Combining estimators

We consider the weighted combination: $a_0\hat{\beta}_1^0 + a_1\hat{\beta}_1^1$, where

$$a^T \equiv (a_0, a_1) = \frac{\mathbf{1}^T V^{-1}}{\mathbf{1}^T V^{-1}\mathbf{1}},$$

$\mathbf{1}^T = (1, 1)$, and V is the variance-covariance matrix of $\hat{\beta}_1^0$ and $\hat{\beta}_1^1$.

Combining estimators

We consider the weighted combination: $a_0\hat{\beta}_1^0 + a_1\hat{\beta}_1^1$, where

$$a^T \equiv (a_0, a_1) = \frac{\mathbf{1}^T V^{-1}}{\mathbf{1}^T V^{-1}\mathbf{1}},$$

$\mathbf{1}^T = (1, 1)$, and V is the variance-covariance matrix of $\hat{\beta}_1^0$ and $\hat{\beta}_1^1$.

Note: Derivation of variance estimator proceeds via a standard Taylor series argument with modifications for case-control sampling (details in manuscript)

Simulated Data

- Genotype G_i is generated using a minor allele frequency 0.3 assuming Hardy-Weinberg equilibrium
- Y_i is generated using $Y_i = \beta_0 + \beta_1 G_i + \epsilon$, where $\epsilon \sim N(0,1)$ or $\epsilon \sim (\chi^2_2 - 2)/2$
- D_i is generated using the logistic model

$$p_{GY} \equiv Pr(D = 1|G, Y) = \frac{e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}{1 + e^{\gamma_0+\gamma_1 G+\gamma_2 Y}}$$

- We set $\beta_0 = \sigma^2 = 1$, and assume that the null hypothesis is $\beta_1 = 0$ and the alternative hypothesis is $\beta_1 =$ -0.12
- $\gamma_0 = log(\frac{\eta_0}{1-\eta_0})$ with $\eta_0 = 0.001$ and 0.1, $\gamma_1 = log(1.0), ..., log(1.5)$, and $\gamma_2 = 0, log(2)/2, log(2)$
- We selected 1000 cases and 1000 controls, and repeated the simulation 10,000 times

DukeMedicine

% Bias, MSE, Type I error, and Power when $\epsilon \sim N(0,1)$

		Rare disease				Common disease				
	γ_1	IPW_R	$SPREG_R$	NAÏVE	$COND$	IPW_K	$SPREG_K$	NAÏVE	$COND$	$IPSW_K$
% Bias	0	0.3333	0.0833	3.0833	0.0833	0.5000	0.4167	1.4167	1.5833	0.0833
	log(1.2)	0.4167	0.8333	10.3333	0.1667	0.4167	0.1667	4.6667	5.0833	0.5000
	log(1.4)	0.0833	1.1667	21.0833	0.0000	0.0833	0.6667	8.4167	12.000	0.0833
MSE	0	0.0013	0.0012	0.0012	0.0012	0.0013	0.0011	0.0012	0.0012	0.0018
	log(1.2)	0.0013	0.0012	0.0013	0.0012	0.0013	0.0012	0.0012	0.0012	0.0017
	log(1.4)	0.0013	0.0011	0.0018	0.0011	0.0013	0.0012	0.0013	0.0013	0.0017
Type I error	0	0.0106	0.0100	0.0121	0.0106	0.0112	0.0130	0.0096	0.0096	0.0115
	log(1.2)	0.0107	0.0140	0.0439	0.0096	0.0113	0.0120	0.0109	0.0131	0.0100
	log(1.4)	0.0093	0.0100	0.1676	0.0090	0.0128	0.0070	0.0141	0.0208	0.0108
Power	0	0.7570	0.8170	0.8267	0.8152	0.7671	0.8190	0.8170	0.7998	0.5994
	log(1.2)	0.7717	0.8240	0.7079	0.8276	0.7828	0.8240	0.7773	0.8727	0.6141
	log(1.4)	0.7825	0.8560	0.5868	0.8404	0.7918	0.8370	0.7460	0.9210	0.6392

% Bias, MSE, Type I error, and Power when $\epsilon \sim (\chi^2_2 - 2)/2$

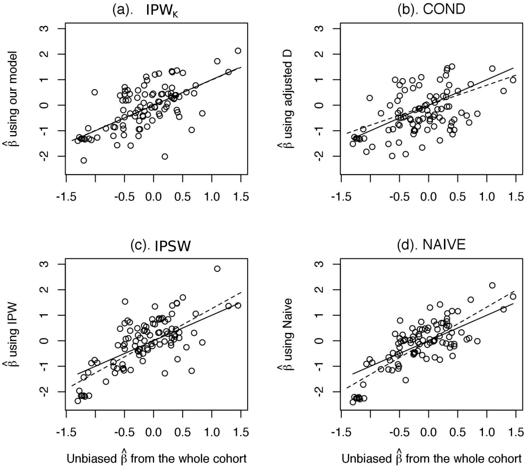
		Rare disease				Common disease				
	γ_1	IPW_R	$SPREG_R$	NAÏVE	$COND$	IPW_K	$SPREG_K$	NAÏVE	$COND$	$IPSW_K$
% Bias	0	0.2500	0.0833	4.0833	0.5833	1.0833	1.0833	1.0000	2.4167	0.2175
	log(1.2)	0.3333	0.5833	15.3333	0.0000	0.4167	2.4167	2.0833	9.2500	0.3775
	log(1.4)	0.7500	1.3333	31.4167	0.6667	0.9167	2.1667	3.0833	20.333	0.1433
MSE	0	0.0010	0.0020	0.0021	0.0020	0.0011	0.0018	0.0015	0.0015	0.0015
	log(1.2)	0.0010	0.0020	0.0024	0.0020	0.0011	0.0019	0.0015	0.0015	0.0014
	log(1.4)	0.0010	0.0018	0.0035	0.0020	0.0011	0.0018	0.0014	0.0020	0.0013
Type I Error	0	0.0093	0.0060	0.0103	0.0108	0.0112	0.0100	0.0115	0.0118	0.0118
	log(1.2)	0.0103	0.0150	0.0200	0.0113	0.0101	0.0090	0.0092	0.0183	0.0117
	log(1.4)	0.0117	0.0130	0.0501	0.0102	0.0121	0.0100	0.0100	0.0393	0.0099
Power	0	0.8867	0.5490	0.5748	0.5474	0.8633	0.5710	0.6957	0.6725	0.7121
	log(1.2)	0.8967	0.5530	0.3890	0.5666	0.8661	0.6090	0.6874	0.8128	0.7310
	log(1.4)	0.9041	0.5810	0.2518	0.5755	0.8703	0.6370	0.6943	0.8949	0.7508

Example

- Extracted case-control sample from unrelated Framingham cohort (case: BMI> 30)
- Diabetics excluded
- Sampled 243 cases and 243 controls from cohort (1114 with GWAS data)
- Fasting blood glucose (FBG) is considered to be secondary phenotype
- FBG and BMI are known to be related
- Estimate relationship (β) between FBG and 100 SNPs most associated with case-control status:
  1. From case-control sample using secondary phenotype analyses
  2. From entire cohort
- Regress βs from 1 against βs from 2

DukeMedicine

Example



DukeMedicine

Example

Results from Framingham example

	Slope	Standard Error	95% CI
IPW_K	0.99	0.1139	[0.7607, 1.2163]
COND	0.78	0.1323	[0.5197, 1.0488]
IPSW	1.26	0.1205	[1.0185, 1.5005]
NAÏVE	1.32	0.1012	[1.1171, 1.5220]

Discussion

- For illustration, we presented our approach in the context of a linear model without covariates
  - Developed approach within a more general restricted moment model framework
  - Can model binary, count data etc.
  - Covariates can also be included
- Our approach is computationally efficient
  - SPREG takes ≈10 times more computing time (worse when null is approximately true)

Manuscript

C Xing, JM McCarthy, J Dupuis, LA Cupples, JB Meigs, X Lin, AS Allen. Robust analysis of secondary phenotypes in case-control genetic association studies. *Statistics in Medicine*. epub 30 May 2016. DOI: 10.1002/sim.6976

References

- Li, H., Gail, M., Berndt, S., and Chatterjee, N. (2010) Using cases to strengthen inference on the association between single nucleotide polymorphisms and a secondary phenotype in genome-wide association studies. *Genet Epidemiol.*, **34**, 427-433.
- Lin, D. Y. and Zeng, D. (2009) Proper analysis of secondary phenotype data in case-control association studies. *Genet Epidemiol.*, **33(3)**, 256-65.
- Monsees, G. M., Tamimi, R. M., and Kraft, P. (2009) Genome-wide association scans for secondary traits using case-control samples. *Genet Epidemiol.*, **33**, 717-728.
- Richardson, D. B., Rzehak, P., Klenk, J., and Weiland, S. K. (2007) Analyses of case-control data for additional outcomes. *Epidemiology*, **18**, 441-445.
- Roeder K, Carroll RJ, Lindsay BG. 1996. A semiparametric mixture approach to case-control studies with errors in covariables. *J Am Stat Assoc*, **91**, 722732.

DukeMedicine