Outline

Costs and Benefits of Environmental Data in Investigations of Gene-Disease Associations

Hao Luo

Department of Statistics, UBC

March 17, 2012

Genetic association study

In the epidemiological world, large-scale genome-wide association studies are conducted to investigate the role of genetics in disease processes and locate disease susceptibility loci.
In such studies, the researchers measure disease status and genotype on subjects and test for the marginal gene effect.

It is generally accepted that complex diseases are likely to be caused by the interplay of both genetic and environmental factors. Many people have suggested that concurrent environmental data should also be collected.

There is no FREE LUNCH!

It is typically costly to obtain exposure data, and this cost could instead be applied to measure disease status and genotype on more subjects.
If the study resources are fixed, we could either measure disease status, genotype, and environmental exposure for a smaller sample or measure only disease status and genotype for a larger sample.

This project aims to compare the statistical power to detect gene-disease associations from two different datatypes, accounting for the cost of collecting data.

(Y,G,X) data

Model:

$$\text{logit}Pr(Y = 1|X, G) = \beta_0 + \beta_1 X + \beta_2 G + \beta_3 XG.$$

Null and alternative hypotheses:

$$H_0 : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad \text{vs.} \quad H_a : \begin{pmatrix} \beta_2 \\ \beta_3 \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

(Y,G) data

Model:

$$\text{logit}Pr(Y = 1|G) = \alpha_0 + \alpha_1 G.$$

Null and alternative hypotheses:

$$H_0 : \alpha_1 = 0 \quad \text{vs.} \quad H_a : \alpha_1 \neq 0$$

"Mendelian Randomization" assumption:

Being exposed is independent of having a specific gene in study population

Under this assumption, $(\beta_2, \beta_3) = (0, 0)$ if and only if $\alpha_1 = 0$. Consequently, we can speak unambiguously about the null hypothesis of no gene effect, or more precisely the null hypothesis of conditional independence between Y and G given X, and this null hypothesis can be tested either using (Y,X,G) data or (Y,G) data.

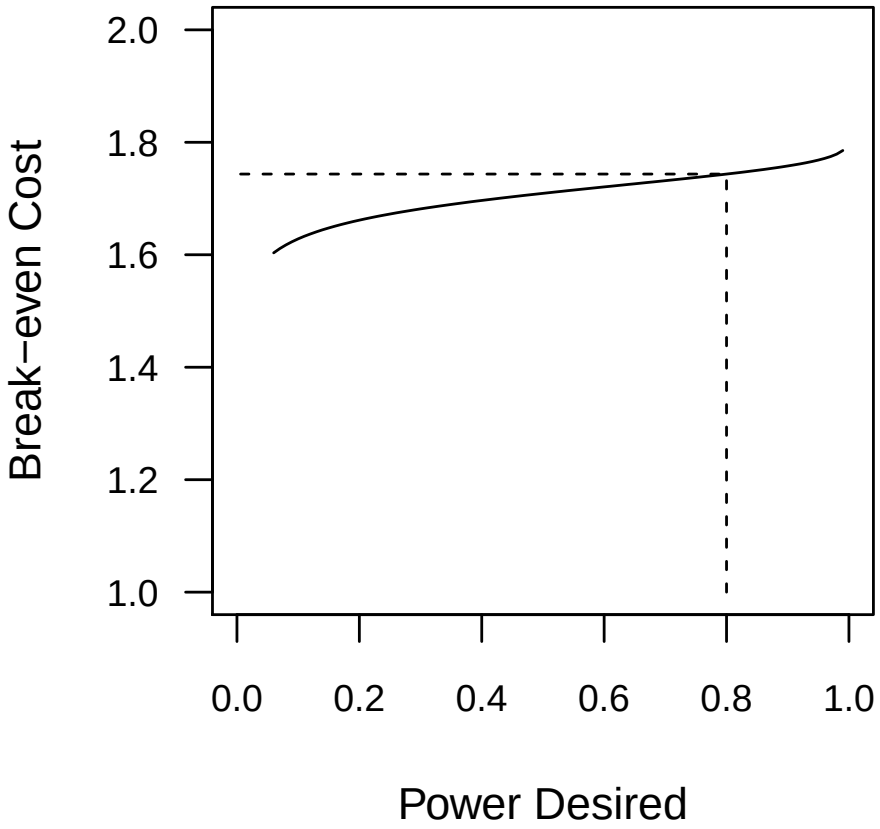
We presume that the cost of measuring Y, X and all the genetic markers on a subject is c times the cost of measuring Y and the markers alone.

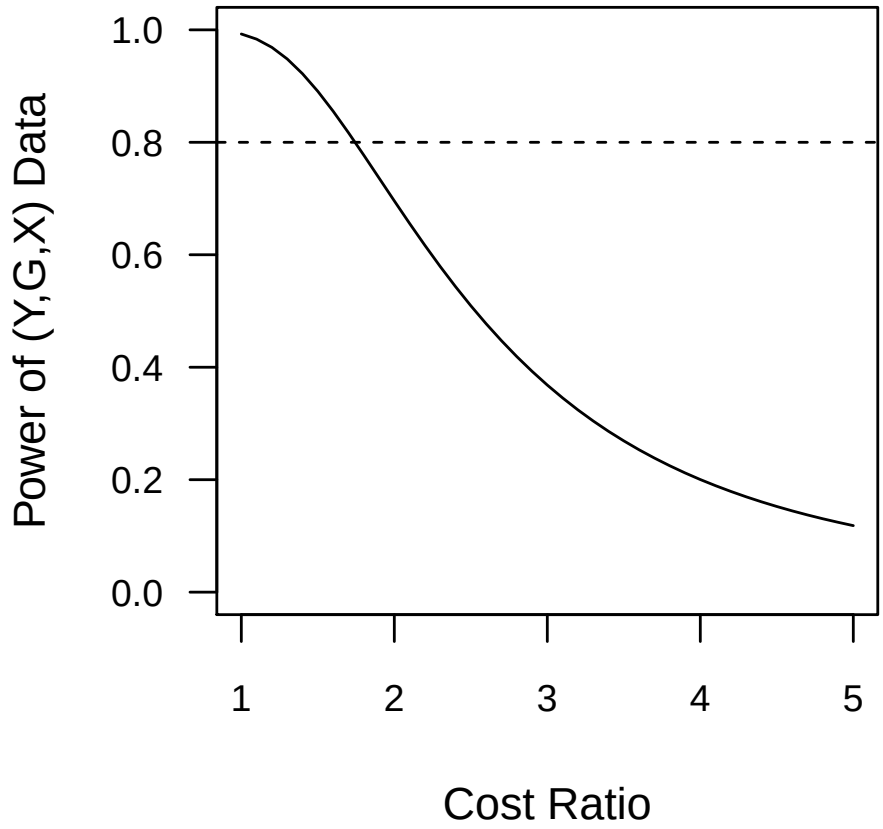
To help our comparison, we introduce the break-even cost c^*, which is the value of c for which the same outlay applied to a smaller sample including X or a larger sample excluding X will yield the same power to detect a gene effect.

If the actual cost ratio c exceeds c^*, then collecting only (Y,G) data and fitting the reduced model is a better use of resources than collecting (Y,X,G) data and fitting the full model.

Break-even Cost

$$c^* = N_{(Y,G)}/N_{(Y,G,X)}.$$



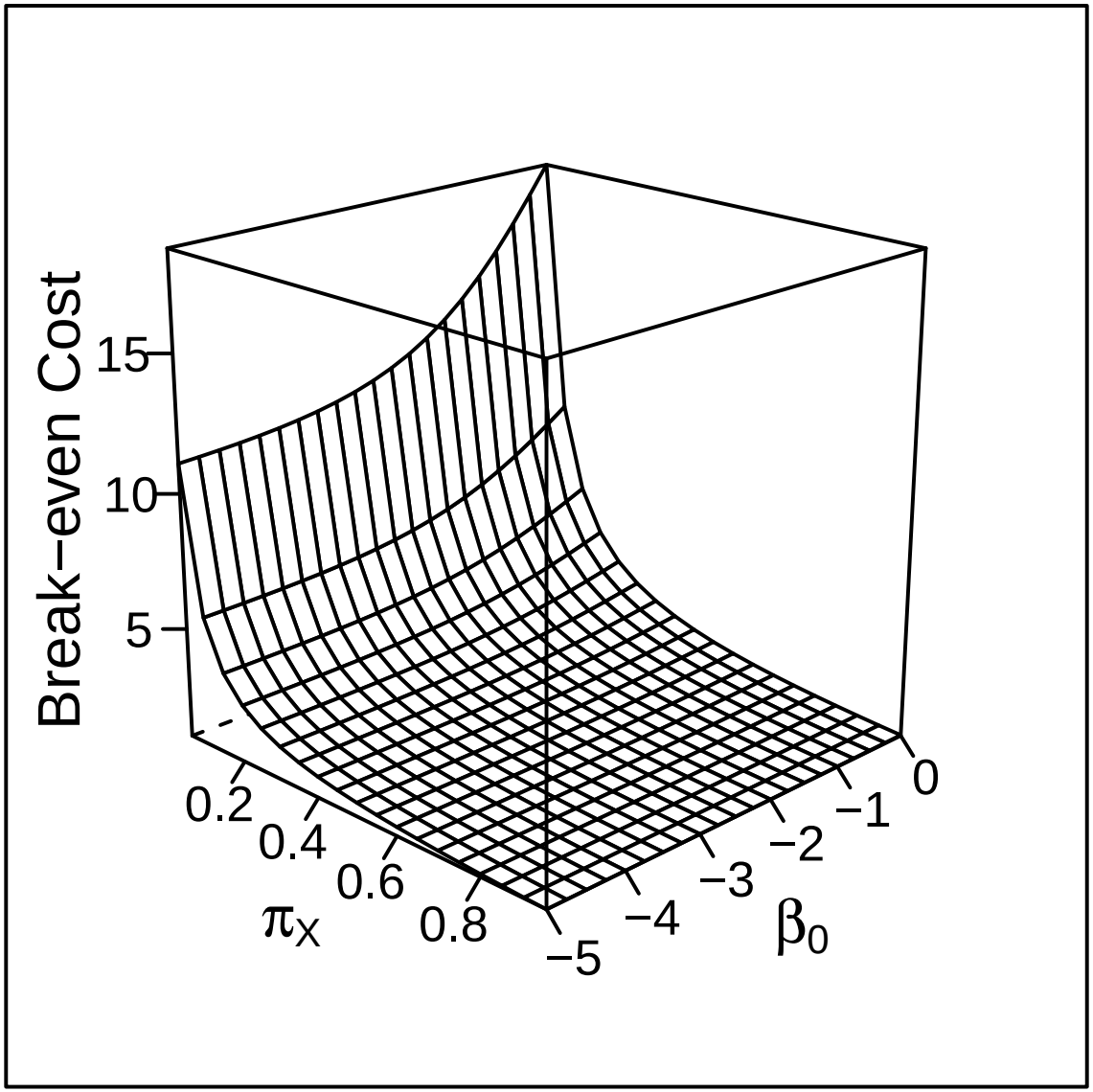


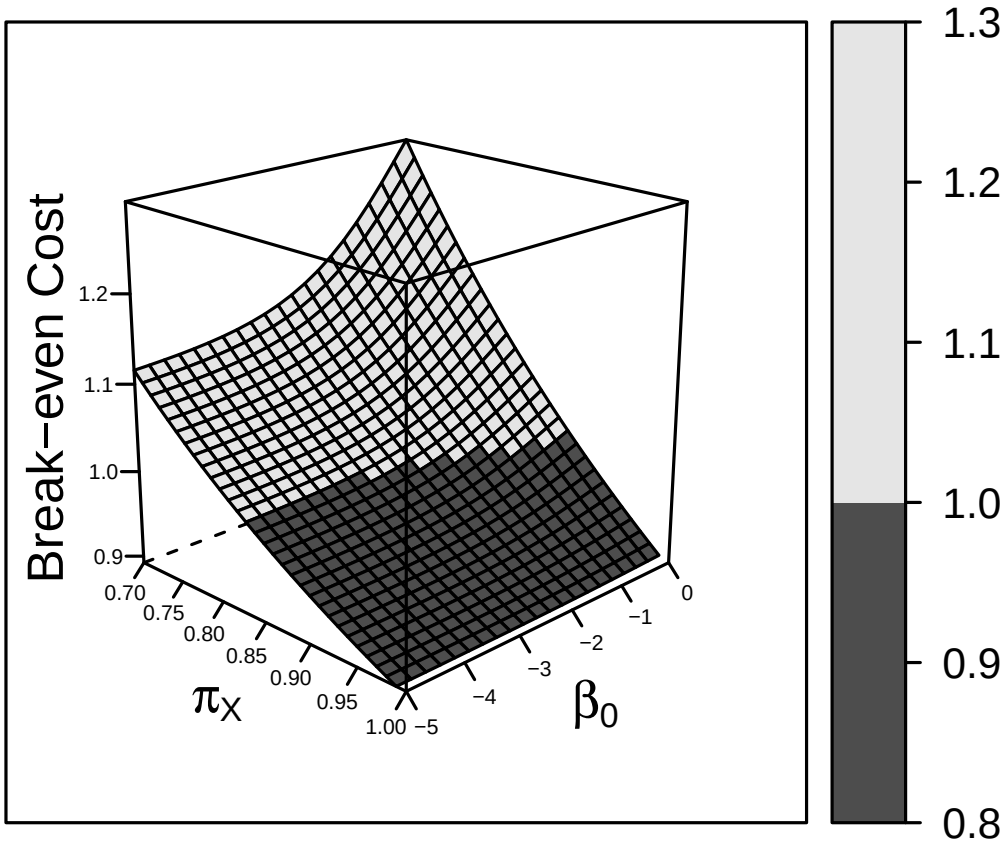
Factorial Experiment: Assess the generality of our findings.

	From	Step	To
π_G	0.05	0.05	0.5
π_X	0.05	0.05	0.5
β_0	-5	0.5	0
β_1	0.1	0.1	2
β_3	0.1	0.1	2

Qualitative Interaction

Gene alone confers no additional disease risk in the absence of exposure.


Break−even Cost
15
10
5
0.2
0.4
0.6
0.8
π_X
−5
−4
−3
−2
−1
0
β_0



Break-even cost can even be below 1.

X data would decrease power even if they could be perfectly measured for free!

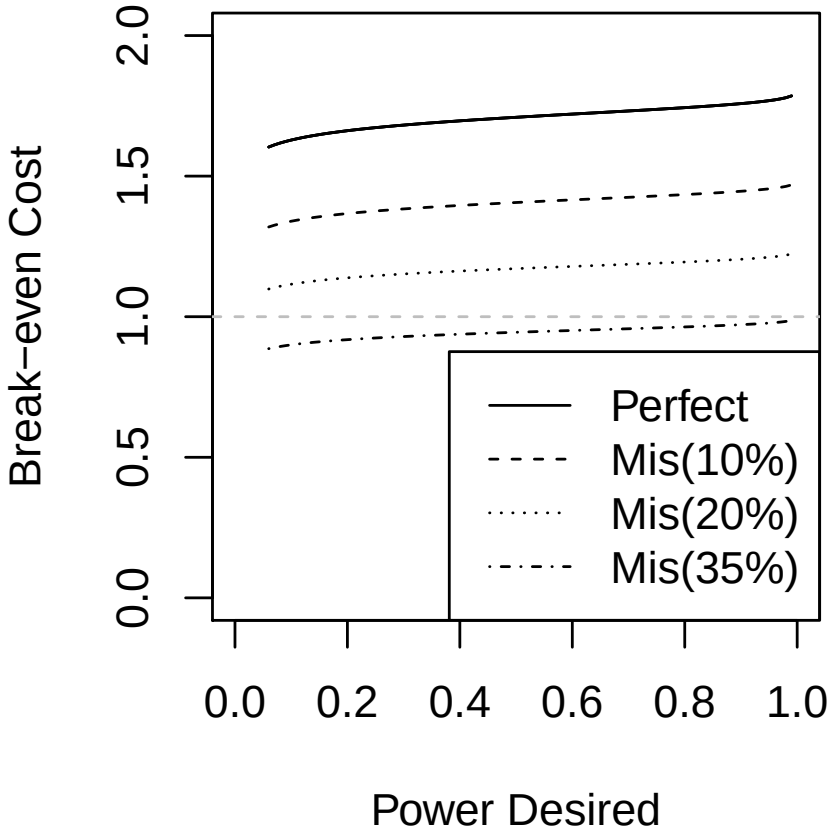
Counter intuitive?

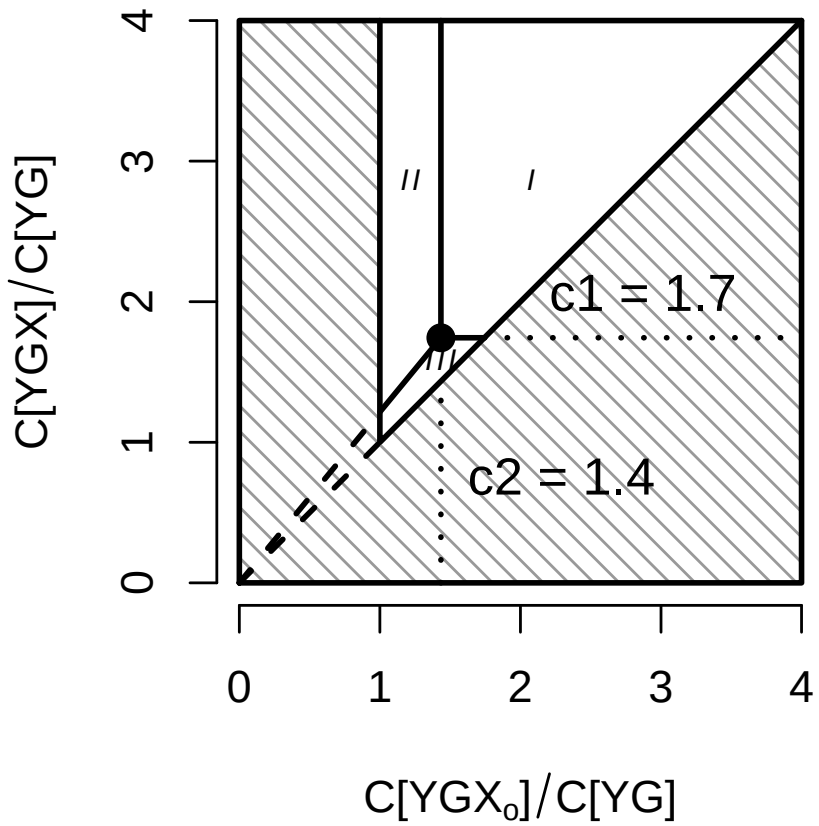
(Y, G, X_o) data

$$\text{logit} Pr(Y = 1|X_o, G) = \beta_0^* + \beta_1^* X_o + \beta_2^* G + \beta_3^* X_o G.$$

When we treat X_o as if it were X, fitting the full model to (Y, G, X_o) data generally gives biased point estimates of the coefficients of interest. However, $(\beta_2, \beta_3) = (0, 0)$ implies $(\beta_2^*, \beta_3^*) = (0, 0)$, so that fitting a model to (Y, G, X_o) data still yields a valid test of the original null hypothesis. However, the use of X_o rather than X will reduce power, and the power calculation should be adjusted for the presence of misclassification.

Misclassification





Conclusion

Under a wide range of circumstances research resources aimed at identification of association between genes and diseases can be more efficiently (in a sense of study power) allocated to genotyping larger groups of individuals rather than investing in exposure assessment, when exposure and genes interact.

We only evaluate our methods with "qualitative" interactions. Both two tests are still valid when there could be a main gene effect (i.e. $\beta_2 \neq 0$).

When we evaluate power in the presence of a main gene effect, we find that the break-even cost decreases with the magnitude of the main effect, whilst other parameters remain fixed.

In fact, when the main gene effect effect comes to dominate the other effects in the model, the marginal gene effect can be easily detected, and the exposure measurements can become harmful for the same reason stated earlier.

Case-Control

While we have presented our results in terms of cohort studies, our power calculations will still be valid in the context of case-control studies, with the proviso that the relevant intercept β_0 would be that induced by the case-control sampling scheme rather than that describing rarity of the disease in the study population.

It should be recognized that these tests may be applied to thousands or millions screened markers to uncover any susceptibility loci. This leads to two problems

- The choice of type I error.
- Different break-even costs.

Thank You!