Variance component analysis of Environmental Health Data

Xuelong Wang

April 10, 2019

1. Background
2. Goal
3. Solution: GCTA method
4. Result
5. Stardardized covaraites: unexpected problem

Background

To understand the effects of the environment (chemical mixtures) on human health.

Figure 1: A complex real world research challenge

Challenge

- lack of traditional epidemiology methodology, e.g. the pathway is not clear
- Many weak signals, hard to identify and select, e.g lasso type is not working

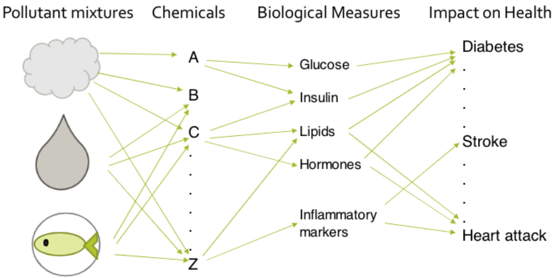

Figure 2:

Environmental Data

Data

- Covariates are concentration of environmental mixtures, e.g. heavy metal, PCBs
  - Continuous
  - The number of predictors are around 30 to 100
  - There are high correlations among those covariates
  - Magnitude levels are very low
- Response are health outcomes, e.g. blood pressure, disease status,etc.

Goal

- Evaluate the relation between the environmental mixture and health outcomes
- More specifically, the variance $Var(X^T\beta)$

What is the GCTA method

- GCTA: Genome-wide complex trait analysis
- GCTA estimates the variance of y related to the covariates.

a working linear mixed effects model

$$Y_i = \mu + \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i \tag{1}$$

$$Y_i = \mu + \sum_{j=1}^{p} X_{ij}\beta_j + \sum_{0\le l<k\le p} \gamma_{lk} X_{il} X_{ik} + \epsilon_i \tag{2}$$

Limitations

Assumption

Covariates have to be independent to each other

Real world

Each covariates are more likely to be correlated to each other

Decorrelation

The linear transformation is

$$\tilde{X} = A^{-1}X,$$

where X are the covariates vector,A is a linear transformation operator which is a full rank square matrix. After transformation, the covariance of the new covariates $\tilde{X}$ will be

$$Var(\tilde{X}) = I_p.$$

Moreover, based on the model from last slide, we have

$$Y = \mu + X^T\beta + \epsilon = \tilde{X}^T A^T\beta + \epsilon = \tilde{X}^T\alpha + \epsilon,$$

where $\alpha = A^T\beta$. Let's look the total effect of X and Z:

$$Var(X^T\beta) = Var(\tilde{X}^T A^T\beta) = Var(\tilde{X}^T\alpha).$$

Simulation result

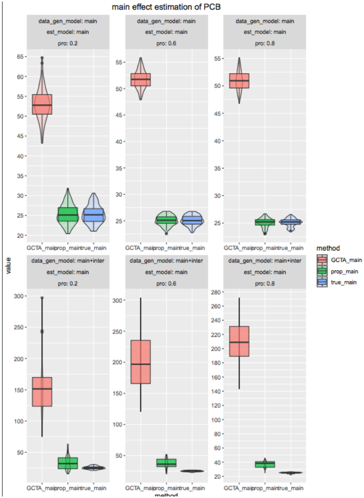

Figure 3: Simulation result

Why X is standardized

- The benefit we get from standardizing data is computational efficiency.
- The the columns of standardized data are in the same scale, so there is no too large or too small values, which may cause computational issues, i.e. rounding.

What is changed if use standardized

$$Z_k = \frac{X_k - \mu_k}{\sigma_k} \implies X_k = \sigma_k Z_k + \mu_k$$

$$\begin{aligned} Y &= \mu + \sum_{k=1}^{p} (\sigma_k Z_k + \mu_k)\beta_k + \epsilon \\ &= \mu + \sum_{k=1}^{p} (\mu_k + \beta_k) + \sum_{k=1}^{p} (Z_k \sigma_k \beta_k) + \epsilon. \end{aligned}$$

By the property of variance, we have

$$Var(\sum_{k=1}^{p} X_k \beta_k) = Var(\sum_{k=1}^{p} Z_k \sigma_k \beta_k).$$

same for the interaction?

$$\begin{aligned}
\sum_{0\le l<k\le p} \gamma_{lk} X_l X_k &= \sum_{0\le l<k\le p} \gamma_{lk}(\sigma_l Z_l + \mu_l)(\sigma_k Z_k + \mu_k) \\
&= \sum_{0\le l<k\le p} (\gamma_{lk}\sigma_l\sigma_k Z_l Z_k) + \sum_{0\le l<k\le p} (\gamma_{lk}\sigma_l Z_l \mu_k) \\
&\quad + \sum_{0\le l<k\le p} (\gamma_{lk}\sigma_k Z_k \mu_l) + \mu^*.
\end{aligned}$$

- $Var(\sum_{k=1}^{p}(Z_k\beta_k^*)) \neq Var(\sum_{k=1}^{p}(X_k\beta_k))$
- $Var(\sum_{0\le l<k\le p}(\gamma_{lk}^* Z_l Z_k)) \neq Var(\sum_{0\le l<k\le p}\gamma_{lk} X_l X_k)$

Solution

Total effect

$$Var\left(\sum_{j=1}^{p} X_j\beta_j + \sum_{0\le l<k\le p} \gamma_{lk}X_lX_k\right) =$$
$$Var\left(\sum_{k=1}^{p}(Z_k\beta_k^*) + \sum_{0\le l<k\le p}(\gamma_{lk}^*Z_lZ_k)\right)$$

Future work

- Separate the main and interaction effects
- Statistical test on the interaction effects