Modeling Phenotype Products through Pre-Computed Summary Statistics

Jack M. Wolf (he/him)

Division of Biostatistics, University of Minnesota School of Public Health

IGES 30th Annual Meeting, General Session #3
October 15, 2021

Acknowledgements

This work was made possible by the generous support of the NIH through grant 2R15HG006915-03

I would also like to thank the wonderful collaborators who have contributed to this research:

- Dr. Nathan Tintle
- Jason Westra
- Martha Barnrd

Table of Contents

1. Introduction

2. Methods

3. Results

4. Discussion

Introduction

A Question

What do we need to consider when we work with large biobank data?

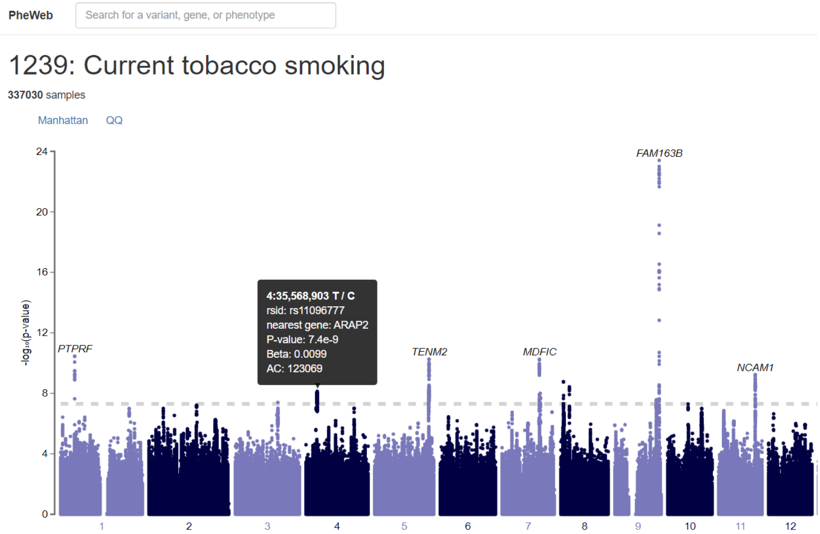
Introduction

A Question

What do we need to consider when we work with large biobank data?

- Data privacy and security
- Data access and availability
- Computational costs

Introduction



1239: Current tobacco smoking

337030 samples

4:35,568,903 T / C
rsid: rs11096777
nearest gene: ARAP2
P-value: 7.4e-9
Beta: 0.0099
AC: 123069

Introduction

Key Idea

How can we leverage pre-computed summary statistics (PCSS) from biobanks to estimate statistical models fit using individual participant data (IPD)?

Existing Methods:

- Multi-trait association tests (Ray & Boehnke, 2018; Dutta et al., 2019; Guo & Wu, 2019)
- Linear combinations of phenotypes (Gasdaska et al., 2019; Wolf et al., 2020)

Introduction

Goal

Approximate linear models for products of phenotypes of the form:

$$\prod_{k=1}^{m} \boldsymbol{y}_k = \boldsymbol{X}\beta + \epsilon$$

using PCSS with flexible choice of covariates.

Introduction

Goal

Approximate linear models for products of phenotypes of the form:

$$\prod_{k=1}^{m} \boldsymbol{y}_k = \boldsymbol{X}\beta + \epsilon$$

using PCSS with flexible choice of covariates.

Why Products?

- Ratios of phenotypes
- Logical combinations of phenotypes

Introduction

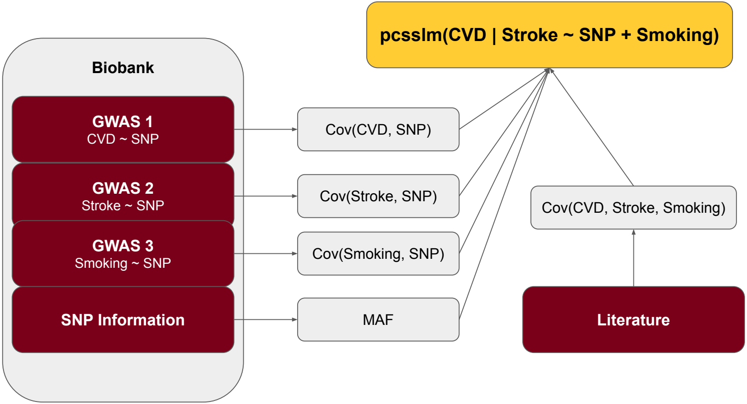


Table of Contents

1. Introduction
2. **Methods**
3. Results
4. Discussion

Assumed PCSS

$$\underbrace{\begin{bmatrix} \sigma_{\boldsymbol{x}_1,\boldsymbol{x}_1} & \sigma_{\boldsymbol{x}_1,\boldsymbol{x}_2} & \cdots & \sigma_{\boldsymbol{x}_1,\boldsymbol{x}_p} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \vdots \\ & & & \sigma_{\boldsymbol{x}_p,\boldsymbol{x}_p} \end{bmatrix}}_{p\times p} \qquad \underbrace{\begin{bmatrix} \sigma_{\boldsymbol{x}_1,\boldsymbol{y}_1} & \sigma_{\boldsymbol{x}_1,\boldsymbol{y}_2} & \cdots & \sigma_{\boldsymbol{x}_1,\boldsymbol{y}_m} \\ \sigma_{\boldsymbol{x_2},\boldsymbol{y_1}} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_{\boldsymbol{x_p},\boldsymbol{y_m}} & \cdots & \cdots & \sigma_{\boldsymbol{x}_p,\boldsymbol{y}_m} \end{bmatrix}}_{p\times m}$$

$$\underbrace{\begin{bmatrix} \bar{x}_1 \bar{x}_2 \cdots \bar{x}_p \end{bmatrix}}_{1\times p}$$

$$\underbrace{\begin{bmatrix} \bar{y}_1 \bar{y}_2 \cdots \bar{y}_m \end{bmatrix}}_{1\times m}$$

$$\underbrace{\begin{bmatrix} \sigma_{\boldsymbol{y}_1,\boldsymbol{y}_1} & \sigma_{\boldsymbol{y}_1,\boldsymbol{y}_2} & \cdots & \sigma_{\boldsymbol{y}_1,\boldsymbol{y}_m} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \vdots \\ & & & \sigma_{\boldsymbol{y}_m,\boldsymbol{y}_m} \end{bmatrix}}_{m\times m}$$

Regression with PCSS

Theorem

For the regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$*, with* $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$*, the ordinary least squares estimate for* $\boldsymbol{\beta}$ *is*

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

This can be computed via PCSS using the facts that:

$$\boldsymbol{X}'\boldsymbol{X} = (n-1)S(\boldsymbol{X}) + n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}' \quad (1)$$

$$\boldsymbol{X}'\boldsymbol{y} = (n-1)(s_{y,x_1}, \ldots, s_{y,x_p})' + n\bar{y}\bar{\boldsymbol{x}} \quad (2)$$

Regression with PCSS

Theorem

*The estimated variance of $\hat{\beta}$ is**

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \hat{\sigma}^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

This can be calculated via PCSS using previous equalities and the fact that:

$$\hat{\sigma}^2 = [(n-1)s_y^2 + n\bar{y}^2 - \hat{\beta}'\boldsymbol{X}'\boldsymbol{y}]/(n-p) \tag{3}$$

Modeling Phenotype Products

To approximate the covariance between $\boldsymbol{x}_j$ and the product $\boldsymbol{w} = \boldsymbol{y}_1 \boldsymbol{y}_2$ we estimate the conditional mean of $\boldsymbol{w}$ given $\boldsymbol{x}_j$ as

$$g(w|x) = g(y_1|x)g(y_2|x) + h(y_1, y_2|x), \tag{4}$$

which gives the covariance estimate

$$s_{x_j,w} \approx \sum_{x \in \mathcal{S}_j} f_j(x)(x - \bar{x}_j)g(w|x) \tag{5}$$

Table of Contents

1. Introduction

2. Methods

3. **Results**

4. Discussion

Simulation Studies

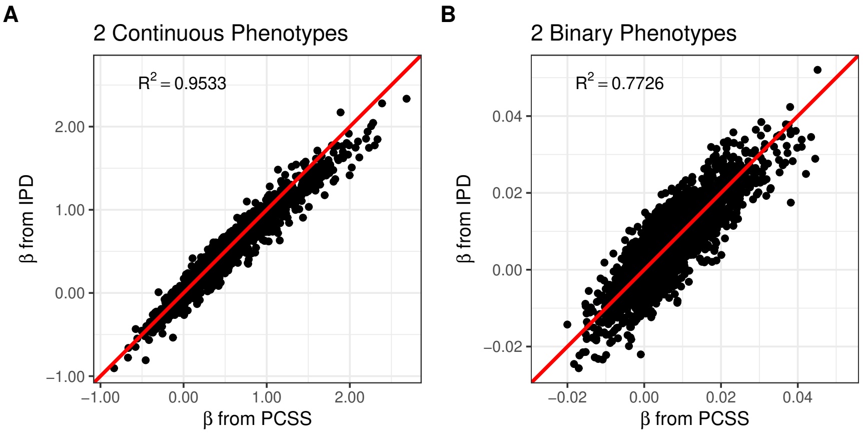
We generated data through the model:

$$u(y_{ik}) = \beta_{k0} + \sum_{j=1}^{3} x_{ij}\beta_{kj} + \epsilon_{ik}$$

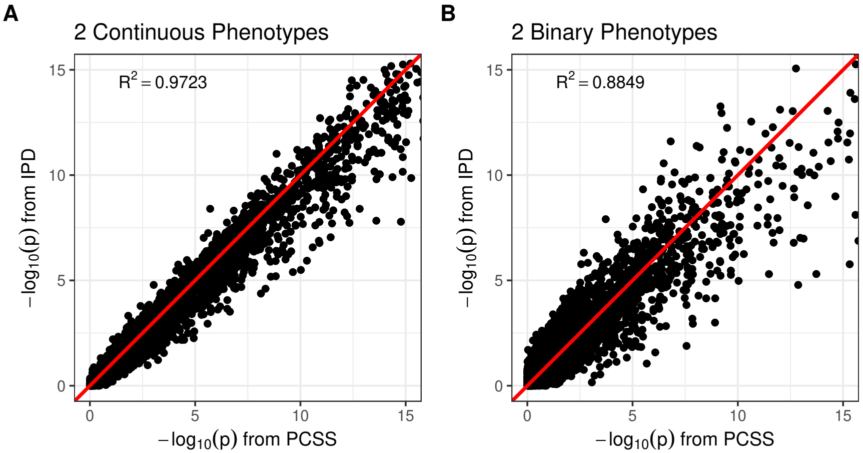
where

- $u(y_{ik}) = y_{ik}$ or $\text{logit}(\Pr(Y_{ik} = 1))$
- $\boldsymbol{x}_1 =$ SNP's minor allele counts
- $\boldsymbol{x}_2 =$ continuous covariate
- $\boldsymbol{x}_3 =$ binary covariate

Simulation Study Estimating β



Simulation Study Estimating p-values



Real Data Analysis

Fatty acids and conversion ratios

- Fatty acids are biomarkers of various cardiometabolic and cognitive health outcomes
- Conversion ratios illustrate how fatty acids are converted from one fatty acid to the next

Real Data Analysis

Framingham Heart Study (Mailman et al., 2007)

- 12 fatty acid conversion ratios
- 362,330 SNPs
- 4,347,960 models: FA Ratio $\sim$ SNP $+$ age $+$ sex

Real Data Analysis

Framingham Heart Study (Mailman et al., 2007)

- 12 fatty acid conversion ratios
- 362,330 SNPs
- 4,347,960 models: FA Ratio $\sim$ SNP $+$ age $+$ sex
- Disagreement rate of $10/(4.3 \times 10^6)$
- Of the 10 disagreements:
  - 4 where PCSS failed to reject when IPD rejected H_0,
  - 6 where PCSS rejected when IPD failed to reject

Table of Contents

1. Introduction
2. Methods
3. Results
4. **Discussion**

Discussion

Takeaway

We can approximate linear models for products and logical combinations of phenotypes with a **flexible choice of covariates** using only readily available pre-computed summary statistics.

Discussion

Limitations and Future Work

- Assessing the compounding of errors when modeling the product of ≥ 4 phenotypes
- Measuring sensitivity to missing data and other assumption violations
- Accounting for related individuals through kinship matrices

Thank you!

Slides:	`http://bit.ly/???`
R Package:	`pcsstools`
Twitter:	`@_jackmwolf`
Email:	`WolfX681@umn.edu`

References

Dutta, D., Scott, L., Boehnke, M., & Lee, S. (2019). Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes. *Genetic Epidemiology*, *43*(1), 4–23.
URL `http://doi.wiley.com/10.1002/gepi.22156`

Gasdaska, A., Friend, D., Chen, R., Westra, J., Zawistowski, M., Lindsey, W., & Tintle, N. (2019). Leveraging summary statistics to make inferences about complex phenotypes in large biobanks. *Pacific Symposium on Biocomputing*, *24*, 391–402.
URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6417828/`

Guo, B., & Wu, B. (2019). Integrate multiple traits to detect novel trait–gene association using GWAS summary data with an adaptive test approach. *Bioinformatics*, *35*(13), 2251–2257.
URL `https://academic.oup.com/bioinformatics/article/35/13/2251/5201342`

Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L., Popova, N., Pretel, S., Ziyabari, L., Lee, M., Shao, Y., Wang, Z. Y., Sirotkin, K., Ward, M., Kholodov, M., Zbicz, K., Beck, J., Kimelman, M., Shevelev, S., Preuss, D., Yaschenko, E., Graeff, A., Ostell, J., & Sherry, S. T. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, *39*(10), 1181–1186.

Ray, D., & Boehnke, M. (2018). Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genetic Epidemiology*, *42*(2), 134–145.
URL `http://doi.wiley.com/10.1002/gepi.22105`

Wolf, J. M., Barnard, M., Xia, X., Ryder, N., Westra, J., & Tintle, N. (2020). Computationally efficient, exact, covariate-adjusted genetic principal component analysis by leveraging individual marker summary statistics from large biobanks. *Pacific Symposium on Biocomputing*, *25*, 719–730.
URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6907735/`

Wolf, J. M., & R Core Team and contributors worldwide (2021). *pcsstools: Tools for Regression Using Pre-Computed Summary Statistics*.
URL `https://CRAN.R-project.org/package=pcsstools`