ADNI1 Dataset and Model

Hao Wu

September 26, 2022

1 How to estimate time-varying coefficient β

1.1 Observations in ADNI1

There are 819 patients and 5122 longitudinal observations when I use MMSE as response variable.

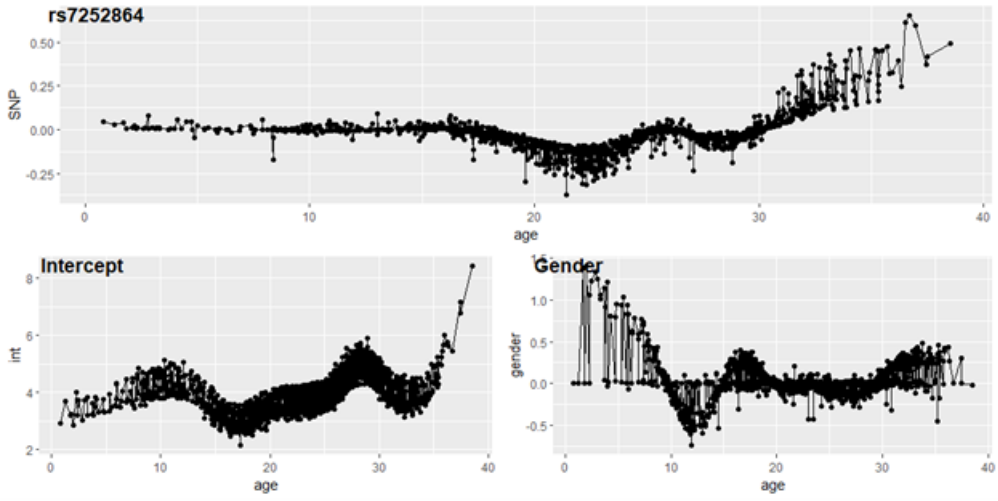
value	n
1	34
2	44
3	54
4	64
5	34
6	74

1.2 How to Set Time Points

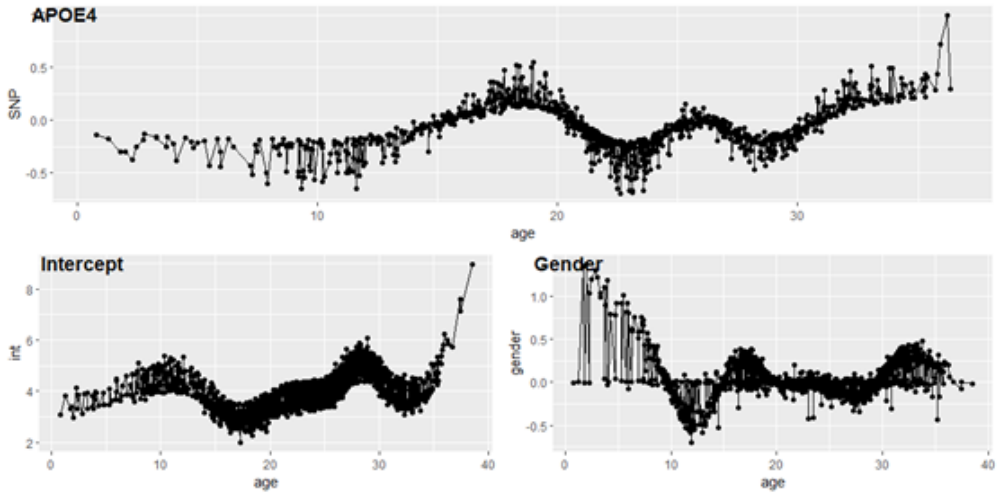
- Treat the smallest age of patients in ADNI1 as 0.
- Treat the age at baseline of each patient as its original age, then add the date difference over 365 to the original age as its new age.

1.3 Time-Varying Coefficient Plots for 2 SNPs

Refer to Chu's **(author?)** [1]idea, we generate our time-varying coefficient plot.



(a) rs7252864



(b) APOE4

2 Construct Models

2.1 Vannucci's Paper Idea

Using the model mentioned in Stingo's [2] paper,

- Y: an n × 1 outcome vector indicating the subjects' phenotype.
- X, an n × p matrix of genotypes.

Let $T(n \times K)$ be the matrix of gene-level summary measures of SNP measurements,

$$y_i = \alpha + \sum_{k=1}^{K} T_{ik}\beta_k + \epsilon_i, \epsilon_i \sim N(0, \sigma^2) \tag{2.1}$$

For gene k, we construct an $n \times 1$ vector T_k of scores calculated based on the vectors X_i of SNP genotypes belonging to gene k. p_k, the number of SNPs in gene k,

$$T_{ik} = \sum_{j=1}^{p_k} w_{ij} X_{ij} \tag{2.2}$$

π, a constant between 0 and 1 determining the influence of the Hardy-Weinberg frequencies on the gene scores. f_{ij}, the expected population genotype frequencies computed according to the Hardy-Weinberg law

$$\tilde{w_{ij}} = \pi \frac{1}{f_{ij}} + (1-\pi)\frac{1}{p_k}, w_{ij} = \frac{\tilde{w_{ij}}}{\sum_{j=1}^{p_k} w_{ij}^{\sim}} \tag{2.3}$$

2.2 One Idea: Measurement Error Model

Consider the model defined by

$$y_i(t) = \beta_0 + \sum_{k=1}^{K} T_{ik}^{*}\beta_k(t) + e_i \tag{2.4}$$

$$T_{ik} = T_{ik}^{*} + u_i \tag{2.5}$$

$$\begin{bmatrix} T_{ik}^{*} \\ e_i \\ u_i \end{bmatrix} \sim N(\begin{bmatrix} \mu_{T_{ik}^{*}} \\ e_i \\ u_i \end{bmatrix}, \begin{bmatrix} \sigma_{T_{ik}^{*}} & \sigma_{Te} & \sigma_{Tu} \\ \sigma_{Te} & \sigma_{ee} & \sigma_{eu} \\ \sigma_{Tu} & \sigma_{eu} & \sigma_{uu} \end{bmatrix}) \tag{2.6}$$

where $i = 1, 2, ..., n$,

- $y_i(t)$ is our response variable, such as MMSE;
- T_{ik} is the observed measurement of T_{ik}^*

$$T_{ik} = \sum_{j=1}^{p_k} w_{ij} X_{ij}, \tilde{w}_{ij} = \pi \frac{1}{f_{ij}} + (1-\pi)\frac{1}{p_k}, w_{ij} = \frac{\tilde{w}_{ij}}{\sum_{j=1}^{p_k} w_{ij}^{\sim}} \tag{2.7}$$

- T_{ik}^* are the true gene-level summary measures of SNP measurements;
- e_i are independent $N(0, \sigma_{ee}^2)$ and potentially be a combination of model and measurement error;
- u_i is a $N(0, \sigma_{uu}^2)$ random variable;
- $\pi = 0.5$ as a default value.

2.3 Toy Longitudinal Results using Time-varying Coefficient Method

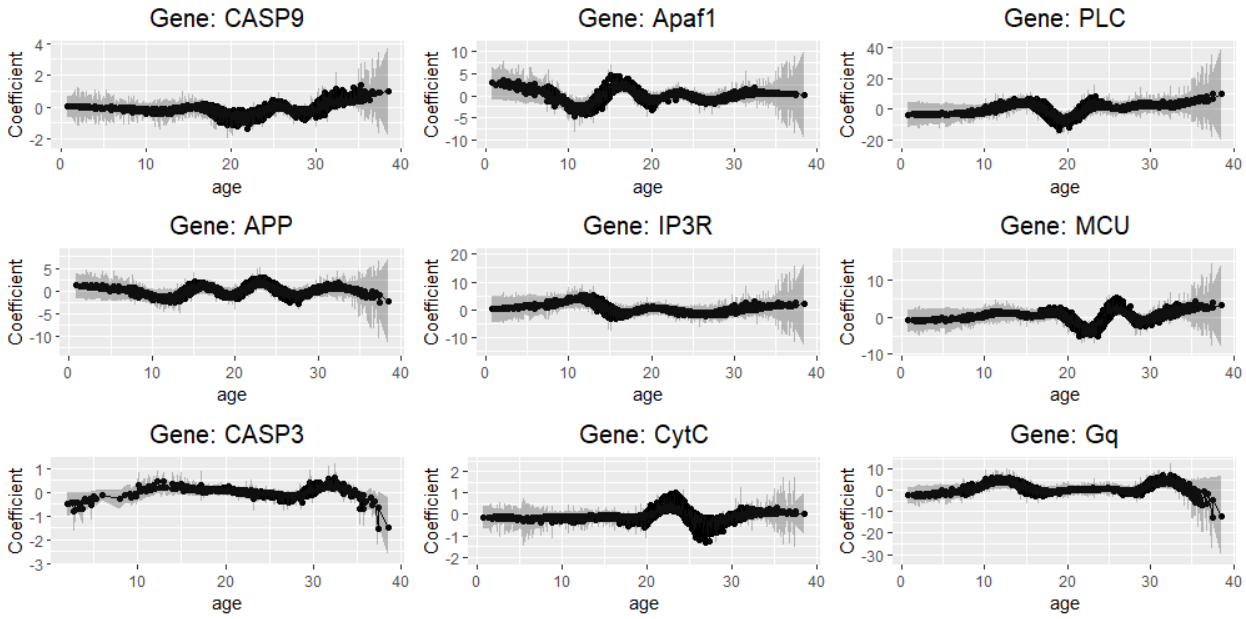
Refer to Chu's **(author?)** [1] idea, we generate our time-varying coefficient.

In our ADNI1 dataset,

- hasing y (MMSE): 819 patients and 5122 longitudinal observations;
- satisfying code requirements for y($4 \leq n \leq 6$): 399 patients and 1935 longitudinal observations
- hasing X (SNP value): 757 patients. Among these patients, there are 14667 missing data out of 528386 (757*698) SNPs when we use Pathway N01002 (including 698 SNPs in ADNI1).
- Combine y and X: 370 patients and 1792 longitudinal observations.

In addition, there are some special transformations,

- treat missing SNP value as 0, and the corresponding f_{ij} also set as 0;
- to avoid the infinity value of w_{ij}, reset all $f_{ij} = 0$ as $f_{ij} = 100000$.


Gene: CASP9
Gene: Apaf1
Gene: PLC
Gene: APP
Gene: IP3R
Gene: MCU
Gene: CASP3
Gene: CytC
Gene: Gq
Coefficient
age

References

[1] Wanghuan Chu, Runze Li, and Matthew Reimherr. Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data. *The annals of applied statistics*, 10(2):596, 2016.

[2] Francesco C Stingo, Michael D Swartz, and Marina Vannucci. A bayesian approach to identify genes and gene-level snp aggregates in a genetic analysis of cancer data. *Statistics and its Interface*, 8(2):137, 2015.