Association Tests for Genetic Effect and Its Interaction with Environmental Factors

Zhengyang Zhou
Southern Methodist University, zhengyangz@smu.edu

Follow this and additional works at: https://scholar.smu.edu/hum_sci_statisticalscience_etds

Part of the Biostatistics Commons, Genetic Phenomena Commons, and the Statistical Models Commons

Recommended Citation
https://scholar.smu.edu/hum_sci_statisticalscience_etds/3

This Dissertation is brought to you for free and open access by the Statistical Science at SMU Scholar. It has been accepted for inclusion in Statistical Science Theses and Dissertations by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.
ASSOCIATION TESTS FOR GENETIC EFFECT AND
ITS INTERACTION WITH ENVIRONMENTAL FACTORS

Approved by:

_______________________________________
Dr. Chao Xing
Associate Professor of Statistical Genetics

_______________________________________
Dr. Daniel Heitjan
Professor of Biostatistics

_______________________________________
Dr. Hon Keung Tony Ng
Professor of Statistics

_______________________________________
Dr. Vinod Mootha
Professor of Ophthalmology
ASSOCIATION TESTS FOR GENETIC EFFECT AND 
ITS INTERACTION WITH ENVIRONMENTAL FACTORS

A Dissertation Presented to the Graduate Faculty of
Dedman College of Humanities and Sciences
Southern Methodist University
in
Partial Fulfillment of the Requirements
for the degree of
Doctor of Philosophy
with a
Major in Biostatistics
by
Zhengyang Zhou
B.S., Beijing Institute of Technology, China, 2014
M.S., Southern Methodist University, 2017

August 7, 2018
ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Dr. Chao Xing, for his fundamental role in my doctoral work. It has been a great honor for me to be his first Ph.D. student. Dr. Xing has provided me with every bit of valuable guidance, insightful ideas and consistent encouragement throughout my research work. Whenever I was trapped in research, Dr. Xing has always made himself available to navigate me to the right direction despite his busy schedules. His passion and enthusiasm in research inspired my desire to pursue a career at academia, which I could never imagine before I joined his lab. I am also thankful for the example he has provided as an excellent researcher and advisor.

I would like to thank the rest of my thesis committee members Dr. Daniel Heitjan, Dr. Hon Keung Tony Ng and Dr. Vinod Mootha, for their valuable comments and advice for my dissertation thesis and unreserved help during my job search. I am also so grateful to all the other faculty members in the Department of Statistical Science at SMU for their excellent education and help throughout my Ph.D. study. They brought their profound understanding in statistics to me and raised my interest in learning statistics.

I would like to thank my friends Dateng Li, Dr. Zhiyun Ge, Dr. Gaoxiang Jia, Shen Yin and others at SMU, who have not only helped me a lot in my coursework, but also made my Ph.D. life much more enjoyable.
I especially thank my parents for their unconditional love, understanding and support. They are always there for me when I need support, and I feel safe and warm because of them. They are always reassuring when I undergo hard times, and they can always convey their strength to me. They are the ones who encouraged me to study aboard and pursue a Ph.D. degree, and I am so thankful for their wise advice. I love them so much and I would not have made it without them. I would also like to thank my wife, Yan Xu, as well as her wonderful family for their support and love. I truly thank Yan for always understanding me and standing by my side even when I was irritable under the pressure of study and research.
My research is in the area of statistical genetics, and it contains three projects: (1) Differentiating the Cochran-Armitage (CA) trend test and Pearson’s $\chi^2$ test: location and dispersion; (2) Decomposing Pearson’s $\chi^2$ test: a linear regression and its departure from linearity; (3) Testing nonlinear gene-environment ($G \times E$) interaction through varying coefficient and linear mixed models.

(1) In genetic case-control association studies, a standard practice is to perform the CA trend test with 1 degree-of-freedom (df) under the assumption of an additive model. However, when the true genetic model is recessive or near recessive, it is outperformed by Pearson’s $\chi^2$ test with 2 df. In this project we analytically reveal the statistical basis that leads to the phenomenon. First, we show that the CA trend test examines the location shift between the case and control groups, whereas Pearson’s $\chi^2$ test examines both the location and dispersion shifts between the two groups. Second, we show that under the additive model the effect of location deviation outweighs that of the dispersion deviation, and vice versa under a near recessive model. Therefore, Pearson’s $\chi^2$ test is a more robust test than the CA trend test and it outperforms the latter when the mode of inheritance evolves to the recessive end.
(2) In genetic case-control association studies, we could identify situations CA trend test outperformed the analysis model consistent with the underlying inheritance mode. In this project we analytically reveal the statistical basis that leads to the phenomenon. By elucidating the origin of the CA trend test as a linear regression model, we decompose Pearson’s $\chi^2$ test statistic into two components—one is the CA trend test statistic that measures the goodness-of-fit of the linear regression model, the other measures the discrepancy between the data and linear regression model. Under this framework we show the additive coding scheme, as well as the multiplicative coding scheme, increases the coefficient of determination of the regression model by increasing the spread of data points. We also obtain the conditions under which the CA trend test statistic equals the MAX statistic and Pearson’s $\chi^2$ test statistic.

(3) We present a novel statistical procedure to detect the nonlinear $G \times E$ interaction with continuous traits in sequencing association studies. Commonly-used approaches for $G \times E$ interaction usually assume linear relationship between genetic and environmental factor, thus they suffer power loss when the underlying relationship is nonlinear. Varying coefficient model is proposed to relax the linear assumption, however, it’s unable to adjust for population stratification, a major source of confounding in genome-wide association studies. To overcome these limitations, we develop the Varying-Coefficient embedded Linear Mixed Model (VC-LMM) for assessing the nonlinear $G \times E$ interaction and accounting for population stratification. The proposed VC-LMM well controls type I error rates when the population stratification is present, and it’s powerful for both common and low frequency variants. We apply computationally efficient algorithms for generating null distributions and estimating parameters in the linear mixed model, thus the computational burden is greatly reduced. Using simulation studies, we demonstrate the performance of VC-LMM.
# TABLE OF CONTENTS

LIST OF FIGURES .................................................................................................................. x

LIST OF TABLES ..................................................................................................................... xi

CHAPTER

1 DIFFERENTIATING THE COCHRAN-ARMITAGE TREND TEST AND PEARSON’S $\chi^2$ TEST: LOCATION AND DISPERSIONISTLIST .................................................................................. 1

1.1 Introduction ..................................................................................................................... 1

1.2 Method .......................................................................................................................... 2

1.3 Simulation studies ......................................................................................................... 5

1.4 Discussion ..................................................................................................................... 11

2 DECOMPOSING PEARSON’S $\chi^2$ TEST: A LINEAR REGRESSION AND ITS DEPARTURE FROM LINEARITY ............................................................................. 13

2.1 Introduction ..................................................................................................................... 13

2.2 Simulation studies ......................................................................................................... 15

2.3 Method .......................................................................................................................... 17

2.4 Discussion ..................................................................................................................... 22

3 TESTING NONLINEAR GENE-ENVIRONMENT INTERACTION THROUGH VARYING COEFFICIENT AND LINEAR MIXED MODELS .......................................... 25

3.1 Introduction ..................................................................................................................... 25

3.2 Method .......................................................................................................................... 27

3.3 Simulation studies ......................................................................................................... 31

3.3.1 Type I error rates ........................................................................................................ 33

3.3.2 Power ........................................................................................................................ 34

3.4 Discussion ..................................................................................................................... 36
APPENDIX

A DIFFERENTIATING THE COCHRAN-ARMITAGE TREND TEST AND PEARSON’S $\chi^2$ TEST: LOCATION AND DISPERSION LIST ................................................................. 37

A.1 Equivalence between the CA trend test statistic $T_{CA}$ and the two-sample mean test statistic $T_{mean}$ ........................................................................................................ 37

A.2 Partitioning $T_P$ by orthogonal polynomials .................................................. 38

B DECOMPOSING PEARSON’S $\chi^2$ TEST: A LINEAR REGRESSION AND ITS DEPARTURE FROM LINEARITY .................................................................................... 39

B.1 Proof of $T_P = T_{CA} + T_{SSE}$ ........................................................................ 39

B.2 Figure: power comparison of the Cochran-Armitage trend test by the additive, multiplicative, and recessive models under different modes of inheritance Simulation studies ........................................................................................................ 40

C TESTING NONLINEAR GENE-ENVIRONMENT INTERACTION THROUGH VARYING COEFFICIENT AND LINEAR MIXED MODELS .................................................. 41

C.1 Algorithm for estimation of parameters ............................................................ 41

REFERENCE ........................................................................................................ 44
LIST OF FIGURES

Figure 1.1 Power comparison of the CA trend test ($T_{CA}$) and Pearson’s $\chi^2$ test ($T_P$) .................. 7

Figure 1.2 Power to detect the location and dispersion shifts of cell counts between cases and controls under additive and near recessive genetic models ........................................... 9

Figure 2.1 Power comparison of the Cochran-Armitage trend test by the additive, multiplicative, and recessive models under recessive modes of inheritance ........................................... 14

Figure 2.2 Comparison of the Cochran-Armitage trend test by different analysis models under the recessive mode of inheritance ............................................................................. 20

Figure 3.1 Power of VC-LMM, LMi and C-LRT for common variants ............................................. 35

Figure 3.2 Power of VC-LMM, LMi and C-LRT for low frequency variants .............................. 35

Figure B.2 Power comparison of the Cochran-Armitage trend test by the additive, multiplicative, and recessive models under different modes of inheritance ........................................... 40
LIST OF TABLES

Table 1  Genotype distribution at a diallelic marker in a case-control study ........................................ 2

Table 2  Type I error rates across VC-LMM, LMi and C-LRT ................................................................. 33
This is dedicated to my family.
CHAPTER 1

DIFFERENTIATING THE COCHRAN-ARMITAGE TREND TEST AND PEARSON’S $\chi^2$ TEST: LOCATION AND DISPERSION

1.1 Introduction

In genetic association studies, in particular for genome-wide screens, the Cochran-Armitage (CA) trend test (Armitage, 1955; W. G. Cochran, 1954) under the assumption of an additive genetic model became the standard practice following Sasieni’s seminal paper (Sasieni, 1997). However, this approach suffers power loss when the true genetic model is non-additive. In contrast, Pearson’s $\chi^2$ test (Pearson, 1900) with 2 degree-of-freedom (df), which is an omnibus test without regard to genetic models, is robust to any underlying models. It outperforms the CA trend test when the mode of inheritance is recessive or near recessive, which was shown by numeric simulation studies (e.g., Gonzalez et al., 2008; Kuo & Feingold, 2010; Li, Zheng, Liang, & Yu, 2009; Loley, Konig, Hothorn, & Ziegler, 2013). A convenient explanation of the phenomenon is that the incorrect model assumption results in significant power loss, which is true yet futile. Here we aim to elucidate the underlying statistical cause that leads to the different performance of the two tests.

Consider a diallelic locus with the major and minor alleles denoted as $\alpha$ and $A$, respectively, in a case-control study (Table 1). Denote by $r_i$ and $s_i$ the numbers of cases and
controls, respectively, in genotype category $G_i$, where $i \in \{0,1,2\}$ reflects the number of $A$ alleles a subject has. Thus $G_0$, $G_1$, and $G_2$ correspond to genotypes $aa$, $Aa$, and $AA$, respectively. Denote by $R$, $S$, and $n_i$ the marginal sums such that $R = \sum_{i=0}^{2} r_i$, $S = \sum_{i=0}^{2} s_i$, and $n_i = r_i + s_i$, and denote by $N$ the total sample size such that $N = R + S = \sum_{i=0}^{2} n_i$. Assume $(r_0, r_1, r_2)$ follow a trinomial distribution with parameters $R$ and $(p_0, p_1, p_2)$, and $(s_0, s_1, s_2)$ follow a trinomial distribution with parameters $S$ and $(q_0, q_1, q_2)$. The null hypothesis of no association between the disease and genotype is $H_0: p_i = q_i$.

Table 1: Genotype distribution at a diallelic marker in a case-control study

<table>
<thead>
<tr>
<th>Phenotype</th>
<th>Genotype</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$aa$</td>
<td>$Aa$</td>
</tr>
<tr>
<td>Cases</td>
<td>$r_0$</td>
<td>$r_1$</td>
</tr>
<tr>
<td>Controls</td>
<td>$s_0$</td>
<td>$s_1$</td>
</tr>
<tr>
<td>Total</td>
<td>$n_0$</td>
<td>$n_1$</td>
</tr>
</tbody>
</table>

1.2 Method

Assign a set of scores $(x_0, x_1, x_2)$ to the three genotypes $aa$, $Aa$ and $AA$, respectively, with the constraints $x_0 \leq x_1 \leq x_2$ and $x_0 < x_2$. The CA trend test statistic is

$$T_{CA} = \frac{R S}{N^2 \left[ N \sum_{i=0}^{2} x_i^2 n_i - (\sum_{i=0}^{2} x_i n_i)^2 \right]^2} \left[ \frac{1}{N} \sum_{i=0}^{2} x_i (Sr_i - Rs_i) \right]^2$$

(1)

Under the null hypothesis, $T_{CA}$ follows a $\chi^2$ distribution with 1 df.
The CA trend test statistic is identical to that of a test for the difference of the average scores between the cases and controls. The derivation details are described in Appendix A.1, and here we summarize the results. Denote by $X_j$ the score of the $j$-th subject ($j = 1, \ldots, R$) in the case group and by $Y_k$ the score of the $k$-th subject ($k = 1, \ldots, S$) in the control group. Then $\bar{X} = \frac{\sum_{i=0}^{R} x_i r_i}{R}$ and $\bar{Y} = \frac{\sum_{i=0}^{S} x_i s_i}{S}$ are the mean scores of the two groups. To test for the difference between the score means, a statistic can be defined as

$$T_{mean} = \frac{(\bar{X} - \bar{Y})^2}{\hat{var}_0 (\bar{X} - \bar{Y})},$$

where $\hat{var}_0 (\bar{X} - \bar{Y})$ is the estimated variance of $(\bar{X} - \bar{Y})$ under the null hypothesis. $T_{mean}$ also follows a $\chi^2$ distribution with 1 df. The numerator $(\bar{X} - \bar{Y})^2 = \frac{\left[\sum_{i=0}^{R} x_i (r_i - R s_i)\right]^2}{R^2 S^2}$. The denominator $\hat{var}_0 (\bar{X} - \bar{Y}) = \frac{\left[N \sum_{i=0}^{R} x_i^2 n_i - (R \sum_{i=0}^{R} x_i n_i)^2\right]}{R S N}$. By some algebraic manipulations, it can be shown that $T_{mean} = T_{CA}$. Therefore, the CA trend test is equivalent to a two-sample mean test for the difference of the average scores. In other words, the CA trend test examines the location shift of cell counts between cases and controls.

Pearson’s $\chi^2$ test is an omnibus test for independence in a contingency table. Unlike the CA trend test, it does not require assignment of a score $x_i$ to each genotype category to reflect assumptions about the genetic model. The test statistic is

$$T_P = \sum_{i=0}^{R} \frac{(r_i - n_i R / N)^2}{n_i R / N} + \sum_{i=0}^{S} \frac{(s_i - n_i S / N)^2}{n_i S / N}. \quad (2)$$

Under the null hypothesis, $T_P$ follows a $\chi^2$ distribution with 2 df.
Here we show how $T_p$ can be partitioned into components that measure the location effect and components that measure the dispersion effect by orthogonal polynomials (Beh, 2001; Rayner & Best, 2000). Define a set of orthogonal polynomials

$$g_0(x_i) = 1, \quad g_1(x_i) = \frac{x_i - \mu}{\sqrt{\mu_2}}, \quad \text{and} \quad g_2(x_i) = \beta \left[ (x_i - \mu)^2 - \frac{\mu_3(x_i - \mu)}{\mu_2} - \mu_2 \right],$$

where

$$\mu = \sum_{i=0}^{2} \frac{x_i n_i}{N}, \quad \mu_j = \sum_{i=0}^{2} \frac{(x_i - \mu)^j n_i}{N}, \quad j \in \{2, 3, 4\}, \quad \text{and} \quad \beta = \frac{1}{\sqrt{\mu_4 - \mu_3^2/\mu_2^2}}.$$

Define

$$V_{u1} = \frac{\sum_{i=0}^{2} r_i g_u(x_i)}{\sqrt{R}} \quad \text{and} \quad V_{u2} = \frac{\sum_{i=0}^{2} s_i g_u(x_i)}{\sqrt{S}},$$

where $u \in \{1, 2\}$. Note that $\mu$ is the mean score of the overall table; $V_{11}$ and $V_{12}$ are functions of the location shift of scores in cases and controls, respectively; $V_{21}$ and $V_{22}$ are functions of the dispersion of scores in cases and controls, respectively. It can be shown that

$$T_p = V_{11}^2 + V_{12}^2 + V_{21}^2 + V_{22}^2 \quad (\text{see Appendix A.2 for details}).$$

Therefore, Pearson’s $\chi^2$ test statistic can be decomposed into two parts, with $V_{11}^2 + V_{12}^2$ measuring the location deviation and $V_{21}^2 + V_{22}^2$ measuring dispersion deviation between cases and controls. In other words, Pearson’s $\chi^2$ test simultaneously examines the location and dispersion shifts of cell counts between cases and controls.

Above we analytically show that the difference between the CA trend test and Pearson’s $\chi^2$ test is that the former only examines the location shift, whereas the latter examines both the
location and dispersion shifts between the case and control groups. Below we show by simulation how this difference leads to distinct performance of the two tests under varying genetic models.

1.3 Simulation studies

To connect the genetic and statistical models, here we parameterize the model using some genetic jargon. Set the genotype group $G_0$, i.e., $aa$, as the reference group. Define penetrance as $f_i = P(\text{Affected} | G_i)$, and genotype relative risk as $\lambda_i = f_i / f_0$ with $\lambda_0 = 1$. The null hypothesis of no association can be expressed as $H_0: \lambda_1 = \lambda_2 = 1$. Under the alternative hypothesis, $\lambda_2 \geq \lambda_1 \geq 1$ and $\lambda_2 > 1$. A genetic model can be described in terms of $\lambda_1$ and $\lambda_2$. Specifically, $\lambda_1 = \lambda_2$, $\lambda_1 = (1 + \lambda_2)/2$ and $\lambda_1 = 1$ correspond to the dominant, additive, and recessive models, respectively. In a two-dimensional space we can re-parameterize the model by defining

$$
\lambda_1 = 1 + \lambda \cos \theta \quad \text{and} \quad \lambda_2 = 1 + \lambda \sin \theta,
$$

where $\lambda \geq 0$ is the distance between point $P = (\lambda_1, \lambda_2)$ and point $O = (1,1)$, and $\theta \in [\pi/4, \pi/2]$ be the angle between $OP$ and the horizontal line (Zheng, Joo, & Yang, 2009). Thus, $\theta$ determines the genetic model and $\lambda$ determines how far the genetic model is from the null. The null hypothesis can be rewritten as $H_0: \lambda = 0$. In terms of genetic models, $\theta = \pi/4$, arctan 2, and $\pi/2$ correspond to dominant, additive, and recessive models, respectively. Note that when $\theta \in (\pi/2, \pi)$, $\lambda_1 < 1$ and $\lambda_2 > 1$, it is a heterozygote advantage model, wherein heterozygous individuals have higher fitness than homozygous individuals. A classic example is that the sickle-cell haemoglobin heterozygote provides a protective advantage against malaria (Allison, 1964). In the simulation study below we arbitrarily set $\theta = 3\pi/5$ as an example of heterozygote advantage model.
We performed simulations under the following alternative settings. Assume a disease prevalence ($K$) of 0.1 and the minor allele $A$ frequency ($p$) of 0.3. Fix the alternative hypothesis as $\lambda = 1$ and vary the genetic models by setting $\theta' = \theta / \pi$ from $1/4$ to $1/2$, i.e., from a dominant model to a recessive model, with an increment of 0.01. Under each genetic model, penetrances are determined by

$$f_0 = K / [(1 - p)^2 + 2\lambda_1 p (1 - p) + \lambda_2 p^2] \hspace{0.1cm} \text{and} \hspace{0.1cm} f_i = \lambda_i f_0.$$  

The probabilities of the two trinomial distributions for cases and controls are then

$$p_i = P(G_i) f_i / K \hspace{0.1cm} \text{and} \hspace{0.1cm} q_i = P(G_i) (1 - f_i) / (1 - K),$$

respectively. The sample size is set to be $R = S = 150$ such that the power of tests ranges from 0.3 to 0.7 at the test significance level of 0.05. For each model, 10,000 replicates are simulated and each dataset is examined by both tests. When performing the CA trend test, the score set $(x_0 = 0, x_1 = 1/2, x_2 = 1)$ is applied to the three genotypes, i.e., an additive model is assumed by convention. The empirical power at the 0.05 level is calculated as the proportion of the 10,000 replicates for which the $P$-value is less than or equal to 0.05. The average power over 10 simulations was plotted in Figure 1.1. Simulations were also performed under the heterozygote advantage model ($\theta = 3\pi/5$).

The results are consistent with previous results (e.g., Gonzalez et al., 2008; Kuo & Feingold, 2010; Li et al., 2009; Loley, Konig, et al., 2013)—the CA trend test outperforms Pearson’s $\chi^2$ test under a dominant model; the power advantage increases as the genetic model evolves into an additive mode; and then the advantage diminishes as the model keeps evolving toward a recessive mode; around a near recessive model ($\theta \approx 0.46\pi$, $\lambda_1 \approx 1.13$, $\lambda_2 \approx 1.99$).
Figure 1.1: Power comparison of the CA trend test ($T_{CA}$) and Pearson’s $\chi^2$ test ($T_P$)
The solid line denotes $T_{CA}$ and the red dotted line denotes $T_P$. Along the $x$-axis $\theta = \pi/4$, arctan 2, and $\pi/2$ correspond to dominant, additive, and recessive models, respectively. The $y$-axis is the average empirical power over 10 simulations at the 0.05 level based on 10,000 replicates each. The disease prevalence equals 0.1; the minor allele frequency equals 0.3; and the sample size is 150 cases and 150 controls.

The two tests have similar power; and Pearson’s $\chi^2$ test outperforms the CA trend test as the model further evolves toward the recessive mode. Under the heterozygote advantage model ($\theta = 3\pi/5$) Pearson’s $\chi^2$ test is far more powerful than the CA trend test—0.81 versus 0.17.
We used two intuitive metrics to measure the location and dispersion shifts of cell counts between cases and controls. The difference of the weighted scores between the two groups, 

$$D_{mean} = \sum_{i=0}^{2} x_i r_i - \sum_{i=0}^{2} x_i s_i,$$

was used to measure the location shift. The score set ($x_0 = 0, x_1 = 1/2, x_2 = 1$) was used. The difference between the standard deviations of the cell counts between the two groups, 

$$D_{sd} = \sqrt{\frac{\sum_{i=0}^{2}(r_i - \bar{r})^2}{2}} - \sqrt{\frac{\sum_{i=0}^{2}(s_i - \bar{s})^2}{2}},$$

where $\bar{r} = \frac{\sum_{i=0}^{2} r_i}{3}$ and $\bar{s} = \frac{\sum_{i=0}^{2} s_i}{3}$, was used to measure the dispersion shift. We empirically measured the deviations of these two metrics from the null under five alternative models—a dominant model ($\theta = \pi/4, \lambda_1 = \lambda_2 \equiv 1.71$) and an additive model ($\theta = \arctan(2), \lambda_1 \equiv 1.45, \lambda_2 \equiv 1.89$), under which the CA trend test is more powerful; a near recessive model ($\theta = 0.48\pi, \lambda_1 \equiv 1.06, \lambda_2 \equiv 2.00$), a recessive model ($\theta = \pi/2, \lambda_1 = 1.00, \lambda_2 = 2.00$), and a heterozygote advantage model ($\theta = 3\pi/5, \lambda_1 \equiv 0.69, \lambda_2 \equiv 1.95$), under which Pearson’s $\chi^2$ test is more powerful. First, the empirical distribution of each metric under the null hypothesis was obtained based on 100,000 replicates with the 2.5% and 97.5% quantiles calculated. Then the empirical distribution under the alternative hypothesis was also calculated based on 100,000 replicates. The area under the alternative distribution curve with values equal to or more extreme than the 97.5% or 2.5% quantiles was calculated, which represented the power of detecting the deviation of the metric from its null distribution at the significance level of 0.05 (Figure 1.2).

The power to detect the $D_{mean}$ and $D_{sd}$ shifts elucidates the power difference of Pearson’s $\chi^2$ test and the CA trend test. Under a dominant model the power to detect the $D_{sd}$ shift is small—0.10. Accounting for the dispersion information cannot offset the cost of one
Figure 1.2: Power to detect the location and dispersion shifts of cell counts between cases and controls under additive and near recessive genetic models

The first row is under a dominant model ($\theta = \pi/4$, $\lambda_1 = \lambda_2 \approx 1.71$); the second row is under an additive model ($\theta = \arctan(2)$, $\lambda_1 \approx 1.45$, $\lambda_2 \approx 1.89$); the third row is under a near recessive model ($\theta = 0.48\pi$, $\lambda_1 \approx 1.06$, $\lambda_2 \approx 2.00$); the fourth row is under a recessive model ($\theta = \pi/2$, $\lambda_1 = 1.00$, $\lambda_2 = 2.00$); and the fifth row is under a heterozygote advantage model ($\theta = 3\pi/5$, $\lambda_1 \approx 0.69$, $\lambda_2 \approx 1.95$). The first column is on the distribution of $D_{mean}$ and the second column is on $D_{std}$. The empirical distribution curves are based on 100,000 replicates. The solid line denotes the null distribution and the dashed line denotes the alternative distribution. The critical values for shaded area are the 2.5% and 97.5% quantiles under the null distribution. Note that the area is calculated under both tails but only one tail is visible in most situations.
extra df; thus Pearson’s $\chi^2$ test is less powerful. Although the dispersion information increases under an additive model, the CA trend test remains more powerful since it is the most efficient test with correct model assumptions. As the genetic model evolves toward the recessive end, Pearson’s $\chi^2$ test becomes more powerful because there is more dispersion information. Under a heterozygote advantage model the CA trend test is less powerful because the location information is small.

1.4 Discussion

Numerous simulation studies showed the 1 df CA trend test is less powerful than the 2 df Pearson’s $\chi^2$ test when the mode of inheritance is recessive or near recessive due to the incorrect model assumption, which, as a phenomenon of ‘dog bites man’, is not newsworthy (Elston, 1989). In this project we analytically reveal the statistical reason of Pearson’s $\chi^2$ test outperforming the CA trend test as the genetic model evolves toward the recessive end. We confirm by simulation that under a near recessive model and a recessive model, the effect of dispersion deviation outweighs that of the location deviation. However, it is not a necessary condition for Pearson’s $\chi^2$ test to outperform the CA trend test. Rather, the relative power of the two tests depends on whether the gain by taking into account of dispersion information can offset the cost of one extra df. There are tests proposed to simultaneously test location and dispersion (Lang & Iannario, 2013; Rayner & Best, 2000). In a genetic association study involving only $2 \times 3$ contingency tables as discussed in this project, these tests are equivalent to Pearson’s $\chi^2$ test. When examining contingency tables of $2 \times M$, where $M > 3$, for example, when testing association for multi-allelic copy number variations with dosage effects, these tests can potentially be more powerful than Pearson’s $\chi^2$ test.
As there is no single best test for all situations (Kuo & Feingold, 2010), researchers have developed the so-called MAX test (Freidlin, Zheng, Li, & Gastwirth, 2002; Gonzalez et al., 2008; Hothorn & Hothorn, 2009; Li, Zheng, Li, & Yu, 2008; Loley, Konig, et al., 2013; So & Sham, 2011; Zang & Fung, 2011), which is more robust than the CA trend test and more powerful than Pearson’s $\chi^2$ test. It is worth noting that the MAX test statistic maximized over $\theta \in [\pi/4, \pi/2]$ is identical to Pearson’s $\chi^2$ test statistic (Zheng et al., 2009). Its power gain lies in the cost of maximization over one nuisance parameter is smaller than that of one extra df when performing testing.
CHAPTER 2

DECOMPOSING PEARSON’S $\chi^2$ TEST: A LINEAR REGRESSION AND ITS DEPARTURE FROM LINEARITY

2.1 Introduction

In genetic association studies, there is no single best test for all situations. If the true genetic model were known, the association test reflecting this model would have optimal power. In practice the CA trend test under the assumption of the additive model is a standard approach in case-control genome-wide association screens (GWAS) for its robustness, as shown in numerous simulation studies (e.g., Gonzalez et al., 2008; Kuo & Feingold, 2010; Li et al., 2009; Loley, König, Hothorn, & Ziegler, 2013). We happened to find by simulation that the CA trend test either under the additive assumption or under the multiplicative assumption outperformed that under the recessive assumption even if the true genetic model is recessive when the minor allele frequency is low and the sample size is not large enough (Figure 2.1). This anomalous phenomenon met the “man bites dog” criterion to be newsworthy (Elston, 1989) and motivated us to investigate the problem analytically. By elucidating the origin of the CA trend test as a test to examine whether there is a trend in binomial proportions across levels of an ordinal variable by a linear regression model, we decompose Pearson’s $\chi^2$ test statistic (2) by the
Figure 2.1: Power comparison of the Cochran-Armitage trend test by the additive, multiplicative, and recessive models under recessive modes of inheritance for low-frequency variants

On the simulation models, the relative risks for $Aa$ and $AA$ are defined as $\lambda_1 = 1 + \cos \theta$ and $\lambda_2 = 1 + \sin \theta$, respectively. Therefore, $\theta \in (0.47\pi, 0.48\pi, 0.49\pi, 0.50\pi)$ ranges from near-recessive models to the recessive model; The labels “Additive”, “Multiplicative”, and “Recessive” indicate the analysis models. The y-axis is the empirical power based on 10,000 replicates at the 0.05 level. The disease prevalence equals 0.1; the minor allele frequency equals 0.05; and the total sample size $N \in (300, 500, 800, 1000)$ with equal numbers of cases and controls.
ordinary least squares (OLS). Under this theoretical framework, the anomalous behavior of the CA trend test can be explained; further, we prove some results previously published in the *Journal* by other investigators (Zheng et al., 2009).

The outcomes of the genetic case-control study are summarized in Table 1, with the same notations in Chapter 1. Assume \( r_i \)'s are drawn from binomial distributions \( B(1, n, \pi_i) \). The null hypothesis of no association between the disease and genotype is \( H_0: \pi_0 = \pi_1 = \pi_2 \). Note that \( \pi_i \) is the probability of a subject with genotype \( G_i \) being affected conditional on the marginal counts \( n_i, R \) and \( S \), different from the penetrance function \( f_i = P(\text{affected}|G_i) \) defined in a general population. Equivalently, we can also assume \( (r_0, r_1, r_2) \) and \( (s_0, s_1, s_2) \) follow the same trinomial distributions as in Chapter 1.

Assume \( G_0, G_1, \) and \( G_2 \) are three ordered categories, a more restricted alternative hypothesis is \( H_1: \pi_0 \leq \pi_1 \leq \pi_2 \) or \( \pi_0 \geq \pi_1 \geq \pi_2 \) with at least one strict inequality. Assign the same scores \( (x_0, x_1, x_2) \) to \( G_0, G_1, \) and \( G_2 \), respectively, as in Chapter 1. The CA trend test (1) examines whether there is a linear relationship between \( \pi_i \)'s and \( x_i \)'s by fitting a linear regression model

\[
\pi_i = \alpha + \beta x_i + \varepsilon_i, \tag{3}
\]

where \( \varepsilon_i \) is a random error term with mean equal to zero and variance equal to \( \sigma^2 \). The null hypothesis is \( H_0: \beta = 0 \). In practice score sets of \( (0, 0, 1), (0, 0.5, 1), \) and \( (0, 1, 1) \) are typically used to represent recessive, additive, and dominant models, respectively.

### 2.2 Simulation studies

Above we review the data structure of a genetic association study and the CA trend test used for analysis; below we first describe the simulation models and results, then derive
analytical results to interpret the simulation results. The same notations and reparametrization as in Chapter 1 are used here. We performed simulations under the following alternative settings. Assume a disease prevalence \((K)\) of 0.1 and the minor allele \(A\) frequencies \(p \in (0.05, 0.1, 0.15, 0.2)\). Fix the alternative hypothesis as \(\lambda = 1\) and vary the genetic models by setting \(\theta' = \theta / \pi\) from \(\arctan 2 / \pi\) to 1/2, i.e., from the additive model to the recessive model, with an increment of 0.01. Assume a balanced design, i.e., \(R = S\) with the total sample size \(N \in (300, 500, 800, 1000)\). For each model, 10,000 replicates are simulated and each dataset is examined by the CA trend test under the additive model with the score set \((0, 0.5, 1)\), denoted as \(T_{CA}^{Add}\), under the recessive model with the score set \((0, 1, 1)\), denoted as \(T_{CA}^{Rec}\), and under the multiplicative model with the score set \((0, \frac{1}{\sqrt{\lambda_2+1}}, 1)\), denoted as \(T_{CA}^{Mul}\). Note that there are unlimited ways to assign the multiplicative model scores. As \(\lambda_2 > \lambda_1\) when \(\theta \in (\pi/4, \pi/2]\) under the alternative hypothesis, we fix the upper limit \(\lambda_2\), which leads to a multiplicative model score set \((1, \sqrt{\lambda_2}, \lambda_2)\). It can be further transformed into \((0, \frac{1}{\sqrt{\lambda_2+1}}, 1)\) because the CA trend test is invariant to the linear transformation of the scores (Tarone & Gart, 1980). The empirical power at the 0.05 level is calculated as the proportion of the 10,000 replicates for which the P-value is less than or equal to 0.05.

First, in most of the parameter space \(T_{CA}^{Add}\) and \(T_{CA}^{Mul}\) have comparable power except that \(T_{CA}^{Mul}\) is more powerful when the underlying model is close to recessive (Figure B.2 in Appendix B). It can be predicted because \(\frac{1}{\sqrt{\lambda_2+1}}\) is constrained between approximately 0.414, when \(\theta = \pi/4\), and 0.421, when \(\theta = \arctan 2\) according to the simulation setting. The multiplicative model lies between the recessive and additive models but in close proximity to the additive end. Therefore overall \(T_{CA}^{Add}\) and \(T_{CA}^{Mul}\) have comparable power, but under near recessive models the
latter is closer to the truth, and thus outperforms the former. Regardless of the simulation setting, the upper limit of \( \frac{1}{\sqrt{\lambda_2+1}} \) is 0.5 because \( \lambda_2 > 1 \), and \( \frac{1}{\sqrt{\lambda_2+1}} = 0.4 \) when \( \lambda_2 = 2.25 \), which is greater than most GWAS hits (see, e.g., Hodge & Greenberg, 2016). Therefore, the similarity between \( T_{CA}^{Add} \) and \( T_{CA}^{Mul} \) in terms of power exists in general.

Second, in most of the parameter space the results are as expected that \( T_{CA}^{Add} \) and \( T_{CA}^{Mul} \) are more powerful when the underlying genetic model is close to additive, whereas \( T_{CA}^{Rec} \) is more powerful when the underlying model is close to recessive (Figure B.2 in Appendix B). However, \( T_{CA}^{Add} \) and \( T_{CA}^{Mul} \) are more powerful than \( T_{CA}^{Rec} \) at the recessive end when the minor allele frequency is low and the sample size is small (Figures 2.1 & Figure B.2), although \( T_{CA}^{Add} \) and \( T_{CA}^{Mul} \) are just above the nominal test level and \( T_{CA}^{Rec} \) is below the nominal level. In other words, the wrong analysis model outperforms the true model that is used for data simulation in those situations. When the sample sizes are 300 or 500, the power of \( T_{CA}^{Add} \) and \( T_{CA}^{Mul} \) are greater than that of \( T_{CA}^{Rec} \); with the increase of sample sizes, the power of \( T_{CA}^{Rec} \) increases to surpass that of \( T_{CA}^{Add} \) and \( T_{CA}^{Mul} \).

2.3 Method

Here we decompose Pearson’s \( \chi^2 \) test statistic (2) by the theory of OLS. Unlike the CA trend test, Pearson’s \( \chi^2 \) test is an omnibus test for independence in a contingency table without pre-assigning scores. In terms of genetics, it makes no assumption about the genetic model. It tests the alternative hypothesis of not all \( \pi_i \)’s are equal. Under the null hypothesis, \( T_p \) follows a \( \chi^2 \) distribution with 2 df. Pearson’s \( \chi^2 \) test statistic \( T_p \) can be decomposed into two components—one is the CA trend test statistic \( T_{CA} \) that measures the goodness-of-fit of the linear regression model (3), the other, denoted as \( T_{SSE} \), measures the discrepancy between the data and
model (3). The derivation details are described in Appendix B.1, and here we summarize the results. Denote by $p_i^o (= r_i / n_i)$ the observed value of $\pi_i$, which has a weight of $n_i$, by $p^o = R/N$ the overall proportion of cases, and by $\bar{x} = \sum_{i=0}^{2} n_i x_i / N$ the average score. By OLS we can estimate for model (3) that

$$\hat{\beta} = \frac{\sum_{i=0}^{2} n_i (p_i^o - p^o)(x_i - \bar{x})}{\sum_{i=0}^{2} n_i (x_i - \bar{x})^2} \quad (4)$$

and

$$\hat{\pi}_i = p^o + \hat{\beta}(x_i - \bar{x}) \quad (5)$$

It can be shown

$$T_P = \frac{1}{p^o(1-p^o)} \sum_{i=0}^{2} n_i (p_i^o - p^o)^2$$

$$= \frac{1}{p^o(1-p^o)} \sum_{i=0}^{2} n_i (\hat{\pi}_i - p^o)^2 + \frac{1}{p^o(1-p^o)} \sum_{i=0}^{2} n_i (p_i^o - \hat{\pi}_i)^2 \quad (6)$$

and

$$T_{CA} = \frac{1}{p^o(1-p^o)} \sum_{i=0}^{2} n_i (\hat{\pi}_i - p^o)^2.$$ 

Note that $\sum_{i=0}^{2} n_i (p_i^o - p^o)^2$, $\sum_{i=0}^{2} n_i (\hat{\pi}_i - p^o)^2$, and $\sum_{i=0}^{2} n_i (p_i^o - \hat{\pi}_i)^2$, are total sum of squares (SSTO), regression sum of squares (SSR), and error sum of squares (SSE), respectively, for model (3); thus we define $T_{SSE} = \frac{1}{p^o(1-p^o)} \sum_{i=0}^{2} n_i (p_i^o - \hat{\pi}_i)^2$. In summary, equation (6) is identical to $T_P = T_{CA} + T_{SSE}$, and also

$$\frac{\text{SSTO}}{p^o(1-p^o)} = \frac{\text{SSR}}{p^o(1-p^o)} + \frac{\text{SSE}}{p^o(1-p^o)} \quad (7)$$

which decompose $T_P$ into a component measuring the goodness-of-fit of the linear regression model (3) and a component measuring the departure of data from linearity. Note that
\( p^o(1 - p^o) \) can be regarded as the pooled variance of \( \pi_i \)'s; by Cochran’s theorem (W.G. Cochran, 1934) \( T_p, T_{CA}, \) and \( T_{SSE} \) follow \( \chi^2 \) distributions with 2, 1, and 1 df, respectively.

According to the analytical work above on decomposing Pearson’s \( \chi^2 \) test statistic \( T_p \), we hypothesize that the additive and multiplicative coding scheme in the linear regression model (3) fits data better than the recessive coding scheme in case of low allele frequencies and relatively small sample sizes, even though the true mode of inheritance is recessive, which explains why \( T_{CA}^{Add} \) and \( T_{CA}^{Mul} \) are more powerful than \( T_{CA}^{Rec} \). We provide a numerical example to illustrate the point (Figure 2.2). A 2 × 3 table with \((r_0, r_1, r_2) = (443, 52, 5)\) and \((s_0, s_1, s_2) = (465, 34, 1)\) was generated by simulation under the recessive model with \( K = 0.1, p = 0.05, \) and \( R = S = 500 \). When analyzing under the additive model, the coefficient of determination \( R^2 (= SSR/SSTO) \), which measures the goodness-of-fit of the linear regression model (3), equals 0.0067; \( R^2 = 0.0069 \) when analyzing under the multiplicative model; in contrast \( R^2 = 0.0027 \) when analyzing under the recessive model. Therefore, both additive and multiplicative models fit the data better than the recessive model. Correspondingly, the association is significant under the additive model (\( P\)-value=0.0095) and under the multiplicative model (\( P\)-value=0.0086) but insignificant under the recessive model (\( P\)-value=0.1014).

Further we show that the additive and multiplicative coding scheme leads to better fit of the linear regression model (3) than the recessive coding scheme partially by increasing the spread of the data points. Denote the sample variance of the scores by \( s_x^2 \), which equals \( \Sigma_{i=0}^2 n_i (x_i - \bar{x})^2 / (N - 1) \). By OLS we know \( \text{Var}(\hat{\beta}) = \sigma^2 / [(N - 1)s_x^2] \). Given a dataset \( SSTO \) is fixed. Thus,
Figure 2.2: Comparison of the Cochran-Armitage trend test by different analysis models under the recessive mode of inheritance: a numerical example

A 2 x 3 table with \((r_0, r_1, r_2) = (443, 52, 5)\) and \((s_0, s_1, s_2) = (465, 34, 1)\) was generated by simulation under the recessive model with \(K = 0.1, p = 0.05\), and \(R = S = 500\). It was analyzed using the CA trend test assuming (A) the additive model \((x_0 = 0, x_1 = 0.5, x_2 = 1)\), (B) the multiplicative model \((x_0 = 0, x_1 = \frac{1}{\sqrt{\lambda_2 + 1}} = \sqrt{2} - 1, x_2 = 1)\), (C) the recessive model \((x_0 = 0, x_1 = 0, x_2 = 1)\), and (D) the “best” model \((x_0 = 0, x_1 = \frac{p_1^0 - p_0^0}{p_2^0 - p_0^0} \approx 0.338, x_2 = 1)\). The x-axis is the scores used for the genotype groups in the CA trend test; the y-axis is the proportion of cases in each genotype group; the area of the circles is proportional to the counts; the straight line is the fitted regression line base on model (3); and \(R^2\) is the corresponding coefficient of determination.

(A) Additive model

(B) Multiplicative model

(C) Recessive model

(D) The "best" model
\[ E(R^2) = E(SSR)/SSTO = E\left(\frac{1}{2} \sum_{i=0}^{2} n_i (\hat{\beta}_i - \beta^o)^2\right)/SSTO \]

\[ = (N - 1)s^2_x E(\hat{\beta}^2)/SSTO = [\sigma^2 + (N - 1)s^2_x \beta^2]/SSTO, \]

which shows \( E(R^2) \) is a monotonically increasing function of \( s^2_x \). Therefore, the more spread out the scores are, the better a linear regression model could potentially fit the data. However, note that \( E(R^2) \) is also a monotonically increasing function of \( \beta \), which will be different with different score sets. The overall goodness-of-fit depends on \( s^2_x \beta^2 \). For the same numerical example above, under the additive model \( s^2_x = 0.025 \), \( E(\hat{\beta}) = 0.259 \), and thus \( s^2_x E(\hat{\beta})^2 = 0.00168 \); under the multiplicative model \( s^2_x = 0.019 \), \( E(\hat{\beta}) = 0.301 \), and thus \( s^2_x E(\hat{\beta})^2 = 0.00173 \); under the recessive model, \( s^2_x = 0.006 \), \( E(\hat{\beta}) = 0.335 \), and \( s^2_x E(\hat{\beta})^2 = 0.00067 \).

With the analytical work elucidating the CA trend test and decomposing Pearson’s \( \chi^2 \) test statistic \( T_p \), we can easily prove some results previously obtained by other investigators. Zheng et al. (2009) showed that the CA trend test with the score set \( (0, \frac{p_1^o - p_0^o}{p_2^o - p_0^o}, 1) \), denoted as \( T_{CA}^{Best} \), equals \( T_p \); moreover, if \( 0 \leq x_1 \leq 1 \), \( T_{CA}^{Best} \) equals the so-called MAX statistic (Freidlin et al., 2002), which is defined as \( \max_{x_1 \in [0,1]} T_{CA}^{(0,x_1,1)} \). By OLS we have equations (4) and (5). When the score set is \( (0, \frac{p_1^o - p_0^o}{p_2^o - p_0^o}, 1) \), it can be shown that equation (4) leads to \( \hat{\beta} = p_2^o - p_0^o \), and equation (5) leads to \( \hat{\pi}_i = p_1^o \). It indicates the three points \( (0, p_0^o), \left(\frac{p_1^o - p_0^o}{p_2^o - p_0^o}, p_1^o\right), \) and \( (1, p_2^o) \), are right on the regression line, and the fitted values equal the observed values (Figure 2.2). As such, \( SSE = 0 \), and thus \( T_p = T_{CA}^{Best} \). Note that this equation is valid regardless of the range of \( x_1 \). Therefore \( \max_{x_1 \in [0,1]} T_{CA}^{(0,x_1,1)} \leq T_{CA}^{Best} \), and the equality holds when \( 0 \leq x_1 \leq 1 \). For the same numerical
example above, if the score set \( \left( 0, \frac{p_1^2 - p_0^2}{p_2^2 - p_0^2}, 1 \right) \) could be used, the association is most significant (P-value=0.0070) among all tested models. Moreover, \( s_x^2 = 0.015 \), which is smaller than that under either the additive model or the multiplicative model; however, \( E(\hat{\beta}) = 0.345 \), and thus \( s_x^2[E(\hat{\beta})]^2 = 0.00174 \), which is the greatest among all the tested models.

2.4 Discussion

We previously showed Pearson’s \( \chi^2 \) test is more powerful than \( T_{CA}^{Add} \) in case of the heterozygote advantage model and provided a theoretical argument from the distribution perspective that Pearson’s \( \chi^2 \) test examines both the location and dispersion shifts, whereas \( T_{CA}^{Add} \) only examines the location shift (Z. Zhou, Ku, Huang, Xing, & Xing, 2017). Under the heterozygote advantage model the dispersion information overshadows the location information, therefore Pearson’s \( \chi^2 \) test is more powerful. The analytical work in this project provides justification from another perspective that \( T_{CA}^{Add} \) only tests the goodness-of-fit of the linear model (3), whereas \( T_p \) also includes a component measuring the departure of data from linearity. Under the heterozygote advantage model \( \lambda_1 < 1 \) and \( \lambda_2 > 1 \), i.e., severe departure from linearity, therefore Pearson’s \( \chi^2 \) test is more powerful. Although we started this project with an observation that the additive coding scheme outperforms the recessive coding scheme even the true genetic model is recessive, note from previous simulation studies (e.g. Gonzalez et al., 2008; Kuo & Feingold, 2010; Li et al., 2009; Loley, König, et al., 2013) that \( T_{CA}^{Add} \) overall performs worse when the true genetic model is recessive than when the model is dominant or additive. The underlying reason is that under a recessive model there is less information on mean score difference between cases and controls, which the CA trend test measures.
The CA trend test under the additive model shows robustness across various genetic models. Intuitively it makes sense in that it lies in the middle of the parameter space between the two extreme ends. By the analytical work in this project we speculate a numerical reason is that the additive coding scheme overall leads to a good fit of the linear regression model (3) by increasing the spread of the data points. Note that the anomalous behavior of the CA trend test observed this project happens only to binary traits, but not to continuous traits (data not shown). It can also be deduced from the linear regression nature of the CA trend test. Compared to the additive coding scheme, the recessive coding scheme not only leads to less spread of the data points in terms of scores, in case of binary traits it also leads to heteroscedasticity because subjects with the same score 0 consist of two groups $\pi_0$ and $\pi_1$ (Figure 2.2C). On the contrary this issue does not arise in case of continuous traits.

Following Sasieni’s seminal paper (Sasieni, 1997) $T_{CA}^{Add}$ became the standard practice in GWAS. In recent years, the generalized linear model, in particular, logistic regression for a binary trait, is advocated for its flexibility in modeling different types of traits and including covariates, as well as providing estimates of parameters of interest such as odds ratio (Dizier, Demenais, & Mathieu, 2017; Loley, König, et al., 2013; So & Sham, 2011; Wellek & Ziegler, 2012). One study closely related to our work is the general regression model proposed by Dizier et al. (Dizier et al., 2017). It tests for both additive effect and deviation from additive effect, which is analogous to our decomposing $T_P$ into a component measuring the goodness-of-fit of the linear regression model and a component measuring the departure of data from linearity. Here we focus on elucidating the characteristics of the CA trend test to explain the anomalous behavior of the test in a certain circumstance. The superior power of the general regression
model in GWAS warrants further investigation on how to modify the CA trend test to increase power based on the current analytical work.
CHAPTER 3

TESTING NONLINEAR GENE-ENVIRONMENT INTERACTION THROUGH VARYING COEFFICIENT AND LINEAR MIXED MODELS

3.1 Introduction

In GWASs, beyond the marginal genetic effects, gene-environment ($G \times E$) interactions have demonstrated great importance in understanding the biological etiology of human traits and improving the ability to detect genetic variants interacted with environmental factors but show little marginal effect (Thomas, 2010). The genetic factors can be gene expression, single-nucleotide polymorphisms (SNPs) or other types of measurement. The environmental factors can be exogenous exposures such as air pollutions, pesticides or treatment in a randomized clinical trial (Mukherjee, Ahn, Gruber, & Chatterjee, 2011) and clinical measurements such as body mass index or some nutrition intake. However, discovering significant $G \times E$ interaction is a hard task. Specifically, the sample size needed for detecting $G \times E$ interaction can be four times larger than detecting the main genetic effect with similar effect magnitude (Thomas, 2011).

Many statistical methods have been proposed at this area, among which the conventional methods are aimed at detecting the $G \times E$ interaction with respect to individual genetic variant like SNP. For the model based methods (Kraft, Yen, Stram, Morrison, & Gauderman, 2007;
Maity, Carroll, Mammen, & Chatterjee, 2009), they usually assume linear $G \times E$ interaction, which can be easily violated when the underlying biological relationship between genetic and environmental factors is nonlinear. For example, based on the study of Levy et al. (2009), Wang and Chen (2012) pointed out that the genetic effect of SNP rs7136259 on systolic blood pressure differs nonlinearly at different ages: higher at early ages, lower at middle ages, and then higher at late ages. In another study, Sparrow et al. (2012) discovered deleterious mutations in gene HES7 and MESP2 for congenital scoliosis, and the severity of the genetic risk is influenced nonlinearly by the level of short-time gestational hypoxia in a mouse model. In these cases, model misspecification could result in large estimation bias and power loss. To account for the nonlinear $G \times E$ interaction, Ma et al. (2011) proposed a varying-coefficient model. This procedure allows the coefficients to change smoothly with the value of other variables, which introduces enough flexibility to evaluate dynamic pattern like interaction.

Besides the model misspecification, another common issue in GWAS is the population stratification, which represents the systematic differentiation in allele frequencies between subpopulations. As a major source of confounding in GWAS, it could lead to spurious associations. Recently, with the development of computationally efficient algorithms, linear mixed models with genetic relatedness as a variance component have become popular for controlling population stratification (Kang et al., 2008; Lippert et al., 2011; Yang, Lee, Goddard, & Visscher, 2011; Zhang et al., 2010; X. Zhou & Stephens, 2012). This variance component would adjust for the subtle relatedness among individuals, thus the correct p-values would be obtained. Genome-wide complex trait analysis (GCTA), a software developed by Yang et al. (Yang et al.), is able to estimate the genetic relatedness matrix (GRM) between individuals by all the genotyped SNPs.
In this project, we proposed a novel procedure called **Varying-Coefficient embedded Linear Mixed Model (VC-LMM)**, which is targeted at detecting nonlinear $G \times E$ interaction when the population stratification is present. The coefficient of genetic effect is a smooth function with respect to the environmental factor of interest. To allow for sufficient flexibility, we estimated the smooth function as a polynomial spline function including one fixed effect part and one random effect part. In addition, the population stratification is adjusted by GRM, which is the second random effect. Under this framework, the linear mixed model has two variance components, and the goal for detecting $G \times E$ interaction becomes simultaneous testing the significance of a subset of fixed effects and a random effect. Unlike current methods for nonlinear $G \times E$ interaction that focus only on the analysis of common variant, the proposed VC-LMM was investigated for both common and low frequency variants. We demonstrated VC-LMM via simulation studies for type I error rates and power comparison.

### 3.2 Method

Let $Y_{n \times 1}, X_{n \times p^c}, E_{n \times 1}, G_{n \times 1}, \Phi_{n \times n}$ be the continuous phenotypes, design matrix for $p^c$ covariates, continuous environmental factors, genotypes and GRM for $n$ subjects, respectively. Assume that $E$ is a subset of $X$. Let $Z_1$ be the Cholesky decomposition of $\Phi$. Thus, the model with $G \times E$ interaction and adjustment for genetic relatedness can be specified as

$$Y = X\eta + f(E) * G + Z_1 b_1 + \epsilon, \quad (8)$$

where $\eta$ are the fixed effects of the covariates, $f(E) = (f(E_1), f(E_2), ..., f(E_n))^T$ and $f(E_i), i = 1, 2, ..., n$ is an unspecified smooth function of environmental factor, “*” denotes element wise multiplication, $Z_1$ is the Cholesky decomposition of $\Phi, b_1 \sim N(0, \sigma_b^2 I_{n \times n})$ are the genetic random effects, $\epsilon \sim N(0, \sigma_\epsilon^2 I_{n \times n})$ are the random errors.
To adjust for a large class of functions, we specify the smooth function to be a flexible spline function

\[ f(x) = \sum_{j=0}^{h} \beta_j x^j + \sum_{l=1}^{L} b_l (x - \tau_l)_+^h, \]  

(9)

where \( \tau_l, l = 1, 2, \ldots, L \) are a sequence of knots and \((\cdot)_+ = \max(\cdot, 0)\). In our model, \( L \) and \( h \) are set to 15 and 1. Plug (9) into (8), we have

\[ Y = X\eta + W\beta + Z_1b_I + Z_2b_{II} + \epsilon, \]  

(10)

where \( W = (G, G \ast E, \ldots, G \ast E^h)^T, \beta = (\beta_0, \beta_1, \ldots, \beta_h)^T, Z_2 = (G \ast (E - \tau_1)_+^h, \ldots, G \ast (E - \tau_K)_+^h)^T, \) and \( b_{II} = (b_1, \ldots, b_L)^T \sim N(0, \sigma_b^2 I_{L \times L}) \) are the interaction random effects. Note that all the minus, power and multiplication are element wise operations.

To test the \( G \times E \) interaction, the null hypothesis is

\[ H_0: f(E) = 0, \]

thus, it’s equivalent to test

\[ H_0: \beta = 0 \text{ and } \sigma_b^2 = 0 \]

in (10). Under this framework, the \( P \)-value can be obtained by the generalized \( F \)-test (Wang & Chen, 2012).

Let \( \gamma \) and \( \xi \) be the variance ratios \( \frac{\sigma_g^2}{\sigma_\epsilon^2} \) and \( \frac{\sigma_b^2}{\sigma_\epsilon^2} \), respectively. Under the null hypothesis, the residual sum of squares is

\[ RSS_0(\gamma) = \frac{1}{\sigma_\epsilon^2} \{Y - X\hat{\gamma}(\gamma)\}^T V(\gamma)^{-1} \{Y - X\hat{\gamma}(\gamma)\}, \]
where \( \hat{\eta}(\gamma) = \{X^TV_0(\gamma)^{-1}X\}^{-1}X^TV_0(\gamma)^{-1}Y \) and \( V_0(\gamma) = I_{n \times n} + \gamma Z_1Z_1^T \). Under the alternative hypothesis, denote \( D = [X, W] \) and \( \rho = [\eta^T, \beta^T]^T \), then the residual sum of squares is

\[
RSS_1(\gamma, \xi) = \frac{1}{\sigma^2} \{Y - D\hat{\rho}(\gamma, \xi)\}^TV_1(\gamma, \xi)^{-1}\{Y - D\hat{\rho}(\gamma, \xi)\},
\]

where \( \hat{\rho}(\gamma, \xi) = \{D^TV_1(\gamma, \xi)^{-1}D\}^{-1}D^TV_1(\gamma, \xi)^{-1}Y \) and \( V_1(\gamma, \xi) = V_0(\gamma) + \xi Z_2Z_2^T \). By Woodbury matrix inversion identity (Woodbury, 1950), \( V_1(\gamma, \xi)^{-1} = V_0(\gamma)^{-1} - \xi\{V_0(\gamma)^{-1}Z_2(I_{L \times L} + \xi Z_2^TV_0(\gamma)^{-1}Z_2)^{-1}Z_2^TV_0(\gamma)^{-1}\} \), so the inverse of an \( n \) by \( n \) matrix can be simplified to the inverse of a \( L \) by \( L \) matrix. Then the test statistic of the generalized \( F \)-test can be defined as

\[
T = \frac{RSS_0(\hat{\varphi}) - RSS_1(\hat{\varphi}, \hat{\xi})}{RSS_1(\hat{\varphi}, \hat{\xi})/n},
\]

where \( \hat{\varphi} \) and \( \hat{\xi} \) can be obtained by the restricted maximum likelihood (REML) under the alternative hypothesis. To accelerate the estimation of parameters, we used the algorithm applied in the factored spectrally transformed linear mixed model (FaST-LMM) (Lippert et al. 2011; Widmer et al., 2014). The reduced computing time is more than 50 folds compared to using R package “nlme” (Pinheiro, Bates, DebRoy, & Sarkar, 2014) for 500 subjects and even more when the sample size is larger. The detailed algorithm for parameter estimation is shown in Appendix C.1.

Under the null hypothesis, \( \sigma_b^2 = 0 \) is on the boundary of the parameter space, so the null distribution of \( T \) cannot be directly derived. Wang et al. (Wang & Chen, 2012) derived a spectral decomposition to obtain the null distribution of the generalized \( F \)-test statistic \( T \). Let \( \varphi_s(\gamma) \) be the \( s \)th eigenvalue of \( Z_2^TP_{V_0}Z_2 \), where \( P_{V_0} = V_0(\gamma)^{-1} - \)
\( V_0(\gamma)^{-1}X_1(X_1^TV_0(\gamma)^{-1}X_1)^{-1}X_1^TV_0(\gamma)^{-1} \). Let \( \omega_s \) be the \( s \)th eigenvalues of \( Z_1^T P_1 Z_1 \), where

\[
P_1 = I_{n \times n} - X_1(X_1^TX_1)^{-1}X_1^T.
\]

Then the null distribution of \( T \) can be simulated in the following steps: for each iteration,

1. Simulate \( u_s \sim \mathcal{N}(0,1) \) independently, for \( s = 1, \ldots, n \) - \( p^c + h + 1 \);

2. Simulate \( v_s \sim \mathcal{N}(0,1) \) independently, for \( s = 1, \ldots, h + 1 \);

3. Obtain \( \hat{\gamma} \) and \( \hat{\xi} \) which maximize

\[
f_n(\gamma, \xi) = -\left( n - (p^c + h + 1) \right) \log \left( \sum_{s=1}^{L} \frac{1}{1 + \xi \varphi_s(\gamma)} u_s^2 + \sum_{s=L+1}^{n-(p^c+h+1)} u_s^2 \right) - \sum_{s=1}^{L} \log(1 + \xi \varphi_s(\gamma)) - \sum_{s=1}^{L} \log(1 + \xi \omega_s),
\]

which is the spectral decomposition of the log restricted likelihood under the alternative hypothesis up to a constant.

4. Obtain the value of \( T \) under the null hypothesis by

\[
T = \frac{\sum_{s=1}^{L} \frac{\hat{\xi} \varphi_s(\hat{\gamma})}{1 + \hat{\xi} \varphi_s(\hat{\gamma})} u_s^2 + \sum_{s=L+1}^{h+1} v_s^2}{\sum_{s=1}^{L} \frac{1}{1 + \hat{\xi} \varphi_s(\hat{\gamma})} u_s^2 + \sum_{s=L+1}^{n-(p^c+h+1)} u_s^2}.
\]

Repeat above steps until the number of replicates is met.

Note that the above computations only involve arithmetic operations so they are extremely fast.
3.3 Simulation studies

To evaluate the performance of the proposed VC-LMM in detecting $G \times E$ interaction, we conducted simulation studies in terms of type I error and power, and compare the results with (1) Crainiceanu’s likelihood ratio test (C-LRT) in linear mixed models with one variance component (Crainiceanu & Ruppert), which can be implemented using R package “RLRsim” (Scheipl, Greven, & Küchenhoff, 2008), and (2) linear regression model for the main genetic and $G \times E$ interaction effects (LMi), where the population stratification is adjusted by top 10 principal components of GRM (Price et al., 2006).

We use dataset of genotyped samples from Dallas Heart Study (DHS) to preserve the realistic patterns of correlated SNPs. After quality control, there are 1,726 subjects and 11,372 SNPs. $N = 500$ and $1000$ subjects are randomly selected from the whole dataset with chromosome Y excluded. For each sample size, the corresponding GRM $\Phi$ is created by GCTA (Yang et al., 2011).

In our simulation, $X = [J, Gender, E]$, where $J$ is a vector of 1, $Gender$ is a vector of Bernoulli trails with probability 0.5, and $E$, a vector of uniform variables ranged within $[-1,1]$, is the continuous environmental factor. Let the corresponding fixed effects $\eta = [1, 0.5, 0.5]^{T}$. To preserve the genetic relatedness among the subjects, we simulate $G$ in the following way:

1. Simulate $G_{\text{continuous}} \sim N(0, \Phi)$ with length $n$, where $G_{\text{continuous}}$ can be regarded as a continuous version of the genotypes for the subjects with correlation equaling to the actual genetic relatedness.

2. Under Hardy-Weinberg equilibrium, discretize $G_{\text{continuous}}$ according to the minor allele frequency (MAF) $p_A$, such that
\[
G = \begin{cases} 
0, & G_{\text{continuous}} \leq G_{\text{continuous, round}}(n(1-p_A)^2) \\
1, & G_{\text{continuous, round}}(n(1-p_A)^2) < G_{\text{continuous}} \leq G_{\text{continuous, round}}(n(1-p_A^2)) \\
2, & G_{\text{continuous, round}}(n(1-p_A^2)) < G_{\text{continuous}} \leq G_{\text{continuous}, n} 
\end{cases}
\]

where \(G_{\text{continuous}, i}\) is the \(i\)-th smallest \(G_{\text{continuous}}\), and \(\text{round}(\cdot)\) takes the value to the nearest integer.

To examine both common and low frequency variants, we set \(p_A = 0.05, 0.1, 0.2, 0.3, 0.4\) and 0.5. Note that \(\text{round}(500 \cdot 0.05^2) = 1\), which means at this extreme case, only one subject is homozygous for minor allele. Without loss of generality, we let \(\sigma_e^2 = \sigma_g^2 = 1\).

For SNPs with the same MAF, the critical values for statistical significance are the same, and they are obtained by the following procedures. To get more precise results, we simulate the genotypes of 10 SNPs with that MAF. For each SNP, we generate the test statistics under the null hypothesis to estimate its null distribution, and the corresponding critical value is the \(100(1 - \alpha)\)% percentile, where \(\alpha\) is the significance level. For the comparison of type I error rates, \(\alpha\) is set to 0.05. For the comparison of power, \(\alpha\) is set to \(10^{-4}\). The number of replicates is set to \(\max(10000, 5/\alpha)\) to ensure the smoothness of the extreme percentiles. By averaging the 10 critical values, we have the final critical value for that MAF. Note that the computations for null test statistics only involve arithmetic operations so the computing time for generating the null distributions is affordable. Further, if we make a grid for the MAF, we can precompute the critical values for each grid so it’s at hand for any SNP in the following hypothesis testing. The number of simulation is 1000 for all configurations.
3.1.1 Type I error rates

We use 5,000 replications to evaluate the type I error rates at each configuration. Table 2 shows the results across the different tests. From this table, it’s seen that the generalized $F$-test has a good control on the type I error rates as they are all close to the nominal level. When $N = 500$, LMi shows moderate inflation of type I error and maintains good control for doubled sample size. It may be because the top principal components couldn’t capture all the information of GRM especially when the sample size is not large. As C-LRT does not allow the adjustment for population stratification, it shows serious inflation of type I error rates.

Table 2: Type I error rates across VC-LMM, LMi and C-LRT based on 5000 simulations

<table>
<thead>
<tr>
<th>$N$</th>
<th>$p_A$</th>
<th>VC-LMM</th>
<th>LMi</th>
<th>C-LRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>0.05</td>
<td>0.0528</td>
<td>0.0550</td>
<td>0.0548</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.0512</td>
<td>0.0546</td>
<td>0.0582</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.0518</td>
<td>0.0558</td>
<td>0.0596</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.0510</td>
<td>0.0566</td>
<td>0.0656</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.0506</td>
<td>0.0530</td>
<td>0.0624</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.0520</td>
<td>0.0530</td>
<td>0.0554</td>
</tr>
<tr>
<td>1000</td>
<td>0.05</td>
<td>0.0518</td>
<td>0.0492</td>
<td>0.0698</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.0470</td>
<td>0.0520</td>
<td>0.0632</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.0496</td>
<td>0.0500</td>
<td>0.0682</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.0502</td>
<td>0.0534</td>
<td>0.0726</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.0500</td>
<td>0.0532</td>
<td>0.0746</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.0468</td>
<td>0.0504</td>
<td>0.0798</td>
</tr>
</tbody>
</table>
3.3.2 Power

For power comparison, the $G \times E$ interaction effect is modeled by a sine function $f(E) = c \cdot \sin(\pi E)$, where $c$ is a constant controlling the effect size. For a fixed heritability $h^2$, $c$ can be obtained by the following formula:

$$h^2 = \frac{\text{var}(\text{Genetic effect})}{\text{var}(\text{Phenotype})}$$

$$= \frac{2f^2(E)p_A(1-p_A)}{2f^2(E)p_A(1-p_A) + \sigma_e^2 + \sigma_g^2}.$$

To simplify the computation, $f^2(E)$ is estimated by its expectation. After some algebraic manipulation, we have

$$c = \sqrt{\frac{2h^2}{p_A(1-p_A)(1-h^2)}}. \quad (11)$$

In the simulation, $h^2$ is set to 0.01, 0.03 and 0.05. From (11), we see that $c$ increases significantly as $p_A$ decreases from 0.1, so we penalize it for $p_A \leq 0.1$ by shrinking $h^2$ to be $h^2p_A/0.2$.

We present the power comparison in Figure 3.1 and Figure 3.2. Figure 3.1 shows the results for the common variants ($p_A \geq 0.2$), and we find that our VC-LMM model outperforms the other two in any configuration. The biggest gap between VC-LMM and C-LRT is at $N = 500, p_A = 0.3, h^2 = 0.03$, where the power of VC-LMM is 154% more than C-LRT. Since C-LRT has inflation of type I errors, the actual gain in power of VC-LMM should be even greater. The results for the low frequency variants ($p_A \leq 0.1$) are depicted in Figure 3.2. We see that the
Figure 3.1: Power of VC-LMM (solid black line), LMi (dotted blue line) and C-LRT (dashed red line) for common variants based on 1,000 replications.

Figure 3.2: Power of VC-LMM (solid black line), LMi (dotted blue line) and C-LRT (dashed red line) for low frequency variants based on 1,000 replications.
comparative results remain the same, and VC-LMM has good power when \( N = 1000 \) and \( pA = 0.1 \).

### 3.4 Discussion

With the advent of the high-throughput sequencing technologies, genetic data can now be generated almost with no limits. Thus, it’s getting more important to understand and take the most advantage of these data. Recently, \( G \times E \) interaction has received considerable attention in understanding the biological etiology for human traits in GWAS. In this study, we proposed a novel procedure, VC-LMM, to detect nonlinear \( G \times E \) interaction, while the population stratification can be accounted for by the GRM in a linear mixed model.

Computational burden is a big issue in GWAS especially when the whole-genome screen is of interest. Fortunately, the proposed procedure VC-LMM has good computational properties. First, the null distribution of test statistic is calculated by arithmetic operations, which are extremely fast. Since it only depends on MAF, we can generate the null distributions for a grid of MAFs beforehand to further expedite the analysis. Second, the linear mixed model fitting can be done by the algorithm used in FaST-LMM (Lippert et al., 2011), which significantly reduces the computing time.

In our simulation, the environmental factor is continuous. One possible extension of our model is investigating the interaction with categorical environmental factor. Naturally, this work can then be extended to gene-gene interaction, since genotypes are also categorical. Another potential direction is applying the generalized linear mixed model framework to study the \( G \times E \) interaction for binary traits, such as diseases.
DIFFERENTIATING THE COCHRAN-ARMITAGE TREND TEST AND PEARSON’S $\chi^2$ TEST: LOCATION AND DISPERSION

A.1 Equivalence between the CA trend test statistic $T_{CA}$ and the two-sample mean test statistic $T_{\text{mean}}$

In $T_{\text{mean}}$, the numerator $(\bar{X} - \bar{Y})^2 = \left( \frac{\sum_{i=0}^x x_i r_i}{R} - \frac{\sum_{i=0}^x x_i s_i}{S} \right)^2 = \sum_{j=0}^x x_j \left( \frac{r_j - \bar{s}_j}{S} \right)^2 = \frac{\sum_{i=0}^x x_i(S_r_i - R_s_i)^2}{R^2S^2}$. Under the null hypothesis, $(r_0, r_1, r_2)$ and $(s_0, s_1, s_2)$ are independent trinomially distributed vectors with $p_i = q_i = p_i$, which are estimated as the homologous sample proportions $\hat{p}_i = \frac{n_i}{N}$. The variance of $(\bar{X} - \bar{Y})$ can be derived as

$$\text{Var}_0(\bar{X} - \bar{Y}) = \text{Var}_0 \left( \frac{\sum_{i=0}^x x_i r_i}{R} \right) + \text{Var}_0 \left( \frac{\sum_{i=0}^x x_i s_i}{S} \right) = \left( \frac{1}{R} + \frac{1}{S} \right) \left[ \sum_{i=0}^x x_i^2 p_i(1 - p_i) - 2 \sum_i x_i x_j p_i p_j \right]$$

Replacing $p_i$ with $\hat{p}_i$, we obtain $\text{Var}_0(\bar{X} - \bar{Y}) = \left[ \sum_{i=0}^x x_i^2 \hat{p}_i(1 - \hat{p}_i) - \sum_{i=0}^x x_i x_j \hat{p}_i \hat{p}_j \right] = \frac{N}{R} \left[ \sum_{i=0}^x x_i^2 \hat{p}_i - (\sum_{i=0}^x x_i \hat{p}_i)^2 \right]$. Thus

$$T_{\text{mean}} = \frac{(X - Y)^2}{\text{Var}_0(\bar{X} - \bar{Y})} = \left\{ \frac{\sum_{i=0}^x x_i(S_r_i - R_s_i)^2}{R^2S^2} \right\} \left\{ \frac{N^2 \sum_{i=0}^x x_i^2 n_i - (\sum_{i=0}^x x_i \hat{p}_i)^2}{R S N^2} \right\} = \frac{\sum_{i=0}^x x_i(S_r_i - R_s_i)^2}{R S N} \frac{N^2 \sum_{i=0}^x x_i^2 n_i - (\sum_{i=0}^x x_i \hat{p}_i)^2}{R S N^2} = T_{CA}.$$
A.2 Partitioning $T_P$ by orthogonal polynomials

Define vectors $U_v = (V_{1v}, V_{2v})^T$, where $v \in \{1,2\}$, $N_1 = (r_0, r_1, r_2)^T$, $N_2 = (s_0, s_1, s_2)^T$, and $p = (\frac{n_0}{N}, \frac{n_1}{N}, \frac{n_2}{N})^T$. Define matrices $H_{2 \times 3} = [g_u(x_i)]$ and $H^*_{3 \times 3} = \begin{bmatrix} H \\ J \end{bmatrix}$, where $J = (1, 1, 1)$.

By the definition of $V_{uv}$, $U_1 = \frac{HN_1}{\sqrt{R}}$ and $U_2 = \frac{HN_2}{\sqrt{S}}$. By the properties of the orthonormal polynomials, $Hp = \left( \sum_{i=0}^{2} \frac{n_i g_1(x_i)}{N}, \sum_{i=0}^{2} \frac{n_i g_2(x_i)}{N} \right)^T = (0,0)^T$ and $H^* diag(p)H^{*T} = l_{3 \times 3}$. By some matrix manipulation, the latter leads to $diag\left( \frac{1}{p} \right) = H^* H^* = H^T H + J^T J$. Therefore, $U_1^T U_1 = \frac{N_1^T H^T H N_1}{R} = \frac{(N_1 - Rp)^T H^T H (N_1 - Rp)}{R} = \frac{\left( \sum_{i=0}^{2} \frac{n_i g_1(x_i)}{N} - \frac{n_i R}{N} \right)^2}{\sum_{i=0}^{2} \frac{n_i R}{N}^2}$.

Similarly, $U_2^T U_2 = \sum_{i=0}^{2} \frac{(s_l - n_l S/N)^2}{n_l S/N}$. Thus $T_P = U_1^T U_1 + U_2^T U_2 = V_{11}^2 + V_{12}^2 + V_{21}^2 + V_{22}^2$. 

38
APPENDIX B

DECOMPOSING PEARSON’S $\chi^2$ TEST: A LINEAR REGRESSION AND ITS DEPARTURE FROM LINEARITY

B.1 Proof of $T_P = T_{CA} + T_{SSE}$

In genetic association studies, there is no single Pearson’s $\chi^2$ statistic $T_P =$

$$
\sum_{i=0}^2 \frac{(r_i - n_i R / N)^2}{n_i R / N} + \sum_{i=0}^2 \frac{(s_i - n_i S / N)^2}{n_i S / N} = \sum_{i=0}^2 \frac{n_i (p_i^o - p^o)^2}{p^o} + \sum_{i=0}^2 \frac{n_i (p_i^o - p^o)^2}{1 - p^o} = \frac{1}{p^o (1 - p^o)} \sum_{i=0}^2 n_i (p_i^o -
$$

$p^o)^2.$ Given $\hat{p}_i = p^o + \hat{\theta} (x_i - \bar{x})$ and $\hat{\theta} = \frac{\sum_{i=0}^2 n_i (p_i^o - p^o)(x_i - \bar{x})}{\sum_{i=0}^2 n_i (x_i - \bar{x})^2},$ by some algebra manipulations, it can be shown $\sum_{i=0}^2 n_i (p_i^o - \hat{p}_i) (\hat{p}_i - p^o) = 0.$ Therefore, $T_P = \frac{1}{p^o (1 - p^o)} \sum_{i=0}^2 n_i (p_i^o - \hat{p}_i + \hat{p}_i -

p^o)^2 = \frac{1}{p^o (1 - p^o)} \sum_{i=0}^2 n_i [(p_i^o - \hat{p}_i)^2 + (\hat{p}_i - p^o)^2 + 2(p_i^o - \hat{p}_i) (\hat{p}_i - p^o)] =

\frac{1}{p^o (1 - p^o)} \sum_{i=0}^2 n_i (\hat{p}_i - p^o)^2 + \frac{1}{p^o (1 - p^o)} \sum_{i=0}^2 n_i (p_i^o - \hat{p}_i)^2 \quad (*) \text{Again by some algebra}

manipulations, it can be shown $\frac{1}{p^o (1 - p^o)} \sum_{i=0}^2 n_i (\hat{p}_i - p^o)^2 = \frac{N(\sum_{i=0}^2 n_i x_i - R \sum_{i=0}^2 n_i x_i)^2}{N \sum_{i=0}^2 n_i x_i (\sum_{i=0}^2 x_i n_i)^2}$ which is exactly the CA trend test statistic $T_{CA}.$ Define $T_{SSE} = \frac{1}{p^o (1 - p^o)} \sum_{i=0}^2 n_i (p_i^o - \hat{p}_i)^2.$ Therefore equation $(*)$ is identical to $T_P = T_{CA} + T_{SSE}.$
Figure B.2: power comparison of the Cochran-Armitage trend test by the additive, multiplicative, and recessive models under different modes of inheritance

On the simulation models, the relative risks for $Aa$ and $AA$ are defined as $\lambda_1 = 1 + \cos\theta$ and $\lambda_2 = 1 + \sin\theta$, respectively. Therefore, along the $x$-axis $\theta$ ranges from $\arctan 2$, the additive model, to $\pi/2$, the recessive model; on the analysis models, the red, blue, and black lines denote the additive, multiplicative, and recessive models, respectively. The $y$-axis is the empirical power based on 10,000 replicates at the 0.05 level. The disease prevalence equals 0.1; the minor allele frequency $q \in (0.05, 0.1, 0.15, 0.2)$; and the total sample size $N \in (300, 500, 800, 1000)$ with equal numbers of cases and controls.
C.1 Algorithm for estimation of parameters

Under the alternative hypothesis,

\[ Y \sim N(D\rho, \sigma_{\theta}^2\Phi + \sigma_b^2Z_2Z_2^T + \sigma_e^2I). \]

Rewrite \( \text{var}(Y) \) as \( \sigma^2(h^2L + (1 - h^2)I) \), where \( \sigma^2 = \sigma_{\theta}^2 + \sigma_b^2 + \sigma_e^2 \), \( h^2 = \frac{\sigma_{\theta}^2 + \sigma_b^2}{\sigma_{\theta}^2 + \sigma_b^2 + \sigma_e^2} \), \( L_c = \Phi + \text{mixing}Z_2Z_2^T \) and \( \text{mixing} = \frac{\sigma_b^2}{\sigma_{\theta}^2 + \sigma_b^2} \). Let \( USU^T = L_c \) be the spectral decomposition of \( L_c \) (mixing). Note that \( UU^T = I \) and \( S \) is a diagonal matrix whose elements are the eigenvalues of \( L_c \). Let \( U \) be a rotation matrix and apply it on the phenotypes \( Y \), then

\[ U^TY \sim N(U^TD\rho, \sigma^2(h^2S + (1 - h^2)I)). \]

Now the variance of \( U^TY \) is a diagonal matrix, which makes the computation much easier.

Based on the distribution of \( U^TY \), the restricted log likelihood parameterized by \( \sigma^2, h^2, \text{mixing} \) and \( \rho \) is
\[
\text{ReLL}(\sigma^2, h^2, \text{mixing}, \rho)
\]
\[
= \text{const.} - \frac{1}{2} \left[ (n - (p^c + h + 1)) \log(\sigma^2) + \log|h^2S + (1 - h^2)I| \\
+ \frac{1}{\sigma^2}(U^TY - U^TD\rho)^T(h^2S + (1 - h^2)I)^{-1}(U^TY - U^TD\rho) \\
+ \log|(U^TD)^T(h^2S + (1 - h^2)I)^{-1}U^TD| \right]
\]
\[
= \text{const.} - \frac{1}{2} \left[ (n - (p^c + h + 1)) \log(\sigma^2) + \sum_{i=1}^{n} \log(h^2[diag(S)]_i + (1 - h^2)) \\
+ \frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{([U^TY]_i - [U^TD]_i; \rho)^2}{h^2[diag(S)]_i + (1 - h^2)} + \log|(U^TD)^T(h^2S + (1 - h^2)I)^{-1}U^TD| \right],
\]

where \([\cdot]_i\) denotes the \(i\)th diagonal element of a vector and \([\cdot]_{i,:}\) denotes the \(i\)th row of a matrix.

Take derivative of the first representation of \text{ReLL} with respect to \(\rho\) and set it to 0, we have
\[
(U^TD)^T(h^2S + (1 - h^2)I)^{-1}U^TY - (U^TD)^T(h^2S + (1 - h^2)I)^{-1}U^TD\hat{\rho} = 0,
\]
so
\[
\hat{\rho} = [(U^TD)^T(h^2S + (1 - h^2)I)^{-1}U^TD]^{-1}(U^TD)^T(h^2S + (1 - h^2)I)^{-1}U^TY
\]
\[
= \left[ \sum_{i=1}^{n} \frac{[U^TD]_{i,:}}{h^2[diag(S)]_i + (1 - h^2)} \right]^{-1} \left[ \sum_{i=1}^{n} \frac{[U^TD]_{i,:}}{h^2[diag(S)]_i + (1 - h^2)} \right]^{T} \left[ \sum_{i=1}^{n} \frac{[U^TY]_i}{h^2[diag(S)]_i + (1 - h^2)} \right] \left[ [U^TD]_{i,:} \right]^{T} \left[ [U^TY]_i \right].
\]

Then substitute \(\rho\) with \(\hat{\rho}\) and take derivative of the second representation of \text{ReLL} with respect to \(\sigma^2\) and set it to 0, we have
\[
\frac{(n - (p_c + h + 1))}{\hat{\sigma}^2} - \frac{1}{\hat{\sigma}^4} \sum_{i=1}^{n} \frac{([U^TY]_i - [U^TD]_i\hat{\rho})^2}{h^2[diag(S)]_i + (1 - h^2)} = 0,
\]

so

\[
\hat{\sigma}^2 = \frac{1}{n - (p_c + h + 1)} \sum_{i=1}^{n} \frac{([U^TY]_i - [U^TD]_i\hat{\rho})^2}{h^2[diag(S)]_i + (1 - h^2)}.
\]

Plug in \(\hat{\rho}\) and \(\hat{\sigma}^2\), the restricted log likelihood is only a function of parameter \(h^2\) and hidden parameter \(mixing\). Using the python package “fastlmm”, we can readily obtain the pair \((h^2, mixing)\) maximizing the \(Rell\) by searching the grid from 0 to 1 with Brent’s method.

With estimates \(\hat{\sigma}^2, \hat{h}^2\) and \(mixing\), we finally have

\[
\hat{\gamma} = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_e^2} = \frac{\hat{h}^2(1 - mixing)}{1 - \hat{h}^2} \text{ and }
\]

\[
\hat{\xi} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_e^2} = \frac{\hat{h}^2 mixing}{1 - \hat{h}^2}.
\]
REFERENCE


Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philosophical Magazine, 50(302), 157-175.


