Genetic association analysis

Jan Graffelman[1]

[1]Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Barcelona, Spain

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

jan.graffelman@upc.edu

December 13, 2018

Contents

Genetic association studies

Goal:

- Investigate associations between markers and a trait (disease).

Designs:

- Unrelated subjects (population-based)
- Related subjects from pedigrees (family-based)

We will focus on population-based association studies

Preliminaries

- The trait (Y_i) (e.g. disease) we wish to understand is binary (dichotomous).
- $Y_i = 1$ individual i has the trait, $Y_i = 0$, individual i does not have the trait.
- The marker is a bi-allelic polymorphism (e.g. AA, Aa and aa)

Levels of analysis

- Tests of association at the level of alleles
  - We are sampling alleles
  - Alleles assumed to be independent
  - Rely on the Hardy-Weinberg equilibrium assumption
  - Chi-square test of the alleles by trait cross table
  - Fisher exact test of the alleles by trait cross table
  - Test on the odds ratio of the alleles by trait cross table
- Tests of association at the level of the genotypes
  - We are sampling individuals
  - Hardy-Weinberg equilibrium assumption is not needed
  - Co-dominant, dominant and recessive Chi-square tests
  - Cochran-Armitage trend test
  - Logistic regression

The data table

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

The alleles test

- Let p be the allele frequency of the A allele.

$$\begin{cases} H_0 : p_{cases} = p_{controls} \\ H_1 : p_{cases} \neq p_{controls} \end{cases}$$

- The test assumes Hardy-Weinberg equilibrium
- The test is a χ^2 test for independence in a 2×2 table of alleles.

	a	A	Total	$\hat{p}$
Cases	$r_a = 2r_0 + r_1$	$r_A = 2r_2 + r_1$	$2r$	$r_A/(2r)$
Controls	$s_a = 2s_0 + s_1$	$s_A = 2s_2 + s_1$	$2s$	$s_A/(2s)$
Total	$n_a = 2n_0 + n_1$	$n_A = 2n_2 + n_1$	$2n$	$n_A/(2n)$

The alleles test

- Chi-square test for independence

$$X^2 = \sum_{i,j}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Expected count e_{ij} = total row $i \times$ total colum j/total of table
- If H_0 is true, then $X^2 \sim \chi^2_1$
- p-value $= P\left(\chi^2_1 \geq X^2\right)$

Example alleles test

A polymorphism in the Dopamine receptor is supposed to be involved in Schizophrenia. In a case-control study, the following data were obtained:

	11	12	22	Total
Cases	7	69	57	133
Controls	20	56	33	109
Total	27	125	90	242

	1	2	Total
Cases	83	183	266
Controls	96	122	218
Total	179	305	484

Example alleles test

	1	2	Total
Cases	83	183	266
Controls	96	122	218
Total	179	305	484

	1	2	Total
Cases	98.38	167.62	266
Controls	80.62	137.38	218
Total	179	305	484

$$X^2 = \frac{(83 - 98.38)^2}{98.38} + \cdots + \frac{(122 - 137.38)^2}{137.38} = 8.4671$$

$$\text{p-value} = P\left(\chi_1^2 \leq 8.4671\right) = 0.0036$$

R code alleles test

```
> X <- matrix(c(7,69,57,20,56,33),byrow=TRUE,ncol=3)
> colnames(X) <- c("11","12","22")
> rownames(X) <- c("Cases","Controls")
> X
         11 12 22
Cases     7 69 57
Controls 20 56 33
> Y <- cbind(2*X[,1]+X[,2],2*X[,3]+X[,2])
> colnames(Y) <- c("1","2")
> Y
          1   2
Cases    83 183
Controls 96 122
> chisq.test(Y,correct=FALSE)

Pearson's Chi-squared test

data:  Y
X-squared = 8.4671, df = 1, p-value = 0.003616
```

Fisher's Exact test

- Often used for cross tables with low counts in the margin, or when $e_{ij} < 5$.
- If the margins are considered fixed, the probability of the table can be calculated, using the hypergeometric distribution.
- The exact p-value is the sum of the probabilities of all possible tables with the same margins that have a probability that is less or equal than the observed table.

For the same data:

```
> Y
          1   2
cases    83 183
controls 96 122
> fisher.test(Y)

Fisher's Exact Test for Count Data

data:  Y
p-value = 0.00448
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.3903016 0.8512261
sample estimates:
odds ratio
 0.5770451
```

Odds ratio test for alleles

- Definition of odds

$$\text{Odds} = \frac{P(\text{success})}{P(\text{failure})} = \frac{P(\text{disease})}{P(\text{no disease})} = \frac{p}{1-p}$$

- The Odds ratio (OR) compares the odds of the disease for the two alleles:

$$OR = \frac{\text{Odss of disease with A allle}}{\text{Odss of disease with B allle}}$$

	A	B
Cases	n_{11}	n_{12}
Controls	n_{21}	n_{22}

$$OR = \frac{(n_{11}/n_{21})}{(n_{12}/n_{22})} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{21}}$$

- An odds ratio based test assumes an additive model: AA doubles the risk of AB.
- OR $= 1$ corresponds to independence; OR > 1 or OR < 1 implies association.
- Known result:

$$V(\ln(OR)) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

- Allows calculation of confidence intervals for the OR.

Odds ratio test for alleles

	1	2	Total
Cases	83	183	266
Controls	96	122	218
Total	179	305	484

$$OR = \frac{83 \cdot 122}{96 \cdot 183} = 0.5764$$

$$se_{ln(OR)} = \sqrt{\frac{1}{83} + \frac{1}{183} + \frac{1}{96} + \frac{1}{122}} = 0.1900$$

$$CI(\text{True ln(OR)}) = \ln(OR) \pm z_{\alpha/2} se_{ln(OR)}$$

$$CI(\text{True OR}) = e^{\ln(OR) \pm z_{\alpha/2} se_{ln(OR)}}$$

$$CI(\text{True OR}) = e^{\ln(0.5764) \pm 1.96 \cdot 0.1900} = (0.397; 0.837)$$

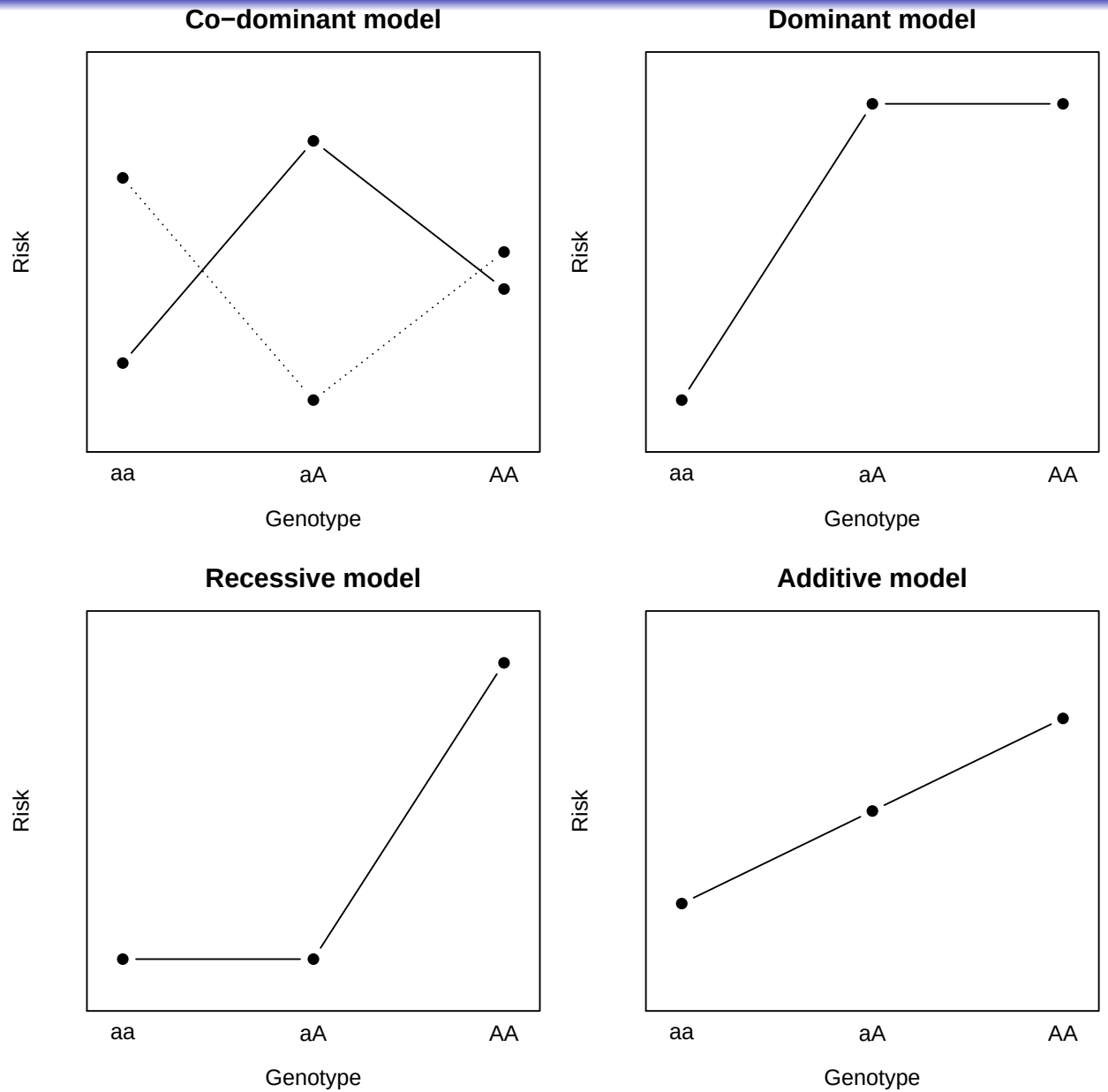
The data table

	aa	aA	AA	Total
Cases	r_0	r_1	r_2	r
Controls	s_0	s_1	s_2	s
Total	n_0	n_1	n_2	n

We can test for association using different genetic models:

- A codominant model
- A dominant model
- A recessive model
- An additive model

Genetic association models



Codominant test

- We test the null hypothesis of no effect of the marker on the trait.
- Formally:

$$\begin{cases} H_0 : P(Y = 1|AA) = P(Y = 1|Aa) = P(Y = 1|aa) \\ H_1 : \text{At least one pair different} \end{cases}$$

- Test statistic

$$X^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Under H_0, we have $X^2 \sim \chi^2_2$
- The test makes no assumptions about the relationship between genotype and trait.
- Under H_1, each genotype can have a different disease rate.
- The test can reject the null if the data support heterozygote advantage (overdominance).

Example codominant test

TNF genotype (G/A polymorphism) in a study on acne patients and controls

	GG	GA	AA	Total
Cases	66	43	4	113
Controls	99	15	0	114
Total	165	58	4	227

R code codominant test

```
> X <- matrix(c(66,43,4,99,15,0),byrow=TRUE,ncol=3)
> colnames(X) <- c("GG","GA","AA")
> rownames(X) <- c("Acne","Contro")
> X
       GG GA AA
Acne   66 43  4
Contro 99 15  0
> results <- chisq.test(X)
Warning message:
In chisq.test(X) : Chi-squared approximation may be incorrect
> print(results)

Pearson's Chi-squared test

data:  X
X-squared = 24.1133, df = 2, p-value = 5.806e-06

> results$expected
             GG       GA       AA
Acne   82.13656 28.87225 1.991189
Contro 82.86344 29.12775 2.008811

> fisher.test(X)

Fisher's Exact Test for Count Data

data:  X
p-value = 1.97e-06
alternative hypothesis: two.sided
```

Dominant test

- Columns in the original table are combined to produce 2×2 tables.
- Dominant model:

	aa	aA or AA	Total
Cases	r_0	$r_1 + r_2$	r
Controls	s_0	$s_1 + s_2$	s
Total	n_0	$n_1 + n_2$	n

- Test:

$$\begin{cases} H_0 : \text{Disease does not depend on presence of A} \\ H_1 : \text{Disease does depend on the presence of A} \end{cases}$$

- Statistic:

$$X^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Under H_0, we have $X^2 \sim \chi^2_1$

R code dominant test

```
> Y <- cbind(X[,1],X[,2]+X[,3])
> colnames(Y) <- c("GG","GA or AA")
> rownames(Y) <- c("Acne","Control")
> Y
        GG GA or AA
Acne    66       47
Control 99       15
> results <- chisq.test(Y)
> print(results)

Pearson's Chi-squared test with Yates' continuity correction

data:  Y
X-squared = 21.7021, df = 1, p-value = 3.184e-06

> results <- chisq.test(Y,correct=FALSE)
> print(results)

Pearson's Chi-squared test

data:  Y
X-squared = 23.1122, df = 1, p-value = 1.528e-06

>
```

Recessive test

- Recessive model:

	aa or aA	AA	Total
Cases	$r_0 + r_1$	r_2	r
Controls	$s_0 + s_1$	s_2	s
Total	$n_0 + n_1$	n_2	n

- Test:

$$\begin{cases} H_0 : \text{Disease does not depend on being homozygote AA} \\ H_1 : \text{Disease does depend on being homozygote AA} \end{cases}$$

- Statistic:

$$X^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

- Under H_0, we have $X^2 \sim \chi^2_1$

The additive genetic model

- Basic idea: disease risk increases as a function of the number of alleles (0,1 or 2).
- There are two tests for the additive genetic model
  - The alleles test
  - Cochran-Armitage trend test

Armitage trend test

- The trend test is based on the linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

- X is the disease status (0 or 1)
- Y is the number of A alleles (0, 1 or 2)
- Tests $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$
- Armitage trend test statistic

$$A = \frac{\hat{\beta}_1^2}{V\left(\hat{\beta}_1\right)} = n \cdot r_{xy}^2$$

- Under H_0, $A \sim \chi_1^2$.

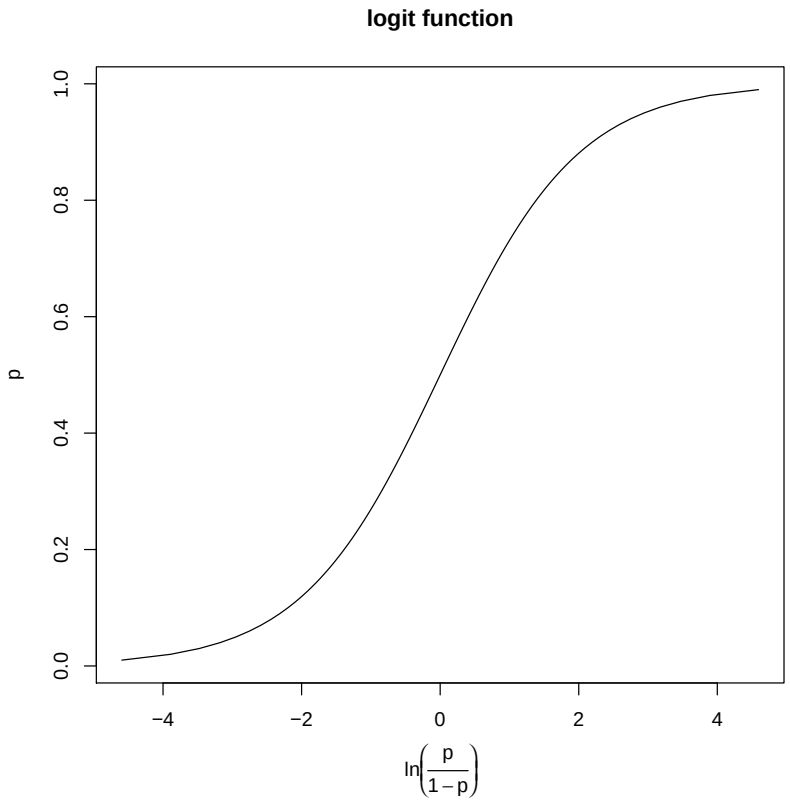
Example Armitage trend test

	GG	GA	AA	Total
Cases	66	43	4	113
Controls	99	15	0	114
Total	165	58	4	227

$$A = 227 \cdot (0.3253)^2 = 24.02$$

$$P\left(\chi_1^2 \geq 24.02\right) = 9.49e - 07$$

Logistic regression



Logit (or logistic) function:

$$logit(\pi) = \frac{\pi}{1-\pi}$$

Inverse of the logit function

$$logit^{-1}(\pi) = \frac{e^{\pi}}{e^{\pi}+1}$$

Using logit(π) as the response is the basis of logistic regression

The logistic regression model

$$\pi(x) = E(Y|x) = P(Y = 1|x)$$

$$y = \pi(x) + \varepsilon \quad y \sim Bin(n, \pi(x))$$

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{e^{\beta_0+\beta_1 x} + 1}$$

Note that

- $0 \leq \pi(x) \leq 1$
- $-\infty \leq g(x) \leq +\infty$

Model and likelihood

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left[1 - \pi(x_i)\right]^{1-y_i}$$

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} \{y_i \ln \left[\pi(x_i)\right] + (1 - y_i) \ln \left[1 - \pi(x_i)\right]\}$$

We maximize $l(\beta_0, \beta_1)$ by numerical methods

Odds ratios and logistic regression with genetic predictors

- One genotype is the reference genotype (e.g. AA)
- Of interest are the odds ratios

$$OR_{BB} = \frac{\text{Odss disease for a BB person}}{\text{Odds disease AA person}}$$

$$OR_{AB} = \frac{\text{Odss disease for a AB person}}{\text{Odds disease AA person}}$$

- These OR are estimated by logistic regression.
- Logistic regression is attractive as it allows to adjust for covariates.
- Model

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_{AB} I_{AB} + \beta_{BB} I_{BB}$$

- $OR_{AB} = e^{\hat{\beta}_{AB}}$ and $OR_{BB} = e^{\hat{\beta}_{BB}}$

Example logistic regression in R

```
Cases    <- c(GG=66,AG=43,AA=4)
Controls <- c(GG=99,AG=15,AA=0)

cas <- rep(c("GG","AG","AA"),Cases)
con <- rep(c("GG","AG","AA"),Controls)

ncas <- length(cas)
ncon <- length(con)

y <- c(rep(1,ncas),rep(0,ncon))
x <- factor(c(cas,con))

out.lm <- glm(y~x, family = binomial(link = "logit"))
summary(out.lm)

or <- exp(coefficients(out.lm))
```

Computer exercise

- A particular SNP is supposed to be involved in Alzheimer's disease. A case control study has been performed, obtaining the following results:

	MM	Mm	mm
Cases	112	278	150
Controls	206	348	150

- Perform the alleles test for this data set.
- Perform Armitage trend test for this data set.
- Plot the risk of disease as a function of the number of m alleles. Fit a linear model and add the regression line to the plot. Test the null hypothesis $\beta_1 = 0$.
- Is there evidence for association of this marker with the disease?
- Also test for association using a codominant, a dominant and a recessive model.
- Which model seems most appropriate?
- Estimate odds ratios using logistic regression

Bibliography

- Laird, N.M. & Lange, C. (2011) The fundamentals of modern statistical genetics. Springer.