Statistical Interactions

University of Liege

Regression models for interactions

- Statistical interaction most easily described in terms a of regression framework

- Suppose X_1 and X_2 are two explanatory variables (factors) that may be associated with a response variable (outcome) Y.

- Regression models the main effects of X_1 and X_2 on Y as:

$$\mathrm{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

- Including interaction effect: the typical treatment of interactions in linear models is to consider the interaction as a product term of the main effect variables

$$\mathrm{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2)$$

Model assumptions

- Quantitative models always rest on assumptions about the way the world works, and regression models are no exception.

- There are four principal assumptions which justify the use of linear regression models for purposes of prediction:
  - linearity of the relationship between dependent and independent variables
  - independence of the errors (no serial correlation)
  - constant variance (homoscedasticity) of the errors
  - normality of the error distribution.

Violation of assumptions

If any of these assumptions is violated (i.e., if there is nonlinearity, serial correlation, heteroscedasticity, and/or non-normality), then

- the forecasts,
- confidence intervals,
- and insights

yielded by a regression model may be (at best) inefficient or (at worst) seriously biased or misleading.

Model diagnostics

- Diagnostic methods can be graphical or numerical.
- Graphical methods tend to be more versatile and informative.
- It is virtually impossible to verify that a given model is exactly correct. The purpose of the diagnostics is more to check whether the model is not grossly wrong.
- Diagnostic plots
  - Plot of residuals against fitted values. This plot can be used to detect lack of fit and to check the constant variance assumption on the errors.
  - QQ plot of residuals to assess normality.
  - Cook's statistics: used for identifying influential observations.

Generalized linear models

- Model

$$g\big(E(Y|X)\big) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

where g is a link function

- The response (outcome) variable Y can be continuous, count, or categorical.
- The distribution of the response variable can be other than the Normal distribution.
  - Continuous outcome and normally distributed → linear regression
  - Binary outcome → logistic regression
  - Count outcome → Poisson regression
  - Categorical outcome → Multinomial regression

Regression analysis in R

- The basic syntax for doing regression in R is

  lm(response ~ covariates) to fit linear models and

  glm(response ~ covariates) to fit generalized linear models.

- Special type of models you can fit using glm() are

  linear regression

  logistic regression

  Poisson regression

GLM fitting in R

- R – code

glm(formula, data=datafile, family=familytype(link=linkfunction))

- formula : y ~ x1 + x2 + x1*x2 (in short y ~ x1*x2)

- data: data file containing the response and explanatory variables.

- family and default link function
  binomial (link = "logit")
  gaussian (link = "identity")
  poisson (link = "log")

Covariates/explanatory variables

In regression models

- Covariates can be continuous, count, binary or categorical.

- Coding is required for categorical variables.

Genetic Syndromes & Gene Therapy

Hartwig, J Genet Syndr Gene Ther 2013, 4:9
http://dx.doi.org/10.4172/2157-7412.1000189

Review Article Open Access

SNP-SNP Interactions: Focusing on Variable Coding for Complex Models of Epistasis

Fernando Pires Hartwig*

Postgraduate Program in Epidemiology, Department of Social Medicine, Faculty of Medicine, Federal University of Pelotas, Brazil

Abstract

Genetic epidemiology is a promising field to identify patterns of disease susceptibility that can be explored in personalized medicine. However, especially for complex traits, the genetic component is likely to be composed of several loci and/or of interactions between them. The last is addressed in this manuscript, which aims to provide an overview of the advantages and disadvantages of statistically-oriented and biologically-oriented approaches for two-SNP interactions. Eight biologically-oriented models of epistasis are discussed, focusing on their implementation, which is exemplified with real data. Additionally, some key technical points (such as reducing statistical power due to multiple testing and use of conceptual considerations) are discussed, and an exploratory step prior to the analysis is proposed to pre-select the models of epistasis to be actually tested. A function (written in R) is provided (under request) to facilitate the implementation of such models (and can be easily modified to implement others). It is stressed that, regardless of the method choice, the biological meaning of the model being tested is critical for correct interpretation of the results.

Genotypes coding schemes

Genotypes*	Genetic models†				
	Genotypic	Overdominant	Dominant	Recessive	Additive
AA	0	0	0	0	0
Aa	1	1	1	0	1
aa	2	0	1	1	2

*A: wild type allele at locus *A*; a: variant allele at locus *A*; †Each column represents the coding system for defining five distinct genetic models

Table 1: Variable coding to define five genetic models for a single SNP.

Categorical

binary

count/ordered

Statistical interactions coding schemes

Combined	Models of statistical interaction†				
genotypes*	Gen-Gen	Over-Over	Dom-Dom	Rec-Rec	Add-Add
AA/BB	0	0	0	0	0
AA/Bb	1	0	0	0	0
AA/bb	2	0	0	0	0
Aa/BB	3	0	0	0	0
Aa/Bb	4	1	1	0	1
Aa/bb	5	0	1	0	2
aa/BB	6	0	0	0	0
aa/Bb	7	0	1	0	2
aa/bb	8	0	1	1	4

*A, B: wild type alleles at loci *A* and *B*, respectively; a, b: variant alleles at loci *A* and *B*, respectively; †Each column represents a variable coding system that is equivalent to including an interaction term in a regression analysis having both SNPs in the specified genetic models; Gen-Gen: Genotypic-Genotypic; Over-Over: Overdominant-Overdominant; Dom-Dom: Dominant-Dominant; Rec-Rec: Recessive-Recessive; Add-Add: Additive-Additive; Of these, only the Add-Add model is actually quantitative (the numbers in the other models represent categories).

Table 2: Statistically-intuitive models of SNP-SNP interactions.

Genotype combinations of two SNPs

SNP2 (X2)

SNP1 (X1)	AA=0	Aa=1	aa=2
BB=0	00	01	02
Bb=1	10	11	12
bb=2	20	21	22

Main effects and interactions

Additive Coding

cell	X1	X2	X1*X2
00	0	0	0
01	0	1	0
02	0	2	0
10	1	0	0
11	1	1	1
12	1	2	2
20	2	0	0
21	2	1	2
22	2	2	4

Main effects and interactions

Co-dominant coding

	X1		X2		X1*X2			
cell	X11	X12	X21	X22	X11*X21	X11*X22	X12*X21	X12*X22
00	0	0	0	0	0	0	0	0
01	0	0	1	0	0	0	0	0
02	0	0	0	1	0	0	0	0
10	1	0	0	0	0	0	0	0
11	1	0	1	0	1	0	0	0
12	1	0	0	1	0	1	0	0
20	0	1	0	0	0	0	0	0
21	0	1	1	0	0	0	1	0
22	0	1	0	1	0	0	0	1

Logistic regression

- Logistic regression models their effect on the log odds of disease as:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1$$

Marginal effect of factor 1

$$\log \frac{p}{1-p} = \beta_0 + \beta_2 x_2$$

Marginal effect of factor 2

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Main effects of factors 1 and 2

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

Main effects and interaction term

Factors (SNPs) from dominant or recessive models

- Expected trait values (log odds of disease) take the form:

	Factor 2	
Factor 1	1	0
1	$\beta_0 + \beta_1 + \beta_2 + \beta_{12}$	$\beta_0 + \beta_1$
0	$\beta_0 + \beta_2$	β_0

  - β_0, β_1, β_2, β_{12} are regression coefficients (numbers) that can be estimated from real data
    - Having factor 1 adds β_1 to your trait value
    - Having factor 2 adds β_2 to your trait value
    - Having both factors adds an additional β_{12} to your trait value

      $\Rightarrow$ Implies that the overall effect of two variables is greater (or less) than the 'sum of the parts'

Factors (SNPs) from co-dominant model

modelling log odds in terms of:

- A baseline effect (β_0)
- Main effects of locus G (β_{G_1}, β_{G_2})
- Main effects of locus H (β_{H_1}, β_{H_2})
- 4 interaction terms

	Locus H		
Locus G	2	1	0
2	$\beta_0+\beta_{G_2}+\beta_{H_2}+\beta_{22}$	$\beta_0+\beta_{G_2}+\beta_{H_1}+\beta_{21}$	$\beta_0+\beta_{G_2}$
1	$\beta_0+\beta_{G_1}+\beta_{H_2}+\beta_{12}$	$\beta_0+\beta_{G_1}+\beta_{H_1}+\beta_{11}$	$\beta_0+\beta_{G_1}$
0	$\beta_0+\beta_{H_2}$	$\beta_0+\beta_{H_1}$	β_0

- Corresponds in statistical analysis packages to coding x_1, x_2 (0,1,2) as a "factor"

GLM for interactions

Install packages:

```
install.packages(c("broom","MASS"), dep=TRUE)

source("http://www.bioconductor.org/biocLite.R")
biocLite(c("Biobase"))
```

Load packages

```
library(Biobase)
library(broom)
library(MASS)
```

Linear model in genetics

```
library(MatrixEQTL)
data(GE)
data(geneloc )

data(SNP)
data(snpsloc)

GE <- data.matrix(GE)
SNP <- data.matrix(SNP)

gex1 <- GE[6,-1]
snp1 <- SNP[4,-1]
snp3 <- SNP[5,-1]
```

#Linear model to see the effect of SNP by SNP interaction on eQTL

```
glm_exp  <-  glm(gex1~ snp1*snp3, family=gaussian)
Summary(glm_exp)
```

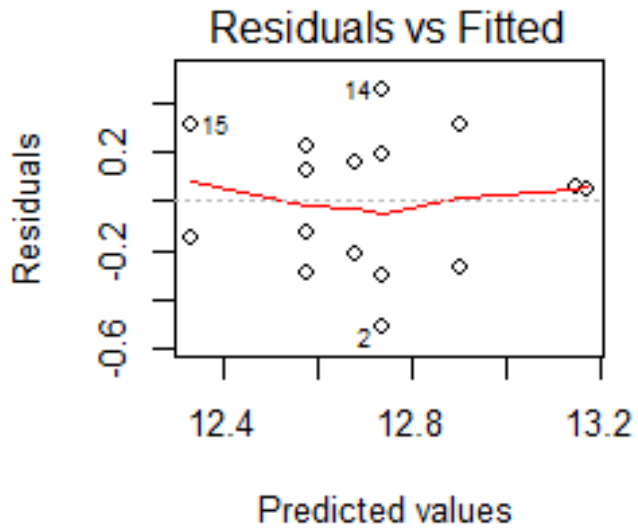
Coefficients:

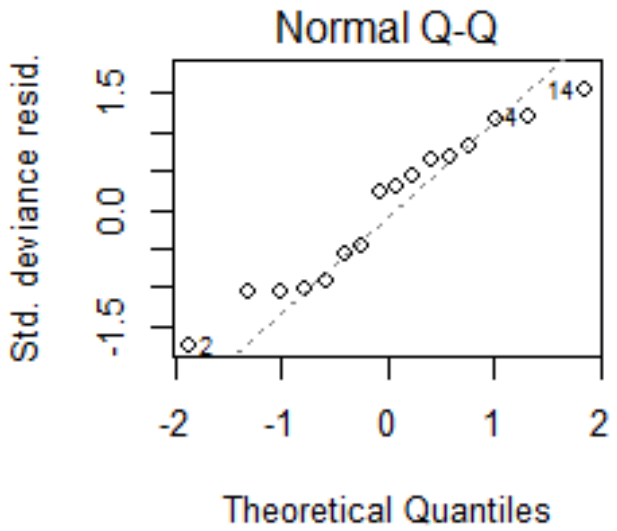
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.16905	0.25747	51.148	2.05e-15 ***
snp1	-0.02138	0.25266	-0.085	0.9340
snp3	-0.59188	0.25411	-2.329	0.0381 *
snp1:snp3	0.18448	0.20219	0.912	0.3795

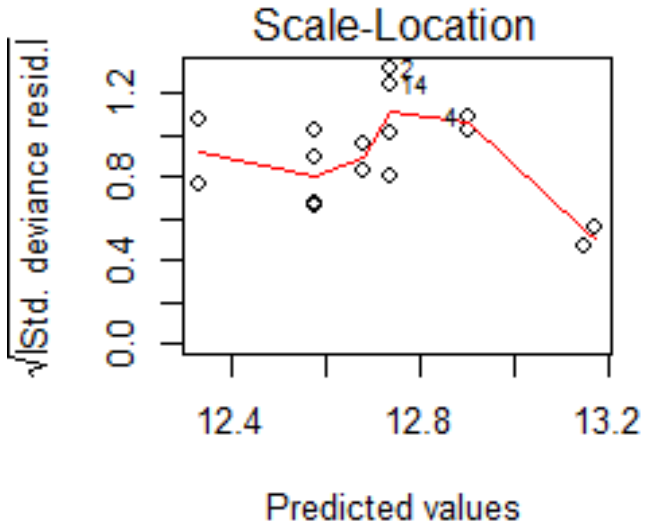
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

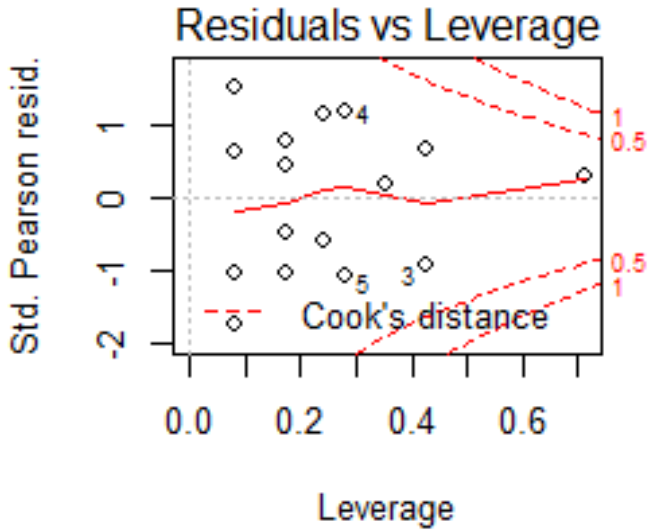
#Model Diagnostics

plot(glm_exp)









Load data files

```
#SNP data
snp_data <- read.table("epistasis_snps_example.txt")
head(snp_data)
dim(snp_data)

#RNA-seq data
rnaseq <- read.table("epistasis_rnaseq_example.txt")
dim(rnaseq)
```

Logistic regression for epistasis detection

We use example SNP data from a case-control genome-wide association study (snp_data)

Response variable: disease (case=1 and control=0)

Use logistic regression model

#Additive model

glm_add = glm(pheno ~ rs7909677*rs4880781, data=snp_data, family="binomial")

summary(glm_add)

Output

```
Call:
glm(formula = pheno ~ rs7909677 * rs4880781, family = "binomial",   data = snp_data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.456  -1.162  -1.073   1.193   1.381

Coefficients:
                    Estimate     Std. Error  z value  Pr(>|z|)
(Intercept)        -0.03583    0.08467  -0.423     0.672
rs7909677          -0.21534    0.24706  -0.872    0.383
rs4880781           0.09236    0.11202   0.824    0.410
rs7909677:rs4880781  0.35071    0.38164   0.919    0.358

(Dispersion parameter for binomial family taken to be 1)
 Null deviance: 1361.3  on 981  degrees of freedom
Residual deviance: 1358.9  on 978  degrees of freedom
  (18 observations deleted due to missingness)
AIC: 1366.9

Number of Fisher Scoring iterations: 3
```

The “broom” package takes the messy output of built-in functions in R, such as lm and glm, and turns them into tidy data frames

```
library(broom)

tidy(glm_add)
```

	term	estimate	std.error	statistic	p.value
1	(Intercept)	-0.03582730	0.08467467	-0.4231171	0.6722098
2	rs7909677	-0.21534232	0.24706118	-0.8716154	0.3834182
3	rs4880781	0.09235736	0.11202029	0.8244700	0.4096726
4	rs7909677:rs4880781	0.35070975	0.38163776	0.9189598	0.3581166

In this data, we have no evidence of additive interaction effect between SNPs rs7909677 and rs4880781 on the disease.

Confidence interval estimates

95% confidence interval

confint.default(glm_add)

	2.5 %	97.5 %
(Intercept)	-0.2017866	0.1301320
rs7909677	-0.6995733	0.2688887
rs4880781	-0.1271984	0.3119131
rs7909677:rs4880781	-0.3972865	1.0987060

```
pheno <- snp_data$pheno

population <- snp_data$population

snp <- snp_data[,3:102]

colnames(snp) <- paste("snp",1:100, sep="")
```

```
# Dominant model
#Coding (AA=0, Aa=1, aa=1)

snp_dom <- (snp == 2 | snp==1)*1
snp_dom <- data.frame(snp_dom)

#two-way table
table(snp_dom$snp1,snp_dom$snp2)
          0         1
  0      527     349
  1       70       36

glm_dom <- glm(pheno ~  snp1*snp2, data=snp_dom, family="binomial")
tidy(glm_dom)

term       estimate                    std.error          statistic            p.value
1     (Intercept)    -0.04175179       0.08714035       -0.4791327        0.6318442
2       snp1             -0.30452442       0.25780135       -1.1812367        0.2375087
3       snp2             0.12776467        0.13811577        0.9250549        0.3549374
4   snp1:snp2       0.55498377        0.43843040        1.2658423        0.2055695
```

```
# Recessive model
# Coding (AA=0, Aa=0, aa=1)
 snp_rec <- (snp == 2)*1
snp_rec <- data.frame(snp_rec)

table(snp_rec$snp1,snp_rec$snp2)

      0     1
 0  927  54
 1    1    0

glm_rec <- glm(pheno ~  snp1*snp2, data=snp_rec, family="binomial")
tidy(glm_rec)
```

	term	estimate	std.error	statistic	p.value
1	(Intercept)	-0.006472515	0.06568896	-0.09853276	0.9215093
2	snp1	12.572535519	324.74370272	0.03871526	0.9691174
3	snp2	0.006472515	0.27998056	0.02311773	0.9815564

No interaction term is fitted.

#Co-dominant model

```
snp6 <- factor(snp$snp6)      # take snp6 in place of snp1
table(snp6)
snp2 <- factor(snp$snp2)
table(snp2)
table(snp6,snp2)

glm_co <- glm(pheno ~  snp6*snp2, family="binomial")
tidy(glm_co)
summary(glm_co)
```

	term	estimate	std.error	statistic	p.value
1	(Intercept)	0.16034265	0.2837516	0.56508110	0.5720186
2	snp61	-0.26413944	0.3088198	-0.85531888	0.3923746
3	snp62	-0.26875195	0.3082172	-0.87195636	0.3832322
4	snp21	0.14981227	0.4879365	0.30703231	0.7588188
5	snp22	-0.16034265	1.0394782	-0.15425302	0.8774102
6	snp61:snp21	0.03154276	0.5329082	0.05918985	0.9528009
7	snp62:snp21	0.09674737	0.5250364	0.18426793	0.8538033
8	snp61:snp22	-0.17769321	1.1304302	-0.15719078	0.8750945
9	snp62:snp22	0.55643401	1.1139233	0.49952633	0.6174086

Likelihood ratio test can be used to test the significance of all 4 interaction terms.

Fit a model without the interaction term

```
glm_main <- glm(pheno ~  snp6 + snp2, family="binomial")
```

Fit a model with interaction terms

```
glm_co <- glm(pheno ~  snp6*snp2, family="binomial")
```

Test

```
anova(glm_main, glm_co, test= "Chisq")
```

Analysis of Deviance Table
Model 1: pheno ~ snp6 + snp2
Model 2: pheno ~ snp6 * snp2

	Resid.	Df Resid.	Dev Df	Deviance	Pr(>Chi)
1	976	1356.4			
2	972	1354.8	4	1.5823	0.812

In this data there is no evidence in support of the effect of interactions.

Adjusting for covariates

```
glm_co <- glm(pheno ~  snp6*snp2 + population, family="binomial")
tidy(glm_co)
```

	term	estimate	std.error	statistic	p.value
1	(Intercept)	0.20734475	0.2848411	0.7279314	0.4666556
2	snp61	-0.16234391	0.3120051	-0.5203246	0.6028374
3	snp62	-0.13083625	0.3135911	-0.4172193	0.6765180
4	snp21	0.32320573	0.4941460	0.6540693	0.5130671
5	snp22	0.12645665	1.0458202	0.1209162	0.9037574
6	populationJPT+CHB	-0.33380141	0.1324656	-2.5199095	0.0117385*
7	snp61:snp21	-0.07517806	0.5360393	-0.1402473	0.8884646
8	snp62:snp21	-0.09948315	0.5322356	-0.1869156	0.8517268
9	snp61:snp22	-0.35385130	1.1331641	-0.3122684	0.7548366
10	snp62:snp22	0.21760515	1.1225014	0.1938574	0.8462876

* Significant at 0.05 level

RNA-seq data analysis using GLM

- This is an example of count data analysis using a Poisson regression
- Data obtained from the paper Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays (Bottomly et al., 2011) . It is a comparative RNA-seq analysis of different mouse strains.
- Using RNA-Seq, an average of 22 million short sequencing reads were generated per sample for 21 samples (10 B6 and 11 D2),
- These reads were aligned to the mouse reference genome, allowing 16,183 Ensembl genes to be queried in striatum for both strains.
- A subset of the data is used here

```
rnaseq <- read.table("epistasis_rnaseq_example.txt")
```

Poisson regression to determine the effect of SNP-SNP interaction on RNA-seq after adjusting for the difference in strain.

```
glm_pois = glm(ENSMUSG00000000001 ~ SNP2*SNP5 + strain, data=rnaseq,
family="poisson")
tidy(glm_pois)
```

	term	estimate	std.error	statistic	p.value
1	(Intercept)	4.7125607	0.07812460	60.321087	0.000000e+00*
2	SNP2	0.9350636	0.05079155	18.409824	1.095777e-75*
3	SNP5	0.7147750	0.04015972	17.798307	7.283733e-71*
4	strainDBA/2J	0.1588196	0.02015197	7.881093	3.245300e-15*
5	SNP2:SNP5	-0.4269630	0.02837443	-15.047456	3.587680e-51**

*significant at 0.05 without correction for multiple testing

*Bonferroni's correction: number of pair-wise SNPs tests for the example data (number of snps=10 and considering only one gene ENSMUSG00000000001) 10*9/2=45

Adjusted level of significance= 0.05/45 = 0.0011

Screening epistasis

- So far we have considered how to test for interaction between two SNPs
- We deal with 500000 to 1 million SNPs across the genome
- One way to search for interactions is to perform an exhaustive search, considering all pairwise combinations
- Computationally possible, but time-consuming and dramatically increases multiple testing burden
- We may need to use filtering approach where only consider a subset of loci chosen based on biological or statistical considerations

Class Exercises

1. Using the GE and SNP data from the R-package MatrixEQTL, determine the interaction effect of Snp_04 and Snp_05 on Gene_09. Check the assumptions related to your model.

2. Use the snp data matrix created (see slide 28) from the SNP_data
    a. assess the interaction effect between snp9 (rs17159711) and snp12 (rs2275677) on the disease given by the variable "pheno" based on the co-dominant genetic model.

    b. repeat the analysis in (a) by adjusting for the covariate "population"

Mapping SNPs to Genes

Mapping SNPs to the corresponding genes allows to have a better understanding and interpretation of the SNPs and their interactions.

Use UCSC Genome browser: Open link
https://genome.ucsc.edu

Go to Tools → Variant Annotation Integrator
Mapping the selected SNPs using UCSC genome browser:
rs4880781 −→ ZMYND11 gene
rs10903439 −→ ADARB2 gene