IBI5086
Introdução a Métodos Estatísticos para a Bioinformática

Profa. Júlia Maria Pavan Soler

pavan@ime.usp.br

IME/USP – 2º Semestre/2020

Inferência sobre Efeitos Genéticos

Já vimos: $Y = f(X) + e$

✓ Ajuste de Modelos para "Comparação de 2 ou mais Populações" ⇒ **Variável resposta (Y) quantitativa** satisfazendo a premissas clássicas (normalidade, independência das observações, homocedasticidade)

Aplicação: Delineamentos Fatoriais (DCA ou DABC) para inferências sobre "Efeitos Genéticos"

- Ajuste de Modelos para "Comparação de 2 ou mais Populações" ⇒ **Variável resposta (Y) qualitativa (especificamente, binária)**

⇒ Estudos Observacionais: Transversal, Prospectivo, Caso-Controle
- Estudos de Associação em Tabelas de Contingência: Testes Qui-Quadrado
- Modelos de Regressão Logística

Variáveis Primárias do Estudo

⇒ **Como definir (selecionar) a variável primária (Y) do estudo (*outcome*)?**

Em geral, essa variável caracteriza uma doença ou fenômeno sob estudo:
Estudo de doenças cardiovasculares: pressão arterial, IMC, peso, circunferência abdominal, glicose, colesterol, etc.
Caracterização biométrica: altura, peso, comprimentos ósseos, perímetro cefálico
Caracterização molecular: concentração de enzimas, expressão de transcritos, abundância de peptídeo (ou proteína), etc.
Melhoramento genético de plantas e animais: produção de grãos, tamanho de bulbilhos de alho, número de grãos na espiga, nota para o sabor do fruto, peso da cria, produção leiteira, etc.

Variáveis Quantitativas (discretas ou contínuas)

Mas, as Variáveis Y podem ser Qualitativas (nominal ou ordinal).

Exemplos?

Tabelas de Contingência

Considere a seguinte distribuição de pacientes de acordo com uma Doença e um Fator de Risco:

Condição	**Fator de Risco**			
Doença	**R1**	**R2**	**R3**	**Total**
D	n10	n11	n12	**n1.**
ND	n20	n21	n22	**n2.**
Total	**n.0**	**n.1**	**n.2**	**n..**

Estes dados podem ter sido gerados dos estudos Caso 1, 2 ou 3. Compare os três tipos de estudos.

Caso 1: Dos prontuários de um Centro de Saúde, ***n..*** foram amostrados e, então, avaliada a condição de uma Doença (presente ou não) e de um Fator de Risco (R1, R2 e R3).

Caso 2: De um Centro de Saúde, pacientes, livres de uma doença, e classificados de acordo com um fator de Risco foram amostrados **(*n.0, n.1 e n.2*)** e acompanhados durante um período de tempo para, então, ser avaliado o desenvolvimento ou não da Doença.

Caso 3: Dentre os pacientes de um Centro de Saúde, com e sem uma Doença, amostras foram extraídas **(*n1. e n2.*)** e então, em cada grupo, a condição de um Fator de Risco foi avaliada.

Estudos de Associação – Testes Qui Quadrado

Fator de Risco Genético para Doenças

Caso 1: Estudo Transversal (n.. fixado)

	Marcador			
Condição	**aa**	**Aa**	**AA**	**Total**
D	n10	n11	n12	n1.
ND	n20	n21	n22	n2.
Total	n.0	n.1	n.2	**n..**

Teste de Associação entre as variáveis Doença e Fator de Risco (Marcador)

Caso 2: Estudo Prospectivo (n.0, n.1, n.2 fix.)

	Marcador			
Condição	**aa**	**Aa**	**AA**	**Total**
D	n10	n11	n12	n1.
ND	n20	n21	n22	n2.
Total	**n.0**	**n.1**	**n.2**	n..

Teste de Homogeneidade entre as Classes Genotípicas de acordo com a probabilidade ou não da Doença

Caso 3. Estudo Retrospectivo (Caso-Controle) (n1. e n2. fixados)

	Marcador			
Condição	**aa**	**Aa**	**AA**	**Total**
D	n10	n11	n12	**n1.**
ND	n20	n21	n22	**n2.**
Total	n.0	n.1	n.2	n..

⇒ Teste de Homogeneidade entre os grupos Doente e Não Doente relativamente às probabilidades de ocorrência do Fator de Risco (Genótipo do marcador)

Testes Qui-Quadrado podem ser usados nestes casos e têm a mesma expressão analítica (independem do plano amostral)

Estudos de Associação – Testes Qui Quadrado

Tabela de contingência com a distribuição de pacientes de acordo com a condição da Doença (D e ND) e do Genótipo do Marcador (0, 1 e 2 alelos A)

	Genótipo			
	aa	**Aa**	**AA**	**Total**
Caso	89	222	189	500
Controle	154	206	150	572

```
Pearson's Chi-squared test
X-squared = 22.375, df = 2, p-value = 0.00001385
Conclusão: Há evidência de associação significante entre doença e
Marcador. A distribuição dos genótipos não é homogênea no grupo
Doente e Não Doente
```

H_0: Não existe associação (equivalentemente: grupos homogêneos)
H_1: Há associação (equivalentemente: grupos não são homogêneos)

$$\chi^2_o = \sum_{ij} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

r: número de linhas; c: número de colunas

$$E_{ij} = \frac{n_{i.} \times n_{.j}}{n_{..}}$$ Frequência esperada

O_{ij}: Frequência observada da casela *ij*

Estudos Observacionais e Estatísticas

Caso 1: Estudo Transversal (n.. fixado)

	Marcador			
Condição	**aa**	**Aa**	**AA**	**Total**
D	n10	n11	n12	n1.
ND	n20	n21	n22	n2.
Total	n.0	n.1	n.2	**n..**

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n_{..}}; \sum_{ij} \pi_{ij} = 1;$$

π_{ij} : prevalência da categoria ij

Caso 2: Estudo Prospectivo (n.0, n.1, n.2 fix.)

	Marcador			
Condição	**aa**	**Aa**	**AA**	**Total**
D	n10	n11	n12	n1.
ND	n20	n21	n22	n2.
Total	**n.0**	**n.1**	**n.2**	n..

$$\hat{\pi}_{1j} = \frac{n_{1j}}{n_{.j}}; \sum_{i} \pi_{ij} = 1; RR_{jj'} = \frac{\pi_{1j}}{\pi_{1j'}}$$: risco relativo para j e j′

π_{1j} : incidência de D na categoria j

Caso 3. Estudo Retrospectivo (Caso-Controle) (n1. e n2. fixados)

	Marcador			
Condição	**aa**	**Aa**	**AA**	**Total**
D	n10	n11	n12	**n1.**
ND	n20	n21	n22	**n2.**
Total	n.0	n.1	n.2	n..

$\Rightarrow$ Estudos Caso Controle são comuns em GWAS

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n_{i.}}; \sum_{j} \pi_{ij} = 1;$$

π_{1j} : prevalência do genótipo j no grupo doente

$$\mathrm{OR}_{jj'} = \frac{\pi_{1j} / \pi_{2j}}{\pi_{1j'} / \pi_{2j'}} = \frac{\pi_{1j}\ \pi_{2j'}}{\pi_{2j}\ \pi_{1j'}} \overset{\mathrm{P(D)}\to 0}{\to} RR_{jj'}$$

OR: odds ratio (razão de chances). É uma medida de associação (ou homogeneidade) válida, independentemente do estudo.

Efeitos Genéticos - Estudos Caso-Controle

Análise de Associação: Dados de Genótipos

Como o Marcador infuencia na doença?

	Genótipo			
	aa	**Aa**	**AA**	**Total**
Caso	89	222	189	500
Controle	154	206	150	510

$$OR_{Aa} = \frac{222/206}{89/154} = 1,865$$

Chance da Doença para indivíduos que carregam o genótipo **Aa** é 1,87 a chance para **aa** (referência é o genótipo aa)

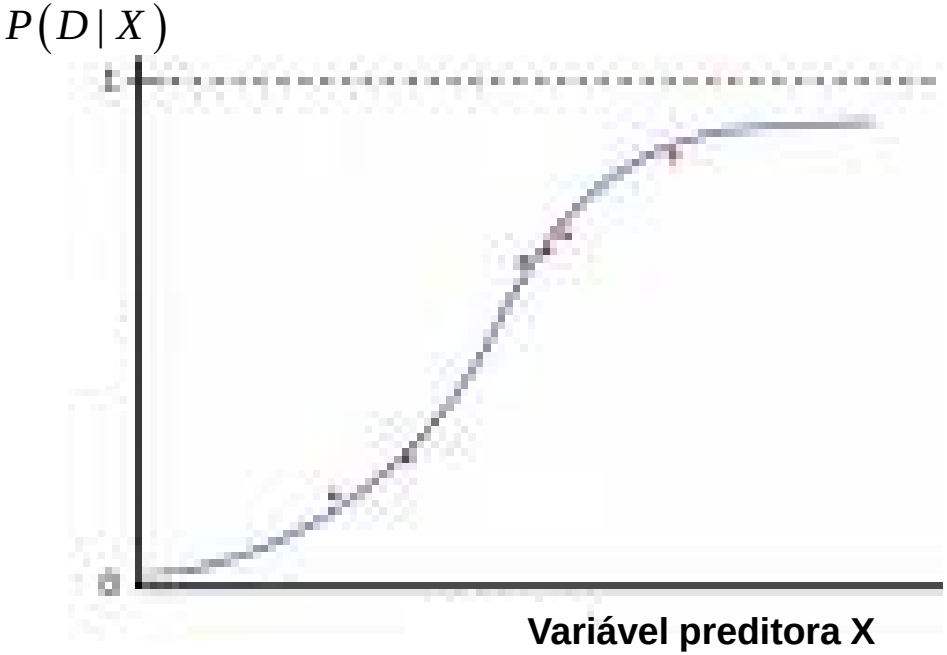
$$OR_{AA} = \frac{189/150}{89/154} = 2,180$$

Chance da Doença para indivíduos que carregam o genótipo **AA** é 2,18 a chance para **aa**

Hipótese H0: $\notin$ associação, não há efeito de marcador na Doença

$$H_0 : \theta_{AA} = \theta_{Aa} = 1$$

Análise de Associação via Modelos de Regressão Logística



$$P(D \mid X) = \frac{1}{1+e^{-(\alpha+\beta X)}}$$

$$\log ito = \ln \frac{P(D \mid X)}{1-P(D \mid X)} = \mu + \beta X$$

Modelo de regressão logística (log-linear)

$$\ln OR_{jj'} = \ln \frac{P(D \mid X_j)/\left(1-P(D \mid X_j)\right)}{P(D \mid X_{j'})/\left(1-P(D \mid X_{j'})\right)} = \log ito_j - \log ito_{j'}$$

$$\ln OR_{jj'} = \beta\left(X_j - X_{j'}\right) \quad \Rightarrow \quad OR = e^{\beta\left(X_j - X_{j'}\right)}$$

Estudos Caso-Controle – Regressão Logística

Distribuição Genotípica nos grupos Caso e Controle

	Genótipo			
	aa	**Aa**	**AA**	**Total**
Caso	89	222	189	500
Controle	154	206	150	510

Regressão Logística: Efeito do Marcador com 2 g.l.

$$Y_{ij} \mid X_{ij} \sim Bernoulli\left(1;\pi_j\right);\ j = D$$

Logito

$$\pi_j = \frac{1}{e^{\mu+\beta_1 X_{1j}+\beta_2 X_{2j}}} \Rightarrow \log\frac{\pi_j}{1-\pi_j} = \mu + \beta_1 X_{1j} + \beta_2 X_{2j}$$

Efeito de AA

Efeito de Aa

Logito (estimativas)		μ	β_1	β_2
Logito $_{aa}$	*log (89/154)*	1	0	0
Logito $_{Aa}$	*log (222/206)*	1	1	0
Logito $_{AA}$	*log (189/150)*	1	0	1

$$\log ito_{aa} = \mu$$

$$\log ito_{Aa} = \mu + \beta_1$$

$$\log ito_{AA} = \mu + \beta_2$$

$$\log OR_{Aa} = \log ito_{Aa} - \log ito_{aa} = \beta_1$$

$$\log OR_{AA} = \log ito_{AA} - \log ito_{aa} = \beta_2$$

$$H_0 : OR_{AA} = OR_{Aa} = 1$$

$$\Downarrow$$

$$H_0 : \beta_1 = \beta_2 = 0$$

Estudos Caso-Controle – Regressão Logística

Distribuição Genotípica nos grupos Caso e Controle

	Genótipo			
	AA	**Aa**	**aa**	**Total**
Caso	89	369	342	800
Controle	56	250	266	572
	145	619	608	

Ajuste do modelo de regressão logística com efeito do Marcador com 2 graus de liberdade.

Coefficientes:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.5483	0.1332	4.118	0.00003822	***
as.factor(x1)[1]	-0.6231	0.1646	-3.786	0.000153	***
as.factor(x1)[2]	-0.7794	0.1723	-4.524	0.00000608	***

$$\hat{\mu} = 0.5481$$

$$\hat{\beta}_1 = -0.6231 \quad \Rightarrow \quad OR_{Aa} = e^{\hat{\beta}_1(X_{0i} - X_{1i})} = e^{\hat{\beta}_1(0-1)} = 1{,}865$$

$$\hat{\beta}_2 = -0.7794 \quad \Rightarrow \quad OR_{AA} = e^{\hat{\beta}_2(X_{0i} - X_{2i})} = e^{\hat{\beta}_2(0-1)} = 2.180$$

Estudos Caso-Controle – Regressão Logística

Distribuição Genotípica nos grupos Caso e Controle

	Genótipo			
	aa	**Aa**	**AA**	**Total**
Caso	89	222	189	500
Controle	154	206	150	510

Regressão Logística: Efeito Linear do Marcador

$$Y_{ij} \mid X_{ij} \sim Bernoulli\left(1;\pi_j\right); j = D$$

$$\pi_j = \frac{1}{e^{\mu+\beta X_j}} \Rightarrow \log\frac{\pi_j}{1-\pi_j} = \mu + \beta\ X_j$$

Logito (estimativas)		μ	β
Logito $_{aa}$	*log (89/154)*	1	0
Logito $_{Aa}$	*log (222/206)*	1	1
Logito $_{AA}$	*log (189/150)*	1	2

Efeito linear de G

$$\log ito_{aa} = \mu$$
$$\log ito_{Aa} = \mu + \beta$$
$$\log ito_{AA} = \mu + 2\beta$$

$$\log OR_{Aa} = \log ito_{Aa} - \log ito_{aa} = \beta$$
$$\log OR_{AA} = \log ito_{AA} - \log ito_{aa} = 2\beta$$

$$H_0 : OR_{AA} = OR_{Aa} = 1$$
$$\Downarrow$$
$$H_0 : \beta = 0$$

Efeitos Genéticos - Estudos Caso-Controle

Distribuição Genotípica nos grupos Caso e Controle

	Genótipo			
	aa	**Aa**	**AA**	**Total**
Caso	89	222	189	500
Controle	154	206	150	510

Modelo de Regressão Logística Reduzido (somente com efeito linear do marcador)

```
Coefficientes:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.42644    0.11303   3.773 0.000162 ***
x1          -0.37073    0.08503  -4.360 0.000013 ***
```

$$\hat{\mu} = 0.42644$$

$$\hat{\beta} = -0.37073 \quad \Rightarrow \quad OR_{Aa} = e^{-\hat{\beta}} = 1.8647$$

$$\Rightarrow \quad OR_{AA} = e^{-2\hat{\beta}} = 2.897497$$

Estudos Caso-Controle – Regressão Logística

Distribuição Genotípica nos grupos Caso e Controle

	Genótipo			
	aa	**Aa**	**AA**	**Total**
Caso	89	222	189	500
Controle	154	206	150	510

Regressão Logística: Marcador com Efeito aditivo e de dominância (2 g.l.)

Efeito aditivo

Efeito de dominância

Logito (estimativas)		μ	β_1	β_2
Logito $_{aa}$	*log (89/154)*	1	-1	0
Logito $_{Aa}$	*log (222/206)*	1	0	1
Logito $_{AA}$	*log (189/150)*	1	1	0

$$\log ito_{aa} = \mu - \beta_1$$

$$\log ito_{Aa} = \mu + \beta_2$$

$$\log ito_{AA} = \mu + \beta_1$$

$$\mu = \left(\log ito_{AA} + \log ito_{aa}\right)/2$$

$$\left(\log ito_{AA} - \log ito_{aa}\right)/2 = \beta_1$$

$$\log ito_{Aa} - \left(\log ito_{AA} + \log ito_{aa}\right)/2 = \beta_2$$

$$H_0 : \beta_1 = \beta_2 = 0$$

Estudos Caso-Controle – Regressão Logística

Distribuição Genotípica nos grupos Caso e Controle

	Genótipo			
	aa	**Aa**	**AA**	**Total**
Caso	89	222	189	500
Controle	154	206	150	510

Regressão Logística: Marcador com Efeito aditivo e de dominância (2 g.l.)

```
Coefficientes:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.15860    0.08615   1.841   0.0656 .
x1a         -0.38971    0.08615  -4.524 6.08e-06 ***
x1d         -0.23340    0.12954  -1.802   0.0716 .
```

Modelo Reduzido: Somente com efeito aditivo

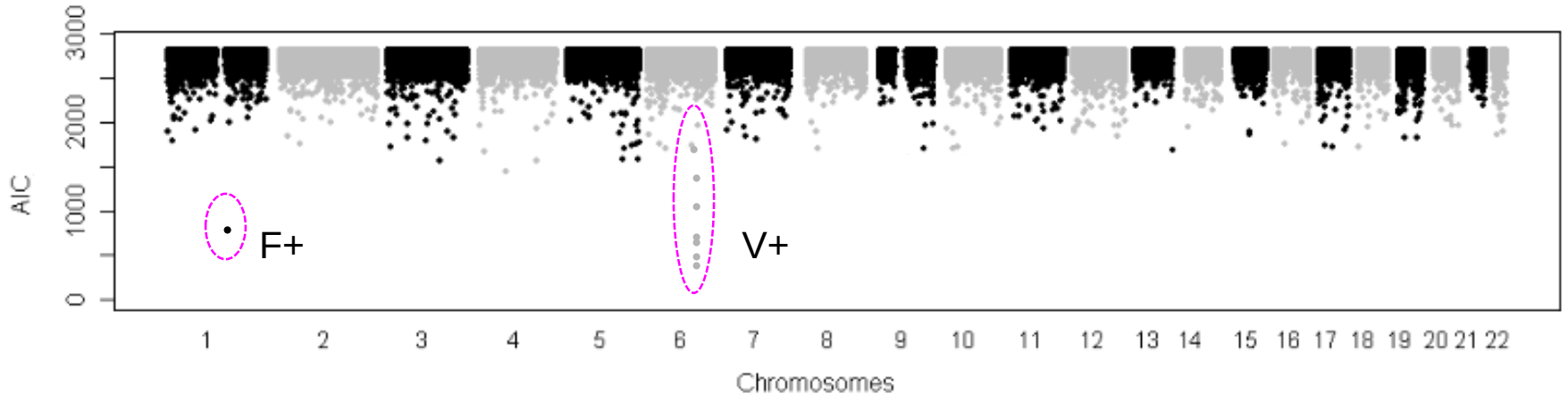
```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.05571    0.06410   0.869    0.385
x1a         -0.37073    0.08503  -4.360  1.3e-05 ***
```

Estudos Caso-Controle - GAW16

Ajuste de Modelos Logísticos para cada SNP : em geral, considerando somente o efeito linear (com 1 g.l.)

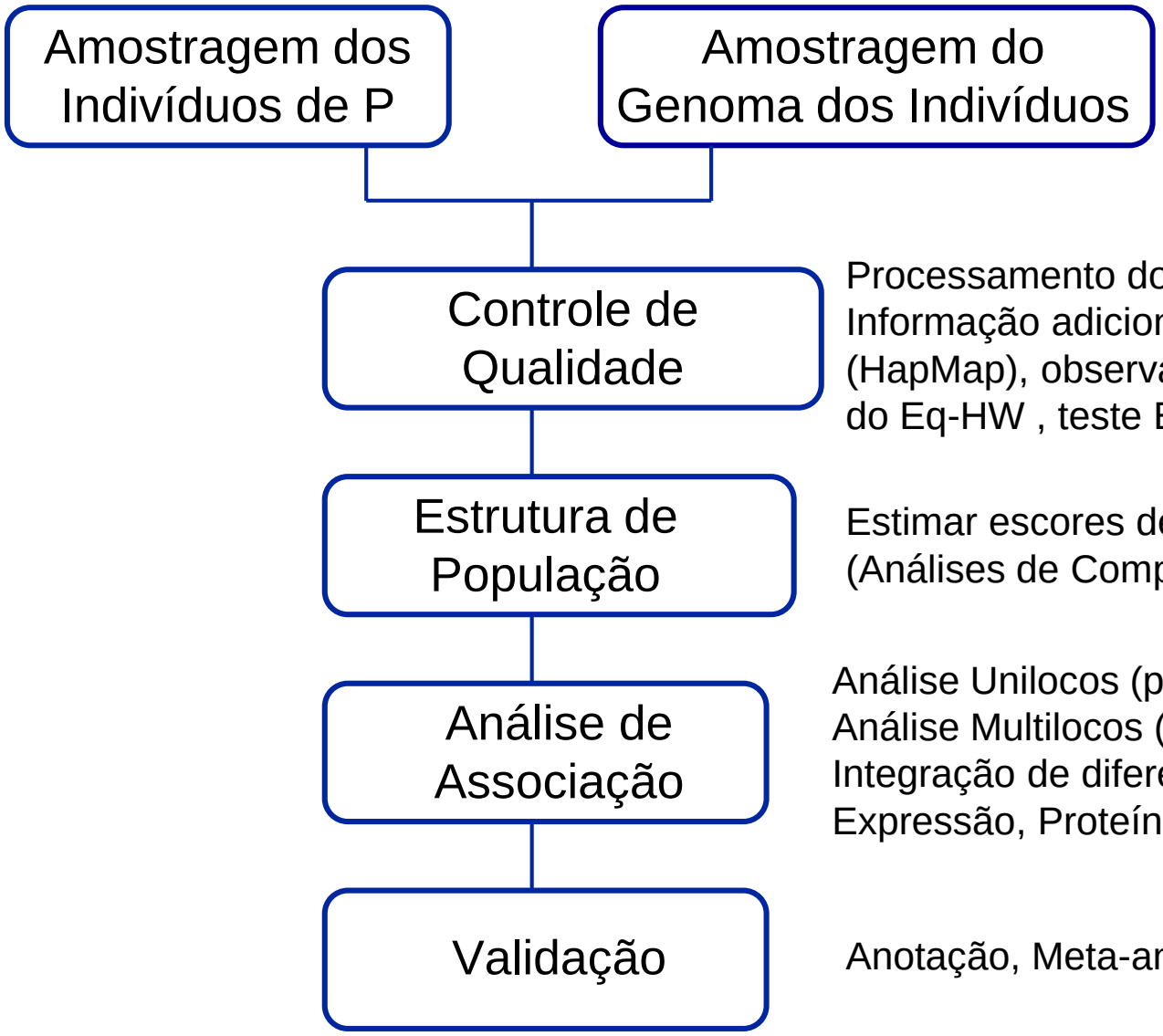
$$\text{logit}[P(Y=1|X_G)] = \beta_0 + \beta_G X_G \qquad X_m = \begin{cases} 0 & se\ aa \\ 1 & se\ Aa \\ 2 & se\ AA \end{cases}$$

m= 1,...,501.464



- SNPs, individualmente, têm pequeno efeito (poucos sinais significantes)
- Correção para Múltiplos testes (Bonferroni, FDR, adotar: $\alpha=10^{-4}$, $\alpha=10^{-8}$)
- Problemas na replicação e validação dos resultados

Estudos de Associação Genéticos - GWAS



Estudos de Associação - Confundimento devido a Populações Estratificadas

Alelo Marcador	A	a	**Total**
Caso	60	40	**100**
Controle	40	60	**100**

"Paradoxo de Simpson"
Falsos Positivos!
Como Combinar Tabelas?

*Razão de chances = (60*60)/(40*40) = 2.25*

Grupo Étnico I

Alelo Marcador	A	a	**Total**
Caso	30	30	**60**
Controle	30	30	**60**

Razão de chances = 1.00

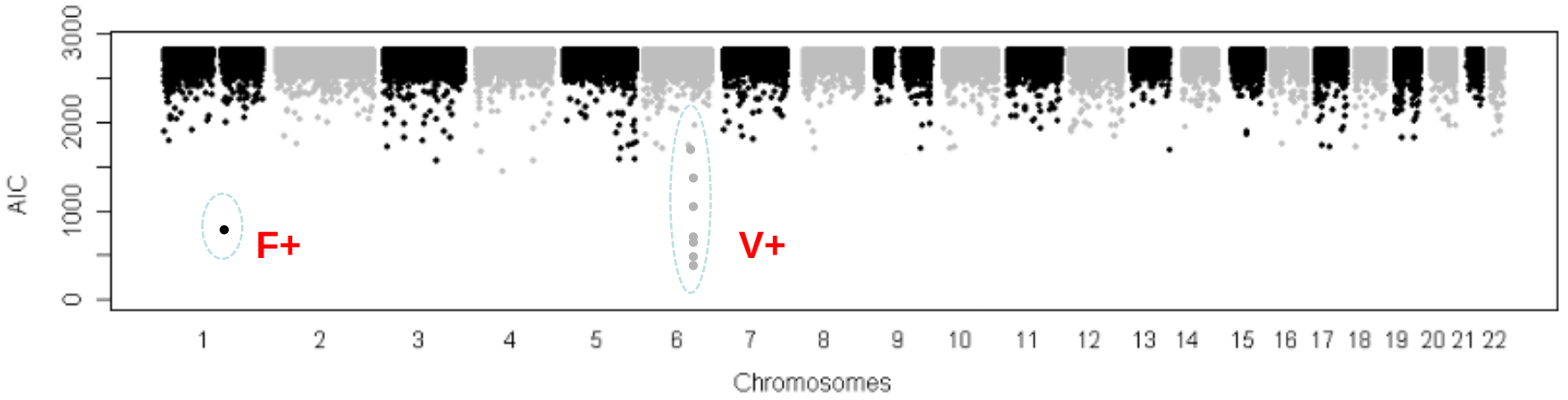
Grupo Étnico II

Alelo Marcador	A	a	**Total**
Caso	30	10	**40**
Controle	10	30	**40**

Razão de chances = 9.00

GWAS – Ajuste para Estrutura de Populações

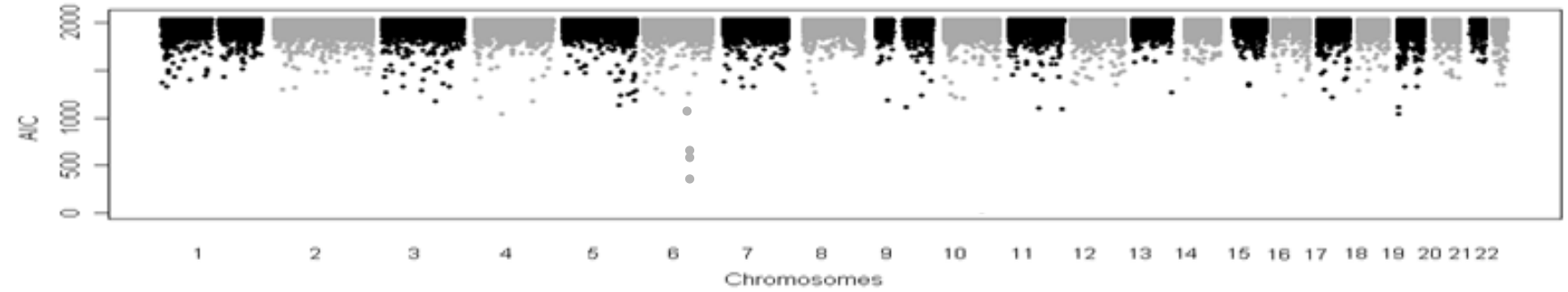
Sem correção para estrutura de população: $\text{logit}[P(Y=1|X_G)] = \beta_0 + \beta_G X_G$



Covariáveis caracterizando miscigenação

Com correção para estrutura de população:

$$\log it\left[P(Y=1|X)\right] = \beta_0 + \beta_1 X_{anc1} + \beta_2 X_{anc2} + \beta_G X_G$$



Modelos mais Gerais

Aplicação: Modelos Multilocos (com efeito de mais de um Marcador, além de covariáveis) para Mapeamento de Genes Associados com Doenças

- Modelo Uni-loco: Efeito Linear do marcador (ajustado ou não por covariáveis)

$$\text{logit}[\text{P(Y=1|X)}] = \beta_0 + \beta_{aG} X_{aG}; \quad G = 1,...,10^6$$

- Modelo Multilocos Aditivo

$$\text{logit}[\text{P(Y=1|X}_1,...,\text{X}_\text{M})] = \beta_0 + \sum_{G=1}^{M} \beta_G X_G$$

- Modelo Multi-locos de Interação (bilocos)

O que é o efeito de interação?

$$\text{logit}[\text{P(Y=1|X}_1,...,\text{X}_\text{M})] = \beta_0 + \sum_{G=1}^{M} \beta_G X_G + \sum_{G=1}^{M} \sum_{l=G+1}^{M} \beta_{Gl} \left(X_G * X_l \right)$$

GWAS – Modelos Multilocos

Análise Bilocos – Efeitos Principais: SNP1 e SNP2

Genótipo Gene1	0 (aa)	1 (Aa)	2 (AA)
Y=0 (Controle)	451	438	111
Y=1 (Caso)	459	436	105

Genótipo Gene2	0 (aa)	1 (Aa)	2 (AA)
Y=0 (Controle)	263	494	243
Y=1 (Caso)	280	487	233

	Chance Observada			Chance Relativa Observada: OR	
Genótipo	0	1	2	1/0	2/0
Gene1	1,0177	0,9954	0,9459	0,9781	0,9294
Gene2	1,0646	0,9858	0,9588	0,9259	0,9006

GWAS – Modelos Multilocos

Análises Bilocos – Efeito Conjunto: SNP1 e SNP2

Gene2	0			1			2		
Gene1	0	1	2	0	1	2	0	1	2
Y=0	106	146	11	247	186	61	98	106	39
Y=1	137	97	46	205	244	38	117	95	21

Gene2	0			1			2		
Gene1	0	1	2	0	1	2	0	1	2
Chance da doença	1,2924	0,6643	4,1818	0,8299	1,31182	0,6229	1,1938	0,8962	0,5384
Chance Relativa		0,5140	3,2356		1,5883	0,7505		0,7507	0,4509
		OR_{Aa_bb}	OR_{AA_bb}		OR_{Aa_Bb}	OR_{AA_Bb}		OR_{Aa_BB}	OR_{AA_BB}

$$\frac{OR_{Aa\_Bb}}{OR_{Aa\_bb}} = \frac{1.5883}{0.5140} = 3.09 \qquad \frac{OR_{Aa\_BB}}{OR_{Aa\_bb}} = \frac{0.7507}{0.5140} = 1.4605$$

$$\frac{OR_{AA\_Bb}}{OR_{AA\_bb}} = \frac{0.7505}{3.2356} = 0.2319 \qquad \frac{OR_{AA\_Bb}}{OR_{AA\_bb}} = \frac{0.4509}{3.2356} = 0.1394$$

GWAS – Ajuste de Modelos Unilocos

Para SNP1:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.01758	0.06630	0.265	0.791
as.factor(x[, 1])1	-0.02216	0.09472	-0.234	**0.815**
as.factor(x[, 1])2	-0.07315	0.15142	-0.483	**0.629**

$$OR_{Aa} = e^{-0,02216} = 0.9781$$
$$OR_{AA} = e^{-0,07315} = 0.9294$$

```
> confint(fit1.l) #IC para os betas
                     2.5 %     97.5 %
(Intercept)        -0.1123876 0.1476030
as.factor(x[, 1])1 -0.2078642 0.1635140
as.factor(x[, 1])2 -0.3705009 0.2236509

#IC(OR) a 95% em cada genótipo, Aa e AA relativo a aa
> lior.Aa <- exp(conf1[2,1]); lsor.Aa <- exp(conf1[2,2])
> (0.8123174; 1.177642)

> lior.AA <- exp(conf1[3,1]); lsor.AA <- exp(conf1[3,2])
> (0.6903884; 1.250634)
```

Não há evidência amostral de efeito (aditivo) significante do SNP1!
Note que o IC(*OR*) a 95% inclui o valor 1!

GWAS – Ajuste de Modelos Unilocos

Para SNP2:

```
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.06264    0.08587   0.729    0.466
as.factor(x[, 2])1 -0.07691    0.10701  -0.719    0.472
as.factor(x[, 2])2 -0.10466    0.12562  -0.833    0.405
```

$$OR_{Bb} = e^{-0,07691} = 0.9259$$
$$OR_{BB} = e^{-0,10466} = 0.9006$$

```
> confint(fit2.rl) #IC para os betas
                         2.5 %     97.5 %
(Intercept)         -0.1056188 0.2311863
as.factor(x[, 2])1 -0.2868074 0.1327894
as.factor(x[, 2])2 -0.3510810 0.1415048

> #IC(OR) a 95% em cada genótipo, Aa e AA relativo a aa
> lior.Aa <- exp(conf2[2,1]); lsor.Aa <- exp(conf2[2,2])
> (0.7506563; 1.142009)
>
> lior.AA <- exp(conf2[3,1]); lsor.AA <- exp(conf2[3,2])
> (0.7039267; 1.152006)
```

Não há evidência amostral de efeito (aditivo) significante do SNP2!
Note que o IC(*OR*) a 95% inclui o valor 1!

GWAS – Ajuste de Modelos Bilocos

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.25654	0.12936	1.983	0.047344
as.factor(x[, 1])1	-0.66544	0.18410	-3.615	0.000301
as.factor(x[, 1])2	1.17420	0.35970	3.264	0.001097
as.factor(x[, 2])1	-0.44292	0.16019	-2.765	0.005692
as.factor(x[, 2])2	-0.07934	0.18837	-0.421	0.673637
as.factor(x[, 1])1:as.factor(x[, 2])1	1.12324	0.22868	4.912	9.02e-07
as.factor(x[, 1])2:as.factor(x[, 2])1	-1.46111	0.42546	-3.434	0.000594
as.factor(x[, 1])1:as.factor(x[, 2])2	0.37867	0.26945	1.405	0.159920
as.factor(x[, 1])2:as.factor(x[, 2])2	-1.97045	0.47052	-4.188	2.82e-05

Efeitos de interação

$$\frac{OR_{Aa\_Bb}}{OR_{Aa\_bb}} = e^{1.1232} = 3.0747$$

Chance da doença para indivíduos AaBb é 3,07 a chance para Aabb

$$\frac{OR_{AA\_Bb}}{OR_{AA\_bb}} = e^{-1.4611} = 0.2319$$

$$\frac{OR_{Aa\_BB}}{OR_{Aa\_bb}} = e^{0.3787} = 1.4604$$

Não significante

$$\frac{OR_{AA\_BB}}{OR_{AA\_bb}} = e^{-1.97045} = 0.1394$$

Chance da doença para indivíduos AABB é 13,94% da chance para AAbb

Note que, o efeito significante dos marcadores, somente foi evidenciado no modelo completo, no qual foi incluído o efeito de interação entre os locos!

GWAS – Ajuste de Modelos Bilocos

Parametrização alternativa

Efeitos Aditivos:

$$\log OR_{AA} = \log ito_{AA} - \log ito_{aa} = 2a_1 \Rightarrow a_1 = \left(\log ito_{AA} - \log ito_{aa}\right)/2 \Rightarrow OR_{AA} = e^{2a_1}$$

$$\log OR_{BB} = \log ito_{BB} - \log ito_{bb} = 2a_2 \Rightarrow a_2 = \left(\log ito_{BB} - \log ito_{bb}\right)/2 \Rightarrow OR_{BB} = e^{2a_2}$$

Efeito de Interação Aditivo*Aditivo:

$$\log ito_{aabb} = \mu - a_1 - a_2 + a_{12}$$
$$\log ito_{aaBB} = \mu - a_1 + a_2 - a_{12}$$
$$\log ito_{AAbb} = \mu + a_1 - a_2 - a_{12}$$
$$\log ito_{AABB} = \mu + a_1 + a_2 + a_{12}$$

Loco1	0			1			2		
Loco2	0	1	2	0	1	2	0	1	2
a_1	-1	-1	-1	0	0	0	1	1	1
a_2	-1	0	1	-1	0	1	-1	0	1
a_{12}	1	0	-1	0	0	0	-1	0	1

$$\log ito_{aabb} - \log ito_{aaBB} - \log ito_{AAbb} + \log ito_{AABB} = 4a_{12}$$
$$\Rightarrow (\log ito_{AABB} - \log ito_{AAbb}) - (\log ito_{aaBB} - \log ito_{aabb}) = 4a_{12}$$

$$4a_{12} = \log \frac{(OR_{BB})_{AA}}{(OR_{BB})_{aa}} \Rightarrow \frac{(OR_{BB})_{AA}}{(OR_{BB})_{aa}} = e^{4a_{12}}$$

GWAS – Ajuste de Modelos Bilocos

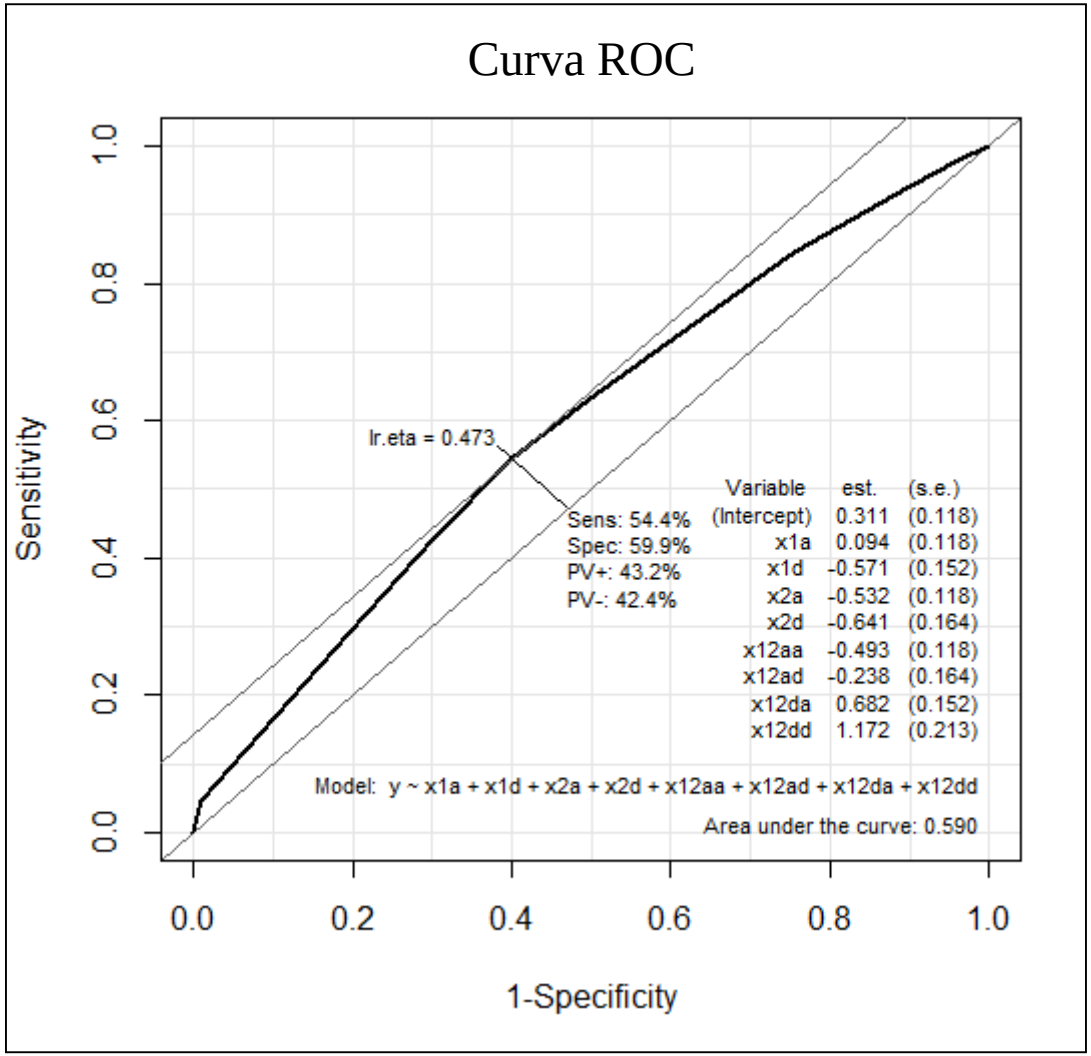
Parametrização alternativa:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	0.31136	0.11763	2.647	0.008122	**
x1a	0.09449	0.11763	0.803	0.421817	
x1d	-0.57059	0.15204	-3.753	0.000175	***
x2a	-0.53228	0.11763	-4.525	6.04e-06	***
x2d	-0.64120	0.16354	-3.921	8.83e-05	***
x12aa	-0.49261	0.11763	-4.188	2.82e-05	***
x12ad	-0.23794	0.16354	-1.455	0.145685	
x12da	0.68195	0.15204	4.485	7.28e-06	***
x12dd	1.17185	0.21331	5.494	3.94e-08	***

$$\frac{(OR_{AA})_{BB}}{(OR_{AA})_{bb}} = e^{4*(-0.49261)}$$

$$= e^{-1.9704} = 0.1394$$

GWAS – Ajuste de Modelos Bilocos



Curva ROC: Curva Característica de Operação

Sensibilidade = P(V+)
= P(Rej H0| H0 é Falso)
= 1 - β = Poder do teste

Especificidade = P(V-)
= P(Não Rej H0| H0 é Verdadeiro)
= 1 - α

α=P(erro tipo I)
= P(Rej H0| H0 é Verdadeiro)

β=P(erro tipo II)
= P(Não Rej H0| H0 é Falso)

Modelos de Regressão Logística
Predição da Doença

$$\hat{P}(D \mid X_i) = \frac{1}{1+e^{-(\mu+\hat{\beta}X_i)}} \geq 0.50 \Rightarrow$$ **Indivíduo predito como Doente**

Modelo para SNP1 (efeito linear)

	y	
test.pred1	0	1
0	549	541
1	451	459

%ClassifCorreta=50.4

Modelo para SNP2 (efeito linear)

	y	
test.pred2	0	1
0	737	720
1	263	280

%classifCorreta=50.9

Modelo para SNP1 e SNP2 com interação

	y	
test.pred12	0	1
0	599	456
1	401	544

%classifCorreta= 0.5715

GWAS – Modelos Multilocos

Problema de Seleção de Variáveis (em geral, sob n<<p)

n: tamanho amostral
p:número de var.

Como pesquisar o espaço genômico (de alta dimensão) para redução de dimensionalidade (subconjuntos de “genes” significantes)

- MDR: Multifactor Dimensionality Reduction (Moore, 2007)
- PLR: Penalized Logistic Regression (Park and Hastie, 2008)
- LASSO-based (Tibshirani, 1996; Valdar et al., 2012)
- SNP Harvester (Yang et al., 2009)
- Análise por haplótipos: construção de blocos de SNPs associados (em desequilíbrio de ligação)

GWAS – Regressão Logística Penalizada

```
> summary(fitip)   # plr(x = xau, y = y)

Coefficients:
          Estimate Std.Error z value Pr(>|z|)
Intercept  0.31136   0.11763   2.647    0.008
x1a        0.09448   0.11763   0.803    0.422
x1d       -0.57058   0.15204  -3.753    0.000
x2a       -0.53227   0.11763  -4.525    0.000
x2d       -0.64118   0.16354  -3.921    0.000
x12aa     -0.49261   0.11763  -4.188    0.000
x12ad     -0.23793   0.16354  -1.455    0.146
x12da      0.68194   0.15204   4.485    0.000
x12dd      1.17183   0.21330   5.494    0.000
```

Park and Hastie, 2008.
Pacote R: PLRModels (plr, step.plr)

Modelo de regressão logística com interações

```
Forward stepwise selection: Variables with nonzero coefficients
x2a
x1d:x2a
x1a:x2a

          Estimate Std.Error z value Pr(>|z|)
Intercept  0.00404   0.04506   0.090    0.928
x2a       -0.47874   0.10599  -4.517    0.000
x1d:x2a    0.63912   0.13851   4.614    0.000
x1a:x2a   -0.43256   0.10750  -4.024    0.000
```

Conjunto das variáveis que permaneceram no modelo.