Introduction to logistic regression
Outline

• General concepts
• Model formulation
• Parameter interpretation
• Estimation
• Model formulation in genetic studies
  - single genotype models
  - multilocus models
  - haplotype models
• A case study in R: CEPH data
General Concepts

• Regression is the study of dependence between a response variable (y), the dependent variable and one or several predictors (x), the independent variables.

• The response variable is binary

• It is a simple representation/modeling of the dependence between one or several variables.
The Logistic Model

- \( \text{Prob}(Y_i=1) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \)
  \( \eta_i = \sum_j x_{ij} b_j \) - Linear Predictor

- \( x_{ij} \) - Design Matrix (genotypes etc)
- \( b_j \) - Model Parameters (to be estimated)

- Model is investigated by
  - estimating the \( b_j \)'s by maximum likelihood
  - testing if the estimates are different from 0
The Logistic Function

\[ \text{Prob}(Y_i=1) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \]
Interpretation of the Parameters

- Logit \( \{Y=1|X\} = \logit(P) = \logit\left[\frac{P}{1-P}\right] = X\beta, \)

Measure of risk: \( \text{OR} = \exp(\beta) \)

- Increase \( X_j \) by \( d \) \( \rightarrow \) increase odds \( Y=1 \) by \( \exp(\beta_j d) \),
  increase log odds by \( \beta_j d \)

- If there is only one predictor \( X \) and that predictor is binary, the model can be written
  \[
  \begin{align*}
  \text{Logit \( \{Y=1|X=0\} = \beta_0 \)}
  \\
  \text{Logit \( \{Y=1|X=1\} = \beta_0 + \beta_1 \).}
  \end{align*}
  \]
Interpretation - cont'd

- One continuous predictor:
  \[ \text{Logit} \{Y=1|X\} = \beta_0 + \beta_1 X \]

- Two treatments (indicated by \( X_1 = 0 \) or 1) and one continuous covariable (\( X_2 \))
  \[ \text{Logit} \{Y=1|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \]
  
  Eg.,
  \[ \text{Logit} \{Y=1|X_1=0, X_2\} = \beta_0 + \beta_2 X_2 \]
  \[ \text{Logit} \{Y=1|X_1=1, X_2\} = \beta_0 + \beta_1 + \beta_2 X_2 \]
Estimation

- In ordinary least squares regression, main objective function is SSE
- If residuals are normally distributed, the resulting least squares estimates are optimal (consistency and lowest variances among unbiased estimates)
- Other fitting criteria such as minimizing sum of absolute errors are needed for non-normal residuals (or residuals not assumed to be symmetrically distributed)
- With binary Y a drastic change is needed
- Need a general way to write down a good fitting criterion for many different types of Y and for any distribution of Y | X
- Maximum likelihood (ML) is a general solution
Maximum likelihood estimation

• Example: 1-sample binomial problem
• Single unknown \( P = \) probability of an event in a population unknown parameter, the probability of an event in a population.
• Occurrence of the event signaled by \( Y = 1 \), non-occurrence by \( Y = 0 \), for an individual with \( \text{Prob} \{ Y = 1 \} = P \)
• Draw a random sample of size \( n = 3 \) from the population and observed the events \( Y = 1; 0; 1 \)
• Assuming individuals in the sample act completely independently, proba. of the 3 events is \( P^2(1-P) \); this joint probability is called the likelihood
• \( P \) is unknown but the ML estimate (MLE) can be computed by solving for \( P \) that makes the likelihood maximum
Maximum likelihood estimation

• In general if $Y$ is binary so that the sample is $Y_1, \ldots, Y_n$ and $s$ is $\Sigma Y_i$, the likelihood is:

$$L = \prod_i P^{Y_i}(1-P)^{1-Y_i} = Ps(1 - P)^{n-s}$$

• For numerical and statistical reasons we work with the log-likelihood function

$$\log L = s \log(P) + (n - s) \log(1 - P)$$

• In logistic regression we allow differences in subject characteristics through $Xs$
Logistic Regression in Genetics

- Applicable to Association Studies
- Data:
  - Binary outcomes (e.g., disease status)
  - Dependent on genotypes [+ sex, environment]
- Aim is to identify which factors influence the outcome
- Rigorous tests of statistical significance
- Flexible modelling language
- Generalisation of Chi-Squared Test
Example

BMC Cancer

Research article

**SNP-SNP interactions in breast cancer susceptibility**

Venüs Ümmiye Onay¹,³, Laurent Briollais¹,²,⁵, Julia A Knight¹,²,⁵, Ellen Shi⁴, Yuanyuan Wang¹,², Sean Wells¹,³, Hong Li¹,³, Isaac Rajendram¹,³, Irene L Andrilis¹,³,⁴,⁶,⁷ and Hilmi Ozcelik*¹,³,⁷

Table 4: Analysis of two-way SNP interaction effects on breast cancer.

<table>
<thead>
<tr>
<th>Two-way Interactions between polymorphisms</th>
<th>Crude P-values</th>
<th>Bootstrap frequency of stepwise variable selection</th>
<th>Bootstrap P-value</th>
<th>FDR adjusted P-value</th>
<th>FPRPI</th>
</tr>
</thead>
<tbody>
<tr>
<td>XPD-[Lys751Gln] and IL10-[G(-1082)A]</td>
<td>0.035</td>
<td>68%</td>
<td>0.001</td>
<td>0.007</td>
<td>0.092</td>
</tr>
<tr>
<td>COMT-[Met108/158Val] and CCND1-[Pro241Pro]</td>
<td>0.010</td>
<td>61%</td>
<td>0.002</td>
<td>0.014</td>
<td>0.169</td>
</tr>
<tr>
<td>GSTP1-[Ile105Val] and COMT-[Met108/158Val]</td>
<td>0.036</td>
<td>54%</td>
<td>0.001</td>
<td>0.007</td>
<td>0.093</td>
</tr>
<tr>
<td>CYP17-[C(518)T] and GADD45-[G(VS3+168)T]</td>
<td>0.024</td>
<td>53%</td>
<td>0.018</td>
<td>0.062</td>
<td>0.999</td>
</tr>
<tr>
<td>BARD1-[Pro24Ser] and ESR1-[Pro323Pro]</td>
<td>0.039</td>
<td>51%</td>
<td>ns¹</td>
<td>ns</td>
<td>-</td>
</tr>
<tr>
<td>TNFA-[G(-308)A] and p27-[Val109Gly]</td>
<td>0.016</td>
<td>49%</td>
<td>0.025</td>
<td>0.079</td>
<td>0.996</td>
</tr>
<tr>
<td>BARD1-[Pro24Ser] and p27-[Val109Gly]</td>
<td>0.021</td>
<td>44%</td>
<td>ns</td>
<td>ns</td>
<td>-</td>
</tr>
<tr>
<td>BARD1-[Pro24Ser] and XPD-[Lys751Gln]</td>
<td>0.024</td>
<td>36%</td>
<td>0.002</td>
<td>0.014</td>
<td>0.671</td>
</tr>
<tr>
<td>ESR1-[Ser105Ser] and ESR1-[Pro323Pro]</td>
<td>0.028</td>
<td>30%</td>
<td>0.097</td>
<td>ns</td>
<td>0.999</td>
</tr>
</tbody>
</table>
**Coding Unphased Genotypes**

- Several possibilities:
  - AA, AG, GG *original genotypes*
  - 12, 21, 22
  - 1, 2, 3
  - 0, 1, 2  *# of G alleles*

- **Missing Data**
  - NA default in R
Case study: genetic association

- CEPH pedigrees which consist of 12 multigenerational Caucasian families from Utah including 107 individuals. DNA for the CEPH family pedigree is available for genetic studies.
- The marker data consist of genotypes for 20 SNP markers, six of them are genotyped for all individuals, the remaining 14 are genotyped in only 50-54 individuals.
- The study looks for an association between these SNPs and a gene expression phenotype (mRNA) that we dichotomized here (>median vs. ≤ median)
- http://www.sph.umich.edu/csg/abecasis/Merlin/tour/assoc.html
Case study: R

R code:

```r
ceph.data<-read.table(paste(my.directory,"ceph_data.txt",sep=""),header=T,na.strings = "0/0")
attach(ceph.data)
#### Create binary trait (Case-Control status) ####
CC<-cut(qt, breaks=c(min(qt),median(qt),max(qt)))
levels(CC)<-c("control","case")
hist(qt)
abline(v=median(qt),col = "red", lty=3)
```

R output
Case study: genotype association

**R code:**
```
snp1.geno<-genotype(snp1)
tab.snp1<-table(CC, snp1)
print(tab.snp1)
print(summary(tab.snp1))
```

**R output**

<table>
<thead>
<tr>
<th>CC</th>
<th>snp1</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1/1</td>
<td>1/3</td>
</tr>
<tr>
<td>control</td>
<td>4</td>
<td>26</td>
</tr>
<tr>
<td>case</td>
<td>5</td>
<td>24</td>
</tr>
</tbody>
</table>

Number of cases in table: 106
Number of factors: 2
Test for independence of all factors:

\[
\text{OR }_{1/1 \text{ vs. } 3/3} = \frac{5 \times 23}{4 \times 24} = 1.20
\]

\[
\text{OR }_{1/3 \text{ vs. } 3/3} = \frac{24 \times 23}{26 \times 24} = 0.88
\]

Chi-squared approximation may be incorrect
Genotype association: logistic model

**R code:**

```r
glm.snp1<-glm(CC ~ snp1.geno, family=binomial, data=ceph.data)
summary(glm.snp1)
```

**R output**

Deviance Residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1.27352</td>
<td>-1.18256</td>
<td>-0.02969</td>
<td>1.15939</td>
<td>1.21159</td>
</tr>
</tbody>
</table>

Coefficients:

|             | Estimate | Std. Error | z value | Pr(>|z|) |
|-------------|----------|------------|---------|---------|
| (Intercept) | 0.2231   | 0.6708     | 0.333   | 0.739   |
| snp1.geno3/1 | -0.3032 | 0.7281     | -0.416  | 0.677   |
| snp1.geno3/3 | -0.1806 | 0.7315     | -0.247  | 0.805   |

(Dispersion parameter for binomial family taken to be 1)

Null deviance:   146.95 on 105 degrees of freedom
Residual deviance: 146.73 on 103 degrees of freedom
(1 observation deleted due to missingness)
AIC: 152.73

Problem with the coding!
Genotype association: logistic model

Test of association:

Null deviance: 146.95 on 105 degrees of freedom
Residual deviance: 146.73 on 103 degrees of freedom

Likelihood ratio test (LRT) =
-2*(Null deviance - Residual deviance) = 0.44

To compare with a $\chi^2$ with 2 d.f.
Case study: simple genotype association with logistic model

**R code:**

```r
snp1.geno<-allele.count(snp1.geno,summary(snp1.geno[CC="control"])[[1]][2])
snp1.geno<-factor(snp1.geno)
glm.snp1<-glm(CC ~ snp1.geno, family=binomial, data=ceph.data)
summary(glm.snp1)
```

**R output**

```
Deviance Residuals:
       Min        1Q    Median        3Q       Max
-1.27352  -1.18256  -0.02969   1.15939   1.21159

Coefficients:
            Estimate Std. Error   z value  Pr(>|z|)  OR  
(Intercept)  0.04256    0.29180   0.146    0.884  OR 1/1 vs. 3/3 =exp(0.18)=1.20
snp1.geno1  -0.12260    0.40654  -0.302    0.763  OR 1/3 vs. 3/3 =exp(-0.12)=0.88
snp1.geno2   0.18058    0.73153   0.247    0.805

(Dispersion parameter for binomial family taken to be 1)
```
## Types of genetic effect at a single locus

A = minor allele frequency

<table>
<thead>
<tr>
<th></th>
<th>AA</th>
<th>AG</th>
<th>GG</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Recessive</strong></td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Dominant</strong></td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Additive</strong></td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td><strong>Codominant</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>var1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>var2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Additive Genotype Model

- Code genotypes for SNP1 as
  - 3/3  \( x=0 \) → baseline
  - 3/1  \( x=1 \)
  - 1/1  \( x=2 \)

- Linear Predictor
  - \( \eta = b_0 + b_1 x \)

- \( P(Y=1|x) = \exp(b_0 + xb_1)/(1+\exp(b_0 + xb_1)) \)

- \( P_{3/3} = P(Y=1|x=0) = \exp(b_0)/(1+\exp(b_0)) \)
- \( P_{3/1} = P(Y=1|x=1) = \exp(b_0 + b_1)/(1+\exp(b_0 + b_1)) \)
- \( P_{1/1} = P(Y=1|x=2) = \exp(b_0 + 2b_1)/(1+\exp(b_0 + 2b_1)) \)
Case study: Additive Genotype Model

**R code:**

```r
snp1.geno<-genotype(snp1)
snp1.geno<-allele.count(snp1.geno,summary(snp1.geno[CC=="control"])[[1]][2])
glm.snp1<-glm(CC ~ snp1.geno, family=binomial, data=ceph.data)
summary(glm.snp1)
```

**R output**

<table>
<thead>
<tr>
<th>Deviance Residuals:</th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1.177</td>
<td>-1.177</td>
<td>0.000</td>
<td>1.177</td>
<td>1.177</td>
</tr>
</tbody>
</table>

| Coefficients:        | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------------|----------|------------|---------|----------|
| (Intercept)          | -6.924e-17 | 2.767e-01 | -2.50e-16 | 1        |
| snp1.geno            | 1.509e-17  | 3.072e-01 | 4.91e-17  | 1        |

(Dispersion parameter for binomial family taken to be 1)

Effect of SNP1 almost null!

Null deviance: 146.95 on 105 degrees of freedom  
Residual deviance: 146.95 on 104 degrees of freedom  
(1 observation deleted due to missingness)  
AIC: 150.95
Case study: Additive Genotype Model

Our model for SNP1

Hypothetical model with $b_0=-2$ and $b_1=0.2$
Dominant Genotype Model

- Code genotypes for SNP1 as
  - 3/3 \( x=0 \) baseline
  - 1/3 \( x=1 \)
  - 1/1 \( x=1 \)

- Linear Predictor
  - \( \eta = b_0 + xb_1 \)

- \( P(Y=1|x) = \frac{\exp(b_0 + xb_1)}{(1+\exp(b_0 + xb_1))} \)
- \( P_{3/3} = P(Y=0|x=0) = \frac{\exp(b_0)}{(1+\exp(b_0))} \)
- \( P_{1/3} = P_{1/1} = P(Y=0|x=1) = \frac{\exp(b_0 + b_1)}{(1+\exp(b_0 + b_1))} \)
Codominant Genotype Model

- Each genotype has a different probability
- Code genotypes as (for example)
  - 3/3 \( x_1=0, x_2=0 \) → baseline risk
  - 1/3 \( x_1=1, x_2=0 \)
  - 1/1 \( x_1=0, x_2=1 \)

- Linear Predictor
  - \( \eta = b_0 + b_1x_1 + b_2x_2 \) → two parameters

\[
P(Y=1|x) = \frac{\exp(b_0 + xb_1+yb_2)}{1+\exp(b_0 + xb_1+yb_2)}
\]
\[
P_{3/3} = P(Y=1|x_1=0,x_2=0) = \exp(b_0)/(1+\exp(b_0))
\]
\[
P_{1/3} = P(Y=1|x_1=1,x_2=0) = \exp(b_0 + b_1)/(1+\exp(b_0 + b_1))
\]
\[
P_{1/1} = P(Y=1|x_1=0,x_2=1) = \exp(b_0 + b_2)/(1+\exp(b_0 + b_2))
\]
## Models in R

<table>
<thead>
<tr>
<th>Genotype</th>
<th>AA</th>
<th>AG</th>
<th>GG</th>
<th>model</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recessive</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>$y \sim \text{dominant}(g)$</td>
<td>1</td>
</tr>
<tr>
<td>Dominant</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>$y \sim \text{recessive}(g)$</td>
<td>1</td>
</tr>
<tr>
<td>Additive</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>$y \sim \text{additive}(g)$</td>
<td>1</td>
</tr>
</tbody>
</table>

### R code:

```R
# Function to get codominant genotype
codominant <- function(x){factor(allele.count(x,summary(x[CC=='control'])[[1]][2]))}

# Function to get additive genotype
additive <- function(x){allele.count(x,summary(x[CC=='control'])[[1]][2])}

# Function to get dominant genotype
dominant <- function(x){carrier(x,summary(x[CC=='control'])[[1]][2])}

# Function to get recessive genotype
recessive <- function(x){homozygote(x,summary(x[CC=='control'])[[1]][2])}
```
Data Transformation

- g <- snp marker
- use these functions to treat a genotype vector in a certain way:
  - a <- additive(g)
  - r <- recessive(g)
  - d <- dominant(g)
  - c <- codominant(g)
Fitting the Model

R code:

- glm.snp1.codom<-glm(CC ~ snp1.geno.codom, family=binomial, data=ceph.data)
- glm.snp1.add<-glm(CC ~ snp1.geno.add, family=binomial, data=ceph.data)
- glm.snp1.dom<-glm(CC ~ snp1.geno.dom, family=binomial, data=ceph.data)
- glm.snp1.rec<-glm(CC ~ snp1.geno.rec, family=binomial, data=ceph.data)

- Equivalent models:
  - codominant = dominant + recessive
  - codominant = additive + recessive
  - codominant = additive + dominant
## Model Comparison

Akaike criteria (AIC): \(-2 \cdot \ln(L) + 2 \cdot \text{number of parameters}\)

<table>
<thead>
<tr>
<th>Model for SNP1</th>
<th>Df</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Codominant</td>
<td>2</td>
<td>152.73</td>
</tr>
<tr>
<td>Additive</td>
<td>1</td>
<td>150.95</td>
</tr>
<tr>
<td>Dominant</td>
<td>1</td>
<td>150.91</td>
</tr>
<tr>
<td>Recessive</td>
<td>1</td>
<td>150.83</td>
</tr>
</tbody>
</table>
Scanning all Markers

R code:

```r
for (i in 5:24){
  snp.geno<-codominant(genotype(ceph.data[,i]))
  model.null<-glm(CC ~ 1, family=binomial, data=ceph.data, na.action=na.omit, subset=!is.na(snp.geno))
  model.alt<-glm(CC ~ snp.geno, family=binomial, data=ceph.data, na.action=na.omit)
  model.anova<-anova(model.null, model.alt)
  snp.stat<-round(model.anova[2,4],3)
  snp.test<-1-pchisq(snp.stat,model.anova[2,3])
  print(c(colnames(ceph.data[i]),round(model.alt$coef,2), round(snp.stat,2), round(snp.test,2)))
}
```
### Scanning all Markers

<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>snp.geno1</th>
<th>snp.geno2</th>
<th>stat</th>
<th>pval</th>
<th>R output</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;snp1&quot;</td>
<td>&quot;0.04&quot;</td>
<td>&quot;-0.12&quot;</td>
<td>&quot;0.18&quot;</td>
<td>&quot;0.21&quot;</td>
<td>&quot;0.9&quot;</td>
</tr>
<tr>
<td>&quot;snp2&quot;</td>
<td>&quot;0.86&quot;</td>
<td>&quot;-0.58&quot;</td>
<td>&quot;-0.86&quot;</td>
<td>&quot;1.37&quot;</td>
<td>&quot;0.5&quot;</td>
</tr>
<tr>
<td>&quot;snp3&quot;</td>
<td>&quot;0.66&quot;</td>
<td>&quot;-0.47&quot;</td>
<td>&quot;-0.66&quot;</td>
<td>&quot;0.61&quot;</td>
<td>&quot;0.74&quot;</td>
</tr>
<tr>
<td>&quot;snp4&quot;</td>
<td>&quot;-0.22&quot;</td>
<td>&quot;0.92&quot;</td>
<td>&quot;15.79&quot;</td>
<td>&quot;5.11&quot;</td>
<td>&quot;0.08&quot;</td>
</tr>
<tr>
<td>&quot;snp5&quot;</td>
<td>&quot;0.78&quot;</td>
<td>&quot;-0.67&quot;</td>
<td></td>
<td>&quot;1.33&quot;</td>
<td>&quot;0.25&quot;</td>
</tr>
<tr>
<td>&quot;snp6&quot;</td>
<td>&quot;0.69&quot;</td>
<td>&quot;-0.37&quot;</td>
<td>&quot;-0.29&quot;</td>
<td>&quot;0.4&quot;</td>
<td>&quot;0.82&quot;</td>
</tr>
<tr>
<td>&quot;snp7&quot;</td>
<td>&quot;-0.07&quot;</td>
<td>&quot;0.32&quot;</td>
<td>&quot;-0.49&quot;</td>
<td>&quot;2.56&quot;</td>
<td>&quot;0.28&quot;</td>
</tr>
<tr>
<td>&quot;snp8&quot;</td>
<td>&quot;0.47&quot;</td>
<td>&quot;-0.32&quot;</td>
<td>&quot;-0.69&quot;</td>
<td>&quot;1.75&quot;</td>
<td>&quot;0.42&quot;</td>
</tr>
<tr>
<td>&quot;snp9&quot;</td>
<td>&quot;0.75&quot;</td>
<td>&quot;-0.55&quot;</td>
<td>&quot;-0.05&quot;</td>
<td>&quot;0.86&quot;</td>
<td>&quot;0.65&quot;</td>
</tr>
<tr>
<td>&quot;snp10&quot;</td>
<td>&quot;0.43&quot;</td>
<td>&quot;-0.09&quot;</td>
<td></td>
<td>&quot;0.02&quot;</td>
<td>&quot;0.89&quot;</td>
</tr>
<tr>
<td>&quot;snp11&quot;</td>
<td>&quot;1.79&quot;</td>
<td>&quot;-2.08&quot;</td>
<td>&quot;-19.36&quot;</td>
<td>&quot;18.6&quot;</td>
<td>&quot;&lt;0.01&quot;</td>
</tr>
<tr>
<td>&quot;snp12&quot;</td>
<td>&quot;1.39&quot;</td>
<td>&quot;-0.64&quot;</td>
<td>&quot;-2.37&quot;</td>
<td>&quot;8.12&quot;</td>
<td>&quot;0.02&quot;</td>
</tr>
<tr>
<td>&quot;snp13&quot;</td>
<td>&quot;1.39&quot;</td>
<td>&quot;-0.64&quot;</td>
<td>&quot;-2.37&quot;</td>
<td>&quot;8.12&quot;</td>
<td>&quot;0.02&quot;</td>
</tr>
<tr>
<td>&quot;snp14&quot;</td>
<td>&quot;0.46&quot;</td>
<td>&quot;0.53&quot;</td>
<td>&quot;0.95&quot;</td>
<td>&quot;2.87&quot;</td>
<td>&quot;0.24&quot;</td>
</tr>
<tr>
<td>&quot;snp15&quot;</td>
<td>&quot;1.47&quot;</td>
<td>&quot;-1.18&quot;</td>
<td>&quot;-1.47&quot;</td>
<td>&quot;3.64&quot;</td>
<td>&quot;0.16&quot;</td>
</tr>
<tr>
<td>&quot;snp16&quot;</td>
<td>&quot;1.47&quot;</td>
<td>&quot;-1.18&quot;</td>
<td>&quot;-1.47&quot;</td>
<td>&quot;3.64&quot;</td>
<td>&quot;0.16&quot;</td>
</tr>
<tr>
<td>&quot;snp17&quot;</td>
<td>&quot;0.14&quot;</td>
<td>&quot;-0.19&quot;</td>
<td>&quot;-0.84&quot;</td>
<td>&quot;1.33&quot;</td>
<td>&quot;0.51&quot;</td>
</tr>
<tr>
<td>&quot;snp18&quot;</td>
<td>&quot;1.32&quot;</td>
<td>&quot;-1.46&quot;</td>
<td>&quot;0.47&quot;</td>
<td>&quot;7.22&quot;</td>
<td>&quot;0.03&quot;</td>
</tr>
<tr>
<td>&quot;snp19&quot;</td>
<td>&quot;0.6&quot;</td>
<td>&quot;-0.11&quot;</td>
<td>&quot;-0.6&quot;</td>
<td>&quot;0.18&quot;</td>
<td>&quot;0.91&quot;</td>
</tr>
<tr>
<td>&quot;snp20&quot;</td>
<td>&quot;0.31&quot;</td>
<td>&quot;-0.28&quot;</td>
<td>&quot;-0.85&quot;</td>
<td>&quot;1.97&quot;</td>
<td>&quot;0.37&quot;</td>
</tr>
</tbody>
</table>
Multilocus Models

- Can test the effects of fitting two or more markers simultaneously
- Several multilocus models are possible
- Interaction Model assumes that each combination of genotypes has a different effect
- \( CC \sim SNP_1 + SNP_2 + SNP_1*SNP_2 + \ldots + SNP_i + \ldots \)
Two-Locus Model with interaction

- \( CC \sim SNP1 + SNP2 + SNP1*SNP2 \)

**R code:**

```r
snp.geno1<-codominant(genotype(snp1))
> snp.geno2<-codominant(genotype(snp2))
> null.model<-glm(CC~ snp.geno1+snp.geno2,family=binomial, data=ceph.data)
> alt.model<-glm(CC~ snp.geno1+snp.geno2+ snp.geno1:snp.geno2,family=binomial, data=ceph.data)
> print(model.anova<-anova(model.null, model.alt))
Analysis of Deviance Table

Model 1: CC ~ 1
Model 2: CC ~ snp.geno
   Resid. Df Resid. Dev  Df Deviance
1     105    146.947
2     103    144.982   2    1.966
> print(snp.test<-1-pchisq(snp.stat,model.anova[2,3]))
[1] 0.3741869
```
Scanning all two-way interactions

R code:

```r
# scanning all two-way interactions
write(c("snp1","snp2","stat","pval"),paste(my.directory,"scan_interactions.txt",sep=""), ncol=4, sep="\t")

for( i in 5:23 ){
  k<-i+1
  for( j in k:24 ){
    col1<-colnames(ceph.data)[i]
    col2<-colnames(ceph.data)[j]
    snp.geno1<-dominant(genotype(ceph.data[,i]))
    snp.geno2<-dominant(genotype(ceph.data[,j]))
    model.null<-glm(CC~ snp.geno1+snp.geno2,family=binomial, data=ceph.data, na.action=na.omit)
    model.alt<-glm(CC~ snp.geno1+snp.geno2+ snp.geno1:snp.geno2,family=binomial, data=ceph.data, na.action=na.omit)
    model.anova<-anova(model.null, model.alt)
    snp.stat<-round(model.anova[2,4],3)
    snp.test<-1-pchisq(snp.stat,model.anova[2,3])
    write(c(col1,col2,snp.stat,snp.test),paste(my.directory,"scan_interactions.txt",sep=""), ncol=4, append=T, sep="\t")
  }
}
```

Day 2 Section 6
Multiple Testing

- Take care interpreting significance levels when performing multiple tests
- Linkage disequilibrium can reduce the effective number of independent tests
- Permutation is a safe procedure to determine significance
- Repeat j=1..N times:
  - Permute disease status y between individuals
  - Fit all markers
  - Record maximum deviance maxdev[j] over all markers

- Permutation p-value for a marker is the proportion of times the permuted maximum deviance across all markers exceeds the observed deviance for the marker
Variable selection

**R code:**

```r
model.fit<-glm(CC~ 1,family=binomial, data=ceph.data)
fit.step <- stepAIC(model.fit, direction="forward", scope = list(upper = ~snp1+snp4+snp14+snp17+snp20, lower = ~1))
```

**R output**

Start:  AIC=148.95
CC ~ 1

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>146.95</td>
<td>148.95</td>
</tr>
<tr>
<td>+ snp4</td>
<td>141.84</td>
<td>147.84</td>
</tr>
<tr>
<td>+ snp14</td>
<td>144.08</td>
<td>150.08</td>
</tr>
<tr>
<td>+ snp20</td>
<td>144.98</td>
<td>150.98</td>
</tr>
<tr>
<td>+ snp17</td>
<td>145.62</td>
<td>151.62</td>
</tr>
<tr>
<td>+ snp1</td>
<td>146.74</td>
<td>152.74</td>
</tr>
</tbody>
</table>

Step1:  AIC=147.84
CC ~ snp4

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>141.84</td>
<td>147.84</td>
</tr>
<tr>
<td>+ snp17</td>
<td>139.64</td>
<td>149.64</td>
</tr>
<tr>
<td>+ snp20</td>
<td>139.81</td>
<td>149.81</td>
</tr>
<tr>
<td>+ snp14</td>
<td>140.31</td>
<td>150.31</td>
</tr>
<tr>
<td>+ snp1</td>
<td>141.72</td>
<td>151.72</td>
</tr>
</tbody>
</table>
Assessing model fit

- Hosmer-Lemeshow test is a commonly used test of goodness-of-fit of a binary logistic model.
- Idea: compares proportion of events with mean predicted probability within deciles of predicted P.

Problems:

- Arbitrary (number of groups, how to form groups).
- Low power (too many d.f.)
Model Validation

• Assess prediction in a new dataset

• use resampling techniques (Bootstrap, Cross-validation)
SNP and Haplotype

- DNA strand 1 differs from DNA strand 2 at a single base-pair location (a C/T polymorphism).

- Haplotype: the ordered allele sequence on a chromosome.
Haplotype Association

- Haplotype Association
  - Different from multiple genotype models
  - Phase taken into account
  - Haplotype association can be modelled in a similar logistic framework

- Treat haplotypes as extended alleles
- Fit additive, recessive, dominant & genotype models as before
  - Eg haplotypes are \( h = \text{AAGCAT, ATGCTT, etc} \)
  - \( y \sim \text{additive}(h) \)
  - \( y \sim \text{dominant}(h) \) etc
Haplotype association

R code:

```r
#### Haplotype analysis ####

snp1.a1<-allele(genotype(snp1),which=1)
snp1.a2<-allele(genotype(snp1),which=2)
snp2.a1<-allele(genotype(snp2),which=1)
snp2.a2<-allele(genotype(snp2),which=2)

haplo.snp1snp2<-data.frame(snp1.a1,snp1.a2,snp2.a1,snp2.a2)
haplo.mat<-setupGeno(haplo.snp1snp2, miss.val=c(0,NA))
haplo.data<-data.frame(haplo.mat,CC)

minhapfreq<-0.01
rarehap<-F

glm.haplo12<- haplo.glm(CC ~ haplo.mat, na.action="na.geno.keep",allele.lev=attributes(haplo.mat)$unique.alleles,
control=haplo.glm.control(haplo.freq.min=minhapfreq,keep.rare.haplo=rarehap),family = binomial,data=haplo.data)
print(glm.haplo12)
```

Day 2 Section 6
Haplotype analysis

R output

Call:
haplo.glm(formula = CC ~ haplo.mat, family = binomial, data = haplo.data, na.action = "na.geno.keep", allele.lev = attributes(haplo.mat)$unique.alleles,
  control = haplo.glm.control(haplo.freq.min = minhapfreq, keep.rare.haplo = rarehap))

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>coef</th>
<th>se</th>
<th>t.stat</th>
<th>pval</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-7.7e-09</td>
<td>0.277</td>
<td>-2.78e-08</td>
<td>1</td>
</tr>
<tr>
<td>haplo.mat</td>
<td>1.2e-08</td>
<td>0.307</td>
<td>3.90e-08</td>
<td>1</td>
</tr>
</tbody>
</table>

Haplotypes:

<table>
<thead>
<tr>
<th>loc.1</th>
<th>loc.2</th>
<th>hap.freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>haplo.mat.2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>haplo.base</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Only 1 haplo created!

Note: Use weights (haplo freq.) to compute the predicted values
Summary

• Various genetic models possible
• Choice of genotype or haplotype based analyses
• Variable selection can help localizing the disease locus
• Multilocus are difficult to fit: many variables
• Model validation is important