Cherry Bud Workshop 2006
Building Models from Data
30 March 2006
in Keio University (Yagami campus)

StaGen Co., Ltd.
Statistical Genetics Analysis Division

Estimation of haplotype associated with several quantitative phenotypic values based on maximization of area under ROC curve

Shigeo Kamitsuji, PhD
Statistical Genetics Analysis Division
StaGen Co., Ltd.
kamitsuji@stagen.co.jp

QTL (Quantitative Trait Locus, loci) analysis

- Aim of QTL analysis
  - To find a locus, loci or haplotype associated with a quantitative phenotype.
- QTL analysis has a long history.
  - Galton (1869)
    - Introducing the notion of regression
  - Fisher (1917)
    - Quantitative phenotypic value is the result of genetic factor, environmental factor, and the interaction of environmental factor with genetic factor

Recent works for QTL analysis

- Yi et al. (2003)
  - Linear Model + MCMC
- Thomas et al. (2004)
  - Graphical Modeling
- Sebastiani et al. (2005)
  - Application of Bayesian Network
- Shibata et al. (2004)
  - [illegible]

Complicated!

Our product

e-mail: kamitsuji@stagen.co.jp

Objectives

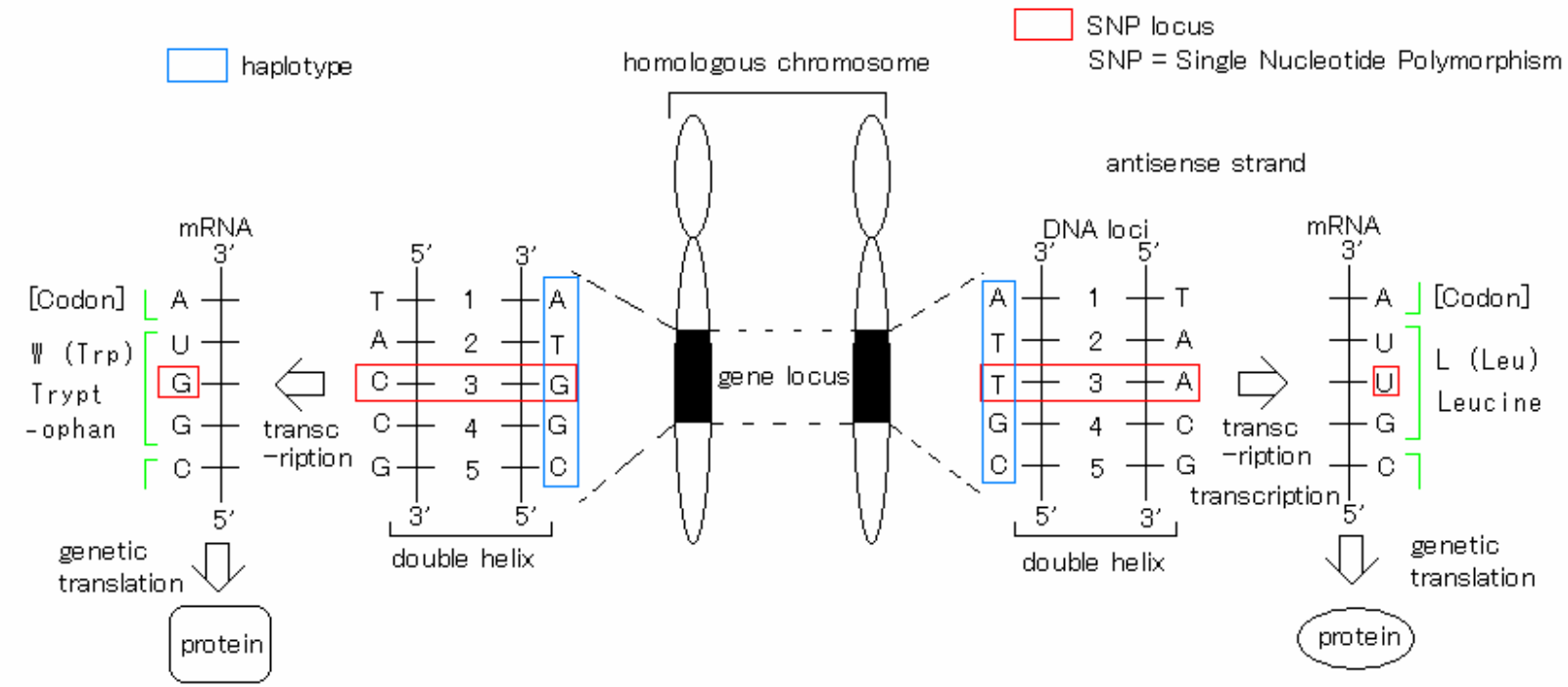
- To estimate a haplotype associated with several quantitative phenotypes
  - Definition of the mixtured phenotype
  - Introduction to the notion of ROC curve and AUC
  - Development of algorithm based on the maximization of AUC
- To validate the effectiveness of our method
  - Analysis of real data; genotypes and phenotypes data for the diabetes patients
  - Comparison with other models, QTLhaplo, Generalized Linear Model and Neural Network

Outline

- Material and Methods
  - ROC curve and AUC
  - Maximization of AUC method
    - Mixtured phenotypes
    - Normality of AUC
    - Algorithm for maximization of AUC
    - Estimating haplotype associated with phenotypes
- Results
  - Example data
  - Real data analysis
    - Genotype and phenotypes data for diabetes
    - Comparison with other models
      - Generalized Linear Model and Neural Network

Materials and Methods

Knowledge:
Locus, Allele, SNP, Genotype, Haplotype



Allele

Observable alelles on SNP locus : G, C

⇒ bi-allelic locus

Genotype

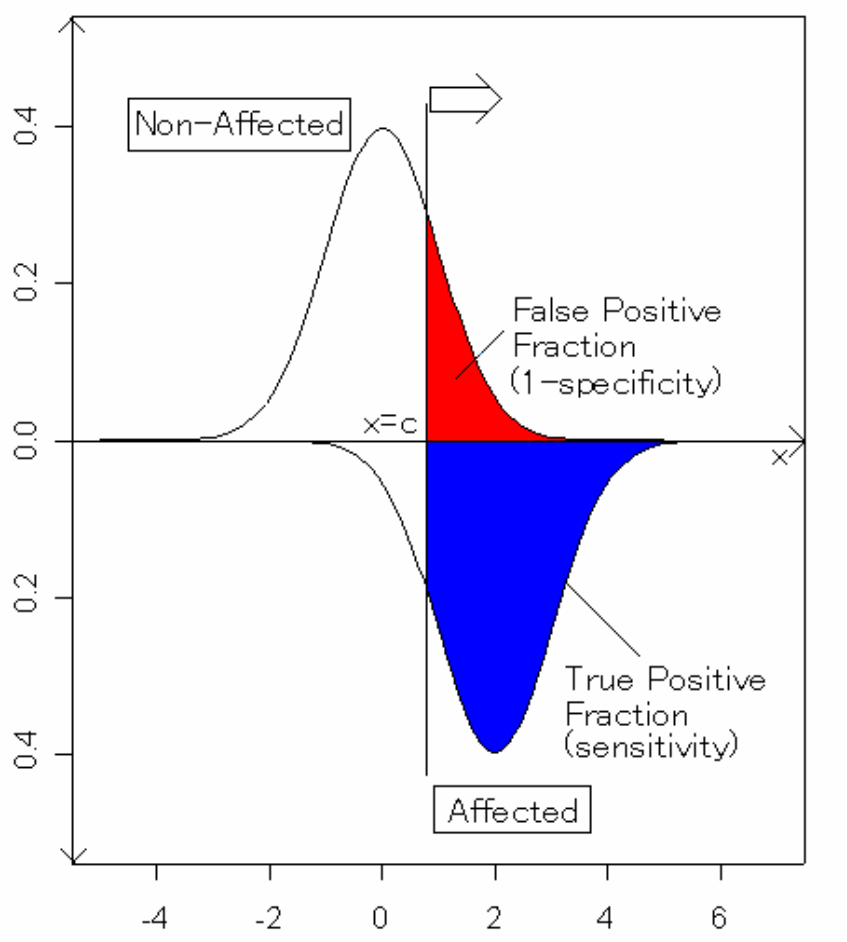
un-ordered : G/G, G/C, C/C

ordered : G-G, G-C, C-G, C-C

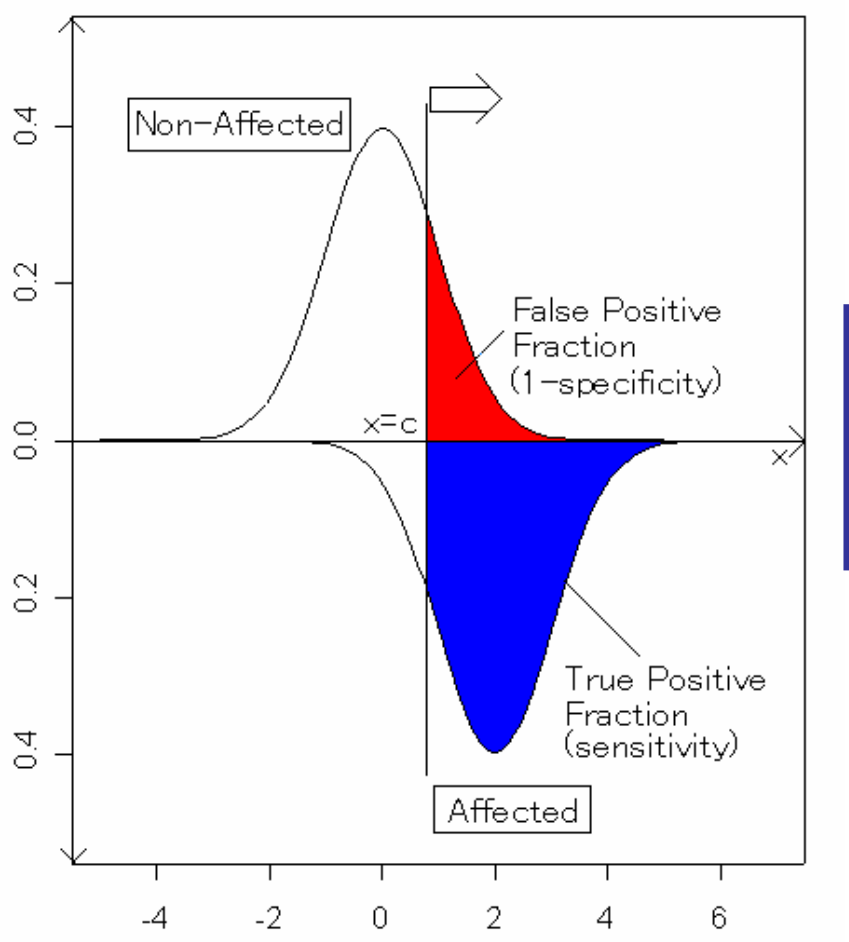
Receiver Operating Characteristic curve (ROC curve) and Area Under the ROC curve (AUC)

- ROC curve
  - A plot to show the trade-off between sensitivity and specificity
  - It is used for evaluating a diagnostic test.
- AUC
  - For evaluating the accuracy of diagnostic test
  - The AUC value varies from 0 to 1, and being close to 1 when the diagnostic test has a high degree of accuracy.

ROC curve and AUC

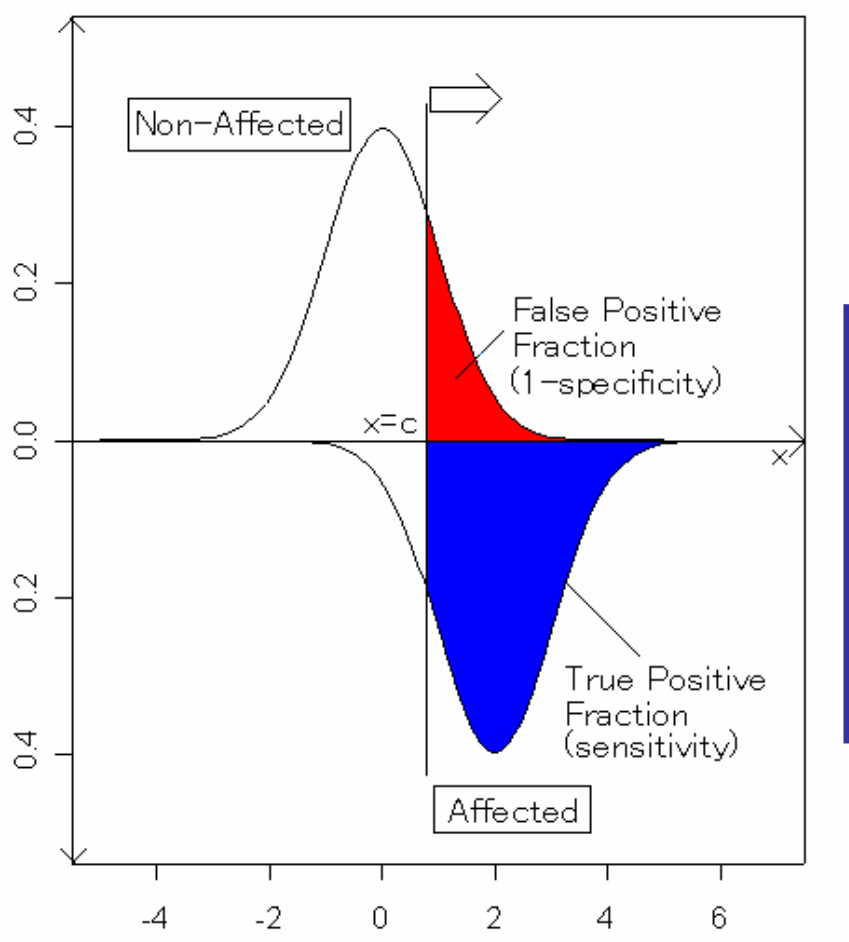


ROC curve and AUC



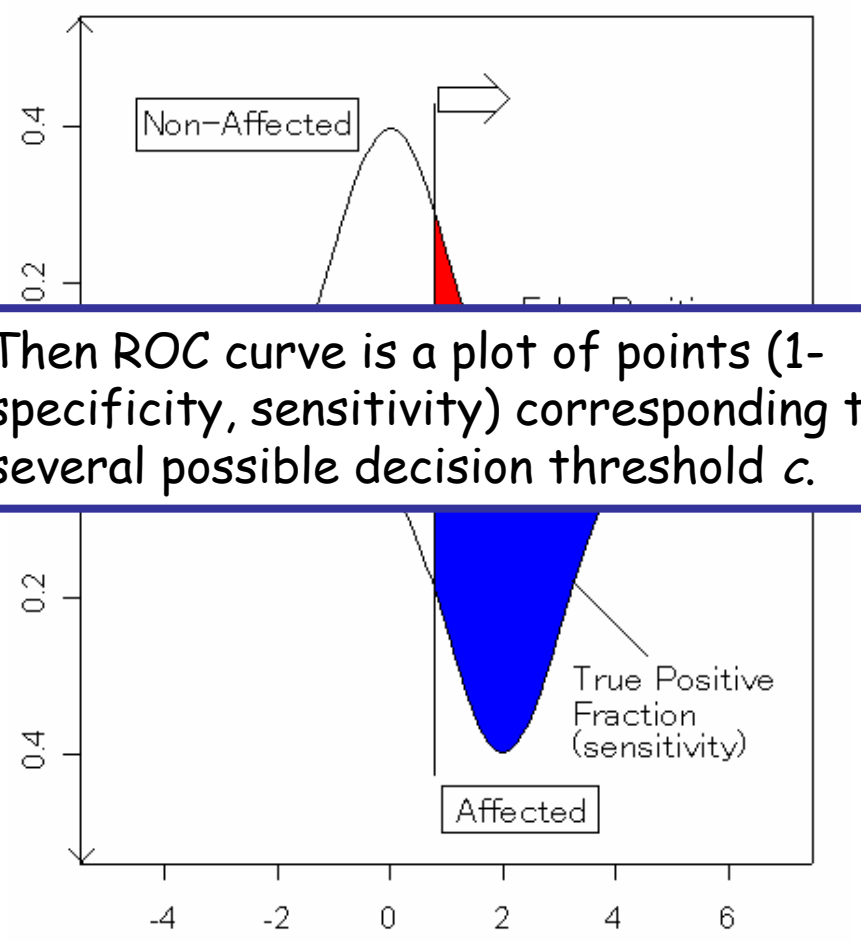
Consider two distributions of a numerical measurement corresponding to non-affected patients and affected patients.

ROC curve and AUC

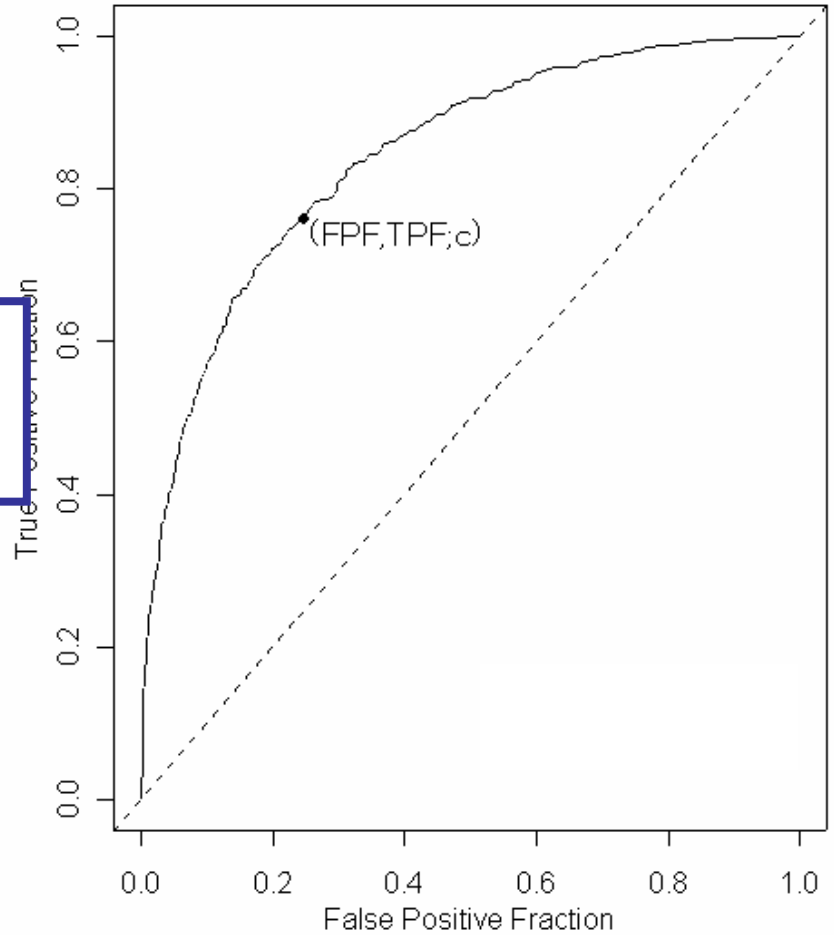


To describe the sensitivity and specificity, we choose a value of threshold c, in which case the patients with gap values greater than *c* are labeled positive and patients with gap values less than or equal to *c* are labeled negative.

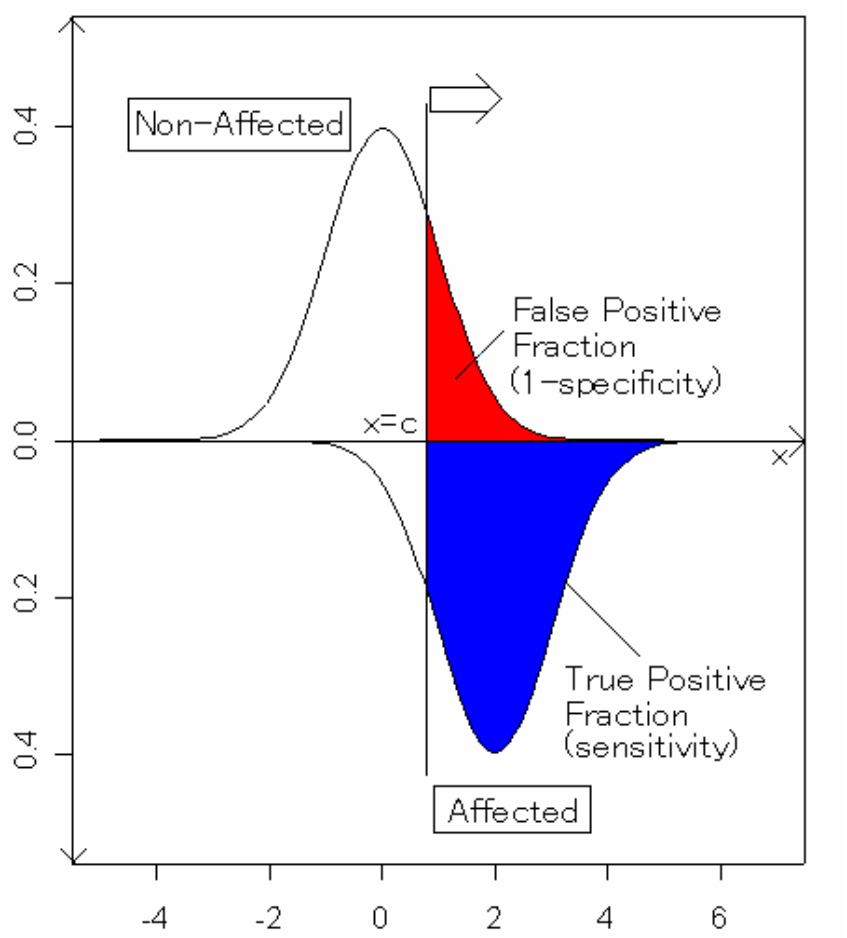
ROC curve and AUC

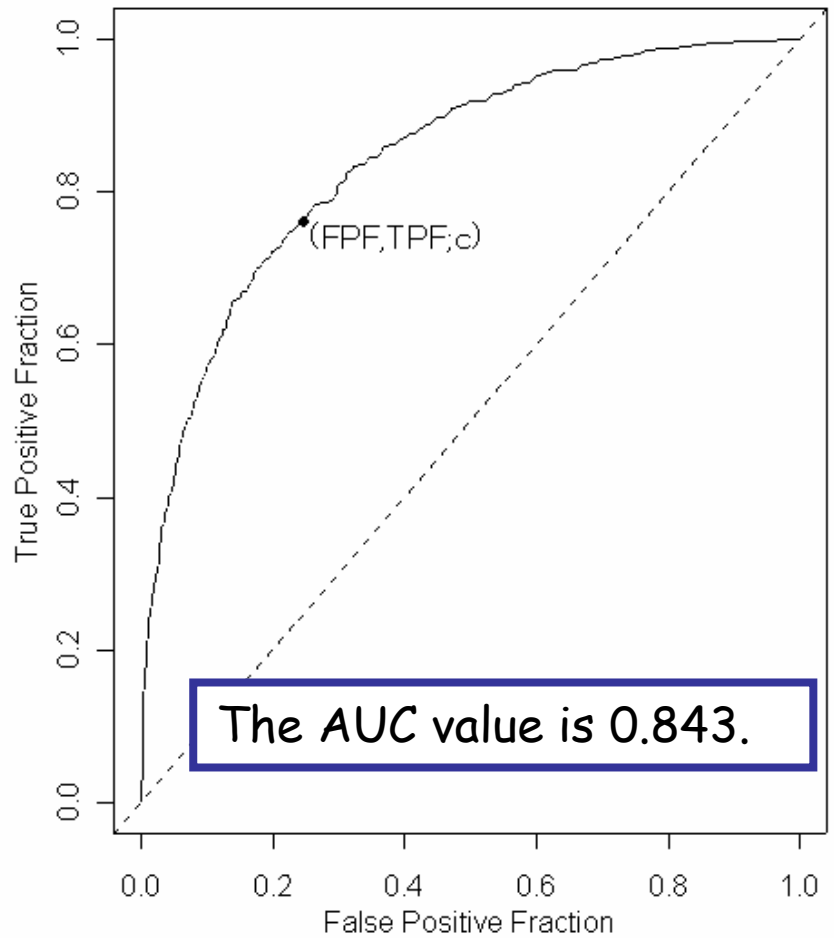


Then ROC curve is a plot of points (1-specificity, sensitivity) corresponding to several possible decision threshold *c*.



ROC curve and AUC





Mixture of quantitative phenotypes

- Notations
  - $\mathbf{x}_j,\ j = 1, \ldots, p$: a quantitative phenotype as an *n*-dimensional vector (*n* observations)
- Mixture of quantitative phenotypes
  - Mixtured phenotype ***z*** can be written as a linear combination of quantitative phenotypes,

$$\mathbf{z} = \sum_{j=1}^{p} \beta_j \mathbf{x}_j, \quad \beta_j,\ j = 1, \ldots p \text{ is the coefficient of } \mathbf{x}_j.$$

Maximization of AUC

- To find the haplotype associated with several quantitative phenotypes
  - Divide mixtured phenotypes $\boldsymbol{z}$ into two parts: $\boldsymbol{z}|G_1$ and $\boldsymbol{z}|G_2$: $\boldsymbol{z}$ for individuals with a haplotype (G1) and without the haplotype (G2)
  - By calculating the AUC value between $\boldsymbol{z}|G_1$ and $\boldsymbol{z}|G_2$, we evaluate the strength of the association between mixtured phenotype and the haplotype.
- Maximization of AUC
  - Estimate the coefficients $\beta_j,\ j=1,\ldots,p$ so as to maximize the AUC value between $\boldsymbol{z}|G_1$ and $\boldsymbol{z}|G_2$.

Normality of ROC curve and AUC (McClish, 1989)

- ROC curve and AUC can be represented by the standard normal distribution function

$$X \sim \mathrm{N}(\mu_X, {\sigma_X}^2),\ Y \sim \mathrm{N}(\mu_Y, {\sigma_Y}^2), \Phi(z) = \Pr(Z \le z),\ Z \sim \mathrm{N}(0,1)$$

$$\mathrm{FPF}(=1-\mathrm{specificity}) = \Pr(X > c) = 1 - \Phi\left(\frac{c-\mu_X}{\sigma_X}\right) =: \xi,$$

$$\mathrm{TPF}(=\mathrm{sensitivity}) = \Pr(Y > c) = 1 - \Phi\left(\frac{c-\mu_Y}{\sigma_Y}\right) =: 1-\eta,$$

$$\mathrm{ROC\,curve}: \eta = \Phi\left(\frac{\sigma_X \Phi^{-1}(1-\xi) + \mu_X - \mu_Y}{\sigma_Y}\right),$$

$$\mathrm{AUC} = \Phi\left(\frac{\mu_Y - \mu_X}{\sqrt{{\sigma_X}^2 + {\sigma_Y}^2}}\right)$$

Representation of ROC curve and AUC

- Assumption
  - Each quantitative phenotype $\mathbf{x}_j|G_k$, j=1, …, p, k=1, 2 given the condition G_k is i.i.d, and $\mathbf{x}_j|G_k$ is normally distributed,

$$\mathbf{x}_j \mid G_k \sim N(\mu_{j,k}, \sigma^2_{j,k}),\ k = 1, 2,$$

$$\mu_{j,k} = \mathrm{E}[\mathbf{x}_j \mid G_k],\ \sigma^2_{j,k} = \mathrm{Var}[\mathbf{x}_j \mid G_k].$$

Representation of ROC curve and AUC

- Mixtured phenotype
  - Then mixture phenotype $\mathbf{z}|G_k$, k=1,2 is also normally distributed,

$$\mathbf{z} \mid G = G_k \sim N\left(\sum_{j=1}^{p} \beta_j \mu_{j,k}, \sum_{j=1}^{p} \beta_j^2 \sigma_{j,k}^2 \right)$$
$$=: N\left(\mu_{z,k}(\boldsymbol{\beta}), \sigma_{z,k}^2(\boldsymbol{\beta})\right), \quad \boldsymbol{\beta} = (\beta_1, \ldots, \beta_1)^T.$$

- AUC $F(\beta)$ as the function of β

$$F(\boldsymbol{\beta}) := \Phi\left(\frac{a(\boldsymbol{\beta})}{\sqrt{1+b^2(\boldsymbol{\beta})}} \right), \quad a(\boldsymbol{\beta}) := \frac{\mu_{z,2}(\boldsymbol{\beta}) - \mu_{z,1}(\boldsymbol{\beta})}{\sigma_{z,2}(\boldsymbol{\beta})}, \quad b(\boldsymbol{\beta}) = \frac{\sigma_{z,1}(\boldsymbol{\beta})}{\sigma_{z,2}(\boldsymbol{\beta})}$$

Algorithm for updating the coefficients

- Quasi-Newton method

$$\boldsymbol{\beta}^{[t+1]} = \boldsymbol{\beta}^{[t]} - \left(\underline{f(\boldsymbol{\beta}^{[t]})}' \right)^{-1} \underline{f(\boldsymbol{\beta}^{[t]})}$$

  - It is necessary to iterate the updating when the convergence of β is observed.

Algorithm for updating the coefficients

Here

$$\underline{f(\boldsymbol{\beta})} := \frac{d}{d\boldsymbol{\beta}} F(\boldsymbol{\beta}) = \phi\left(\frac{a(\boldsymbol{\beta})}{\sqrt{1+b^2(\boldsymbol{\beta})}}\right) \cdot \frac{\underline{a'(\boldsymbol{\beta})}(1+b^2(\boldsymbol{\beta})) - 0.5\, a(\boldsymbol{\beta})\,\underline{(b^2(\boldsymbol{\beta}))'}}{(1+b^2(\boldsymbol{\beta}))^{1.5}}$$

$$\left(\underline{f(\boldsymbol{\beta}^{[t]})}\right)' := \frac{d}{d\boldsymbol{\beta}} f(\boldsymbol{\beta}) = diag(\underline{\Delta})^{-1} diag(\underline{f(\boldsymbol{\beta}+\underline{\Delta})} - \underline{f(\boldsymbol{\beta})}),\ \underline{\Delta} = (\Delta,...,\Delta)^T$$

$$\underline{a'(\boldsymbol{\beta})} = \frac{-\sigma^2_{z,2}(\boldsymbol{\beta})(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - \{\mu_{z,2}(\boldsymbol{\beta}) - \mu_{z,1}(\boldsymbol{\beta})\}\Sigma_{x,2}\boldsymbol{\beta}}{\sigma^3_{z,2}},$$

$$\underline{\left(b^2(\boldsymbol{\beta})\right)'} = \frac{2\Sigma_{x,1}\boldsymbol{\beta}\,\sigma^2_{z,2}(\boldsymbol{\beta}) - 2\Sigma_{x,2}\boldsymbol{\beta}\,\sigma^2_{z,1}(\boldsymbol{\beta})}{\sigma^4_{z,2}(\boldsymbol{\beta})},$$

$$\sigma^2_{z,k}(\boldsymbol{\beta}) = \boldsymbol{\beta}^T\Sigma_{x,k}\boldsymbol{\beta}, \quad \boldsymbol{\mu}_k = (\mu_{1,k},...,\mu_{p,k})^T, \quad \Sigma_{x,k} = diag(\sigma^2_{1,k},...,\sigma^2_{p,k}),$$

Test of AUC

- Whether the AUC value is significantly large?
  - AUC becomes 0.5 in the case of two identical distributions
- Hypothesis test
  - Null hypothesis H_0: AUC = 0.5
  - Alternative hypothesis H_1: AUC≠0.5

Test of AUC

- Test statistic (Zhou et al. 2002)

$$Z = \frac{F(\hat{\boldsymbol{\beta}}) - 0.5}{\sqrt{Var[F(\hat{\boldsymbol{\beta}})]}} \sim N(0,1).$$

Here

$$Var[F(\hat{\boldsymbol{\beta}})] = g_1(\hat{\boldsymbol{\beta}})^2 Var[a(\hat{\boldsymbol{\beta}})] + g_2(\hat{\boldsymbol{\beta}})^2 Var[b(\hat{\boldsymbol{\beta}})] + 2g_1(\hat{\boldsymbol{\beta}})g_2(\hat{\boldsymbol{\beta}})Cov[a(\hat{\boldsymbol{\beta}}), b(\hat{\boldsymbol{\beta}})],$$

$$g_1(\boldsymbol{\beta}) = \frac{\exp\left(-a(\boldsymbol{\beta})^2 / 2(1+b(\boldsymbol{\beta})^2)\right)}{\sqrt{2\pi(1+b(\boldsymbol{\beta})^2)}},\ g_2(\boldsymbol{\beta}) = -\frac{a(\boldsymbol{\beta})b(\boldsymbol{\beta})\exp\left(-a(\boldsymbol{\beta})^2 / 2(1+b(\boldsymbol{\beta})^2)\right)}{\sqrt{2\pi(1+b(\boldsymbol{\beta})^2)^3}},$$

$$Var[a(\boldsymbol{\beta})] = \frac{n_2(a(\boldsymbol{\beta})^2 + 2) + 2n_1 b(\boldsymbol{\beta})^2}{2n_1 n_2},\ Var[b(\boldsymbol{\beta})] = \frac{(n_1+n_2)b(\boldsymbol{\beta})^2}{2n_1 n_2},$$

$$Cov[a(\boldsymbol{\beta}), b(\boldsymbol{\beta})] = \frac{a(\boldsymbol{\beta})b(\boldsymbol{\beta})}{2n_1},$$

n_j : the number of elements in G_j.

Algorithm for haplotype estimation

- Haplotype Estimation (HE) step
  - HE-Step 1
    - The individuals in the data are divided into two parts: individuals with a haplotype h_i (G_1) and without the haplotype h_i (G_2).
  - HE-Step 2
    - By using MARC method, the coefficient β is estimated so as to maximize the value of AUC $F(\beta)$ between $\boldsymbol{z}|G_1$ and $\boldsymbol{z}|G_2$.
  - For all hi$\in$H, coefficient β is estimated and $F(\beta)$ is calculated.
    - H is denoted as a set of all possible haplotypes from data.
- Model Selection (MS) step
  - MS-Step 1
    - Consider a model that omits the single term $\boldsymbol{x}_j$ from current model $\boldsymbol{z}$.
  - MS-Step 2
    - Re-estimate β by HE-Step 1 and 2 for each omitted model.
  - If AUC value for omitted model is larger than that for previous model, MS-Steps are continued.

QTLMARC
(QTL Maximization of Area under the ROC Curve)

Results

Example: Plane for partitioning

- For the help to understand the coefficients β
  - The number of samples is 1,000.

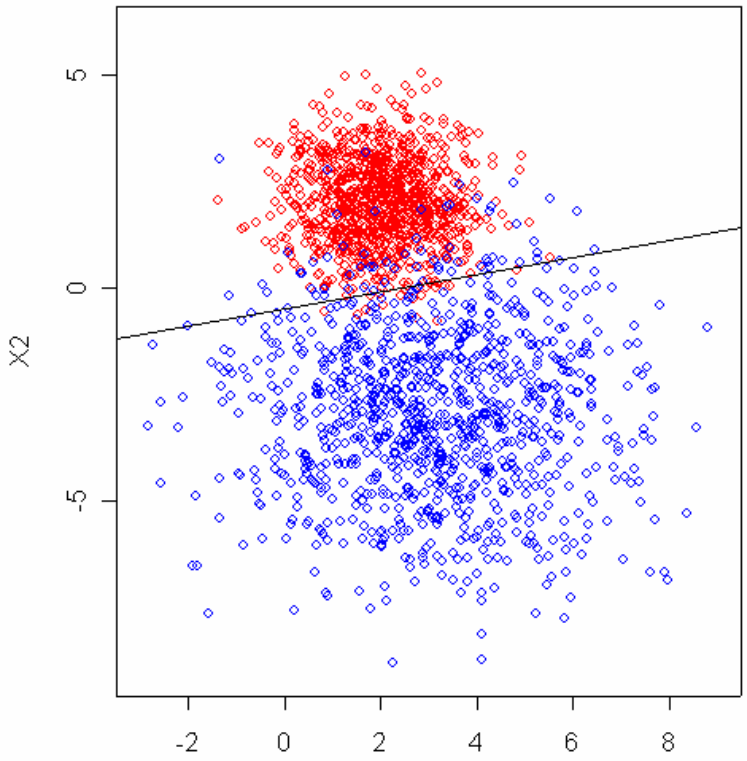
$$\mathbf{X}_1 \mid G_1 \sim N(2,1^2), \quad \mathbf{X}_2 \mid G_1 \sim N(3,\sqrt{2}^2),$$

$$\mathbf{X}_1 \mid G_2 \sim N(-2,1^2), \; \mathbf{X}_2 \mid G_2 \sim N(-3,\sqrt{2}^2),$$

$$\mathbf{Z} = \beta_0 \mathbf{1} + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2, \;\; \mathbf{Z} \mid G_1 \; vs. \; \mathbf{Z} \mid G_2$$

- Estimation of β

Intercept	β_1	β_2
-0.5000	-1.3957	6.9527



Plane

$$-0.5000 \times 6.9527 = -1.3957\, X_1 + 6.9527\, X_2$$

Analysis of real data

- Diabetes data (Iwasaki et al., 2005)
  - 3 SNP data on Calpine-10 gene
    - three bi-allelic loci with alleles 1 and 2
  - 12 quantitative phenotypes
    - height, weight, Body Mass Index (BMI), Hemoglobin A1c (HbA1c)
    - Blood glucose levels (BS) at 0 min (BS0), 30 min (BS30), 60 min (BS60), and 120 min (BS120)
    - Immunoreactive insulino levels (IRI) at 0 min (IRI0), 30 min (IRI30), 60 min (IRI60), and 120 min (IRI120)
  - 2 others
    - Sex, and Age

Aim

- Estimation of the haplotype associated with phenotypes
  - The strength of association of each phenotype is also evaluated.

Normality of phenotypes

- It is necessary to evaluate whether each quantitative phenotype we here used is normally distributed.
  - AUC function can be written as the standard normal distribution function based on the assumption of normality of phenotypes.
- 8 Phenotypes are used
  - Weight, BMI, BS0, BS30, BS60, BS120, IRI0, and HbA1c
    - Quantile-Quantile plot (QQ plot)

Homogeneity of data

- Variable Age is a risk factor in diabetes
  - Iwasaki et al. 2005
  - Since Age is not a phenotype, Therefore it is inappropriate to add to the mixtured phenotype **z**.
- Divide the data
  - The diabetes data is divided into two parts: patients aged 50 years and over, and those aged less than 50 years.
  - In this study, diabetes data corresponding to patients aged ≥ 50 years is applied.

Results for data from patients aged ≥ 50 years

age>=50 ~BMI+BS0+BS60+BS120

haplotype	Number of carriers	Coefficients				AUC	p-value
		BMI	BS0	BS60	BS120		
221	9	7.1367	9.4493	−26.2790	−8.4942	0.7116	0.0041
222	4	1.3692	3.4324	−3.2579	−4.0592	0.7087	0.0026
111	64	0.1532	−0.6446	1.0344	−0.5304	0.6954	0.0000
211	4	−0.2473	2.1815	−1.5759	0.3523	0.6687	0.0509
112	54	1.1523	0.8362	3.1134	0.1982	0.6489	0.0004
212	3	−0.8632	0.2621	−0.2221	−0.7076	0.6254	0.2396
122	42	−0.0572	0.0499	−8.9660	−0.1150	0.6224	0.0058
121	108	−18.2788	2.2698	0.9149	−10.2325	0.5988	0.0837

Results for data from patients aged ≥ 50 years

age>=50 ~BMI+BS0+BS60+BS120

haplotype	Number of carriers	Coefficients				AUC	p-value
		BMI	BS0	BS60	BS120		
221	9	7.1367	9.4493	-26.2790	-8.4942	0.7116	0.0041
222	4	1.3692	3.4324	-3.2579	-4.0592	0.7087	0.0026
111	64	0.1532	-0.6446	1.0344	-0.5304	0.6954	0.0000
211	4	-0.2473	2.1815	-1.5759	0.3523	0.6687	0.0509
112	54	1.1523	0.8362	3.1134	0.1982	0.6489	0.0004
212	3	-0.8632	0.2621	-0.2221	-0.7076	0.6254	0.2396
122	42	-0.0572	0.0499	-8.9660	-0.1150	0.6224	0.0058
121	108	-18.2788	2.2698	0.9149	-10.2325	0.5988	0.0837

- Haplotype 111 gives the minimum P value.
  - The haplotype 111 has a high diabetes risk in older Japanese subjects (Iwasaki et al., 2005.)

Results for data from patients aged ≥ 50 years

age>=50	~BMI+BS0+BS60+BS120						
haplotype	Number of carriers	Coefficients				AUC	p-value
		BMI	BS0	BS60	BS120		
221	9	7.1367	9.4493	−26.2790	−8.4942	0.7116	0.0041
222	4	1.3692	3.4324	−3.2579	−4.0592	0.7087	0.0026
111	64	0.1532	−0.6446	1.0344	−0.5304	0.6954	0.0000
211	4	−0.2473	2.1815	−1.5759	0.3523	0.6687	0.0509
112	54	1.1523	0.8362	3.1134	0.1982	0.6489	0.0004
212	3	−0.8632	0.2621	−0.2221	−0.7076	0.6254	0.2396
122	42	−0.0572	0.0499	−8.9660	−0.1150	0.6224	0.0058
121	108	−18.2788	2.2698	0.9149	−10.2325	0.5988	0.0837

- Haplotype 111 gives the minimum P value.
  - The haplotype 111 has a high diabetes risk in older Japanese subjects (Iwasaki et al., 2005.)
  - The coefficient of BS0 is the largest
    - The association of haplotype 111 with BS60 is stronger than any other phenotype

Results for data from patients aged ≥ 50 years

age>=50 ~BMI+BS0+BS60+BS120

haplotype	Number of carriers	Coefficients				AUC	p-value
		BMI	BS0	BS60	BS120		
221	9	7.1367	9.4493	−26.2790	−8.4942	0.7116	0.0041
222	4	1.3692	3.4324	−3.2579	−4.0592	0.7087	0.0026
111	64	0.1532	−0.6446	1.0344	−0.5304	0.6954	0.0000
211	4	−0.2473	2.1815	−1.5759	0.3523	0.6687	0.0509
112	54	1.1523	0.8362	3.1134	0.1982	0.6489	0.0004
212	3	−0.8632	0.2621	−0.2221	−0.7076	0.6254	0.2396
122	42	−0.0572	0.0499	−8.9660	−0.1150	0.6224	0.0058
121	108	−18.2788	2.2698	0.9149	−10.2325	0.5988	0.0837

- Haplotype 111 gives the minimum P value.
  - The haplotype 111 has a high diabetes risk in older Japanese subjects (Iwasaki et al., 2005.)
  - The coefficient of BS0 is the largest
    - The association of haplotype 111 with BS60 is stronger than any other phenotype
  - Coefficients of BMI and BS60 are positive and coefficients of BS0 and BS120 are negative.
    - If a patient has a haplotype 111, then BMI and BS60 tends to increasing.

Comparison with other models

- Variables
  - Response variable ***y***
    - Binary data for a patient with or without a haplotype h_i (1 or 0)
  - Explanatory variables
    - 13 phenotypes
- Models
  - Generalized Linear Model (GLM)
  - Deterministic Neural Network (NN)

Comparison with GLM

Haplotype	Number of carriers	Phenotypes							AIC
		Intercept	BMI	BS0	BS30	BS60	BS120	HbA1c	
212	3	−3.7590	0.2894	−0.1502	−0.3963	0.0632	0.4418	−0.7601	40.2117
222	4	−3.8055	−0.1932	−0.7961	−0.4085	0.4320	0.9490	−0.9102	41.4962
211	4	−3.4838	0.2994	−0.7166	−0.2411	0.7109	−0.2706	−0.8674	44.2789
221	9	−2.9488	−0.2738	−0.4436	−1.0408	1.5379	0.1287	−0.6048	66.6565
121	108	2.2292	−0.4876	0.1835	−0.0837	0.1544	−0.1105	−0.5136	104.6632
122	42	−0.8324	0.0962	−0.2290	0.0988	0.5748	−0.2166	0.0432	165.3384
111	64	0.0514	0.3474	−0.5259	0.0335	0.7983	−0.5611	0.1332	169.1515
112	54	−0.5748	0.3954	−0.1541	0.2278	0.6335	−0.3486	0.2299	169.2884

- Model selection based on AIC
  - 6 phenotypes are selected.

Comparison with GLM

Haplotype	Number of carriers	Phenotypes							AIC	AUC
		Intercept	BMI	BS0	BS30	BS60	BS120	HbA1c		
212	3	−3.7590	0.2894	−0.1502	−0.3963	0.0632	0.4418	−0.7601	40.2117	0.7507
222	4	−3.8055	−0.1932	−0.7961	−0.4085	0.4320	0.9490	−0.9102	41.4962	0.8031
211	4	−3.4838	0.2994	−0.7166	−0.2411	0.7109	−0.2706	−0.8674	44.2789	0.7271
221	9	−2.9488	−0.2738	−0.4436	−1.0408	1.5379	0.1287	−0.6048	66.6565	0.8053
121	108	2.2292	−0.4876	0.1835	−0.0837	0.1544	−0.1105	−0.5136	104.6632	0.6563
122	42	−0.8324	0.0962	−0.2290	0.0988	0.5748	−0.2166	0.0432	165.3384	0.6619
111	64	0.0514	0.3474	−0.5259	0.0335	0.7983	−0.5611	0.1332	169.1515	0.7022
112	54	−0.5748	0.3954	−0.1541	0.2278	0.6335	−0.3486	0.2299	169.2884	0.6981

- The goodness of fitted GLM is evaluated by AUC value.
  - The AUC value between two sets of fitted values for carrier and non-carrier is calculated.
- Minimization of AIC and the maximization of AUC are not equivalent as criteria of parameter estimation and model selection.
  - It is difficult to compare the results by QTLMARC and GLM.

Comparison with GLM

Haplotype	Number of carriers	Phenotypes							Predicted Probability	
		Intercept	BMI	BS0	BS30	BS60	BS120	HbA1c	non-carrier	carrier
212	3	-3.7590	0.2894	-0.1502	-0.3963	0.0632	0.4418	-0.7601	0.0165	0.0341
222	4	-3.8055	-0.1932	-0.7961	-0.4085	0.4320	0.9490	-0.9102	0.0135	0.0940
211	4	-3.4838	0.2994	-0.7166	-0.2411	0.7109	-0.2706	-0.8674	0.0176	0.0612
221	9	-2.9488	-0.2738	-0.4436	-1.0408	1.5379	0.1287	-0.6048	0.0360	0.1466
121	108	2.2292	-0.4876	0.1835	-0.0837	0.1544	-0.1105	-0.5136	0.8506	0.8901
122	42	-0.8324	0.0962	-0.2290	0.0988	0.5748	-0.2166	0.0432	0.3086	0.3702
111	64	0.0514	0.3474	-0.5259	0.0335	0.7983	-0.5611	0.1332	0.4442	0.5854
112	54	-0.5748	0.3954	-0.1541	0.2278	0.6335	-0.3486	0.2299	0.3762	0.4997

- The mean of the probability for the existence of a patient with or without the haplotype is calculated.
  - If it is appropriate to assume that the response variable y is distributed from binomial distribution.

Comparison with GLM

Haplotype	Number of carriers	Phenotypes							Predicted Probability	
		Intercept	BMI	BS0	BS30	BS60	BS120	HbA1c	non-carrier	carrier
212	3	-3.7590	0.2894	-0.1502	-0.3963	0.0632	0.4418	-0.7601	0.0165	0.0341
222	4	-3.8055	-0.1932	-0.7961	-0.4085	0.4320	0.9490	-0.9102	0.0135	0.0940
211	4	-3.4838	0.2994	-0.7166	-0.2411	0.7109	-0.2706	-0.8674	0.0176	0.0612
221	9	-2.9488	-0.2738	-0.4436	-1.0408	1.5379	0.1287	-0.6048	0.0360	0.1466
121	108	2.2292	-0.4876	0.1835	-0.0837	0.1544	-0.1105	-0.5136	0.8506	0.8901
122	42	-0.8324	0.0962	-0.2290	0.0988	0.5748	-0.2166	0.0432	0.3086	0.3702
111	64	0.0514	0.3474	-0.5259	0.0335	0.7983	-0.5611	0.1332	0.4442	0.5854
112	54	-0.5748	0.3954	-0.1541	0.2278	0.6335	-0.3486	0.2299	0.3762	0.4997

- It is hard to assume that the response variable is distributed from binomial distribution.
  - Sum of two probabilities is far to 1.

Comparison with GLM

Haplotype	Number of carriers	Phenotypes							Predicted Probability	
		Intercept	BMI	BS0	BS30	BS60	BS120	HbA1c	non-carrier	carrier
212	3	−3.7590	0.2894	−0.1502	−0.3963	0.0632	0.4418	−0.7601	0.0165	0.0341
222	4	−3.8055	−0.1932	−0.7961	−0.4085	0.4320	0.9490	−0.9102	0.0135	0.0940
211	4	−3.4838	0.2994	−0.7166	−0.2411	0.7109	−0.2706	−0.8674	0.0176	0.0612
221	9	−2.9488	−0.2738	−0.4436	−1.0408	1.5379*	0.1287	−0.6048	0.0360	0.1466
121	108	2.2292	−0.4876	0.1835	−0.0837	0.1544	−0.1105	−0.5136	0.8506	0.8901
122	42	−0.8324	0.0962	−0.2290	0.0988	0.5748	−0.2166	0.0432	0.3086	0.3702
111	64	0.0514	0.3474	−0.5259	0.0335	0.7983*	−0.5611*	0.1332	0.4442	0.5854
112	54	−0.5748	0.3954	−0.1541	0.2278*	0.6335	−0.3486	0.2299	0.3762	0.4997

- The GLMs for the halotype 111 and 112 are acceptable.
  - Sum of probabilities is close to 1
- BS60 in GLM for the haplotype 111 is significantly large.
  - In the table, "*" means that p-value obtained F-test by analysis of variance is smaller than 0.05.
  - The results obtained by QTLMARC indicate that the haplotype 111 is associated with BS60.

Comparison with NN

- Three-layer feed-forward neural network
  - one input layer with four neurons, one output layer with one neuron, and one hidden layer with three neurons.
  - Response variable y can be written as

$$y = \sum_{k=1}^{m} h_k \cdot f_k\left(\sum_{j=1}^{n} w_{jk} x_j\right),$$

where w_{jk} and h_k, $j = 1,...,4, k = 1,...,3$ are connection weights,

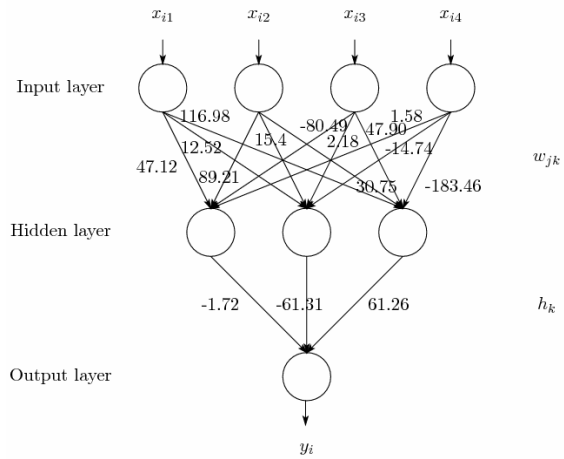
and $f(x)$ is the sigmoid function

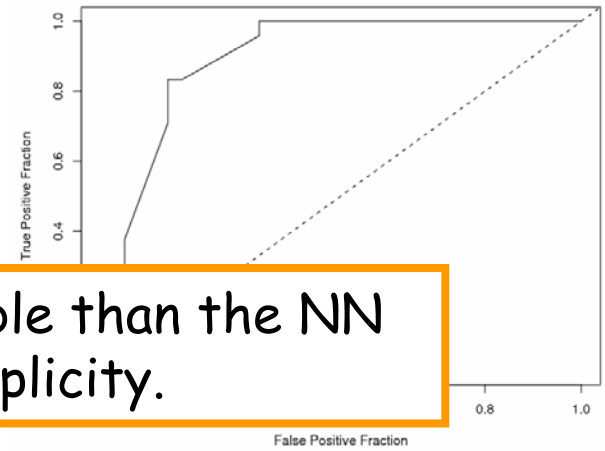
$$f(x) = 1/(1+\exp(-x))$$

as the activation function.

Comparison with NN

- Results obtained by NN given data for the haplotype 111
  - ROC curve based on the fitted value by NN is shown.
    - AUC = 0.8981
- Too complicated to explain the association between phenotypes and haplotype
  - Model
  - Over-f





QTLMARC is more suitable than the NN model with regard to simplicity.

Advantages of QTLMARC

- Several phenotypes are available for estimating the associated haplotype.
  - Introducing the notion of ROC curve and AUC
    - Alternative model for response variable with binomial?
- QTLMARC is suitable for analysis of genetic data
  - QTLMARC is developed based on medical science and genetics
    - It is easy to understand the results obtained from QTLMARC
      - Model is not complicated

Future scopes

- Haplotype estimation
  - Haplotype estimation by QTLMARC is not the originial haplotype estimation
    - In QTLMARC, diplotype of individual is not estimated
- Interaction of phenotypes
  - Multiplicative phenotypes is introduced(?)
- Non-parametric AUC
  - It is not assumed that phenotype is normally distributed.
  - AUC can be asymptotically calculated (Zhou, 2002)

References

- Iwasaki N, Horikawa Y, Tsuchiya T, Kitamura Y, Nakamura T, Tanizawa Y, Oka Y, Hara K, Kadowaki T, Awata T, Honda M, Yamashita K, Oda N, Yu L, Yamada N, Ogata M, Kamatani N, Iwamoto Y, Del Bosque-Plata L, Hayes M G, Cox N J, and Bell G I (2005). Genetic variants in the calpain-10 gene and the development of type 2 diabetes in the Japanese population. Journal of Human Genetics 2: 92--98.
- Chambers J M, Hastie T J (1992). Statistical Models in S, Wadsworth, CA, U.S.A.
- Copas J B, Corbett P (2002). Overestimation of the receiver operating characteristic curve for logistic regression. Biometrika 89 (2): 315--331.
- Cortes C, Mohri M (2004). AUC optimization vs. error rate minimization, Advances in Neural Information Processing Systems (NIPS 2003).
- Fisher R A (1918). The correlation between relatives on the supposition of Mendelian inheritance. Transaction of Royal Society of Edinburgh 52:399--433.
- Fisher R A (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics 7:179--188.
- IBM TJ Watson Research Center (2004). Model selection via the AUC. in Proceeding of the 21st International Conference on Machine Learning, Banff, Canada.

References

- Kamitsuji S. and Kamatani N. (2006) Estimation of haplotype associated with several quantitative phenotypic values based on maximization of area under ROC curve, Journal of Human Genetics, 15 Feb (Online First).
- Kitamura Y, Moriguchi M, Kaneko H, Morisaki H, Morisaki T, Toyama K, Kamatani N (2002). Determination of probability distribution of diplotype configuration (diplotype distribution) for each subject from genotypic data using the EM algorithm. Annuals of Human Genetics. 66: 183-93.
- McClish D K (1989). Analyzing a portion of the ROC curve. Medical Decision Making. 9: 190--195.
- Pepe M S (1997). A regression modelling framework for receiver operating characteristic curve in medical diagnostic testing. Biometrika. 84:595--608.
- Pepe M S (1998). Three approaches for regression analysis of receiver operating characteristic curves for continuous test results. Biometrics. 54:124--135.
- Pepe M S (2000). Interpretation, estimation and regression for ROC curves. Biometrics. 56:352--359.
- Qin J (2000). Using logistic regression procedures for estimating receiver operating characteristic curves. Biometrika. 90 (3): 585--596.

References

- Ransohoff D J, Feinstein A R (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. The New England Journal of Medicine. 299:926--930.
- Sebastiani P, Ramoni M F, Nolan V, Baldwin C T, Steinberg M H (2005). Genetic dissection and prognostic modeling of overt stroke in sickle cell anemia. Nature Genetics. 37 (4):435-400.
- Shibata K, Ito T, Kitamura Y, Iwasaki N, Tanaka H and Kamatani N (2004). Simultaneous estimation of haplotype frequencies and quantitative trait parameters: applications to the test of association between phenotype and diplotype configuration. Genetics. 168 (1): 525--539.
- Thomas A, Camp N J (2004). Graphical modeling of the joint distribution of alleles at associated loci. Am. J. Hum. Genet. 74: 1088--1101.
- Toesteson A A N, Begg C B (1988). A general regression methodology for ROC curve estimation. Medical Decision Making. 8:204--215.
- Yi N, George V, Allison D B (2003). Stochastic search variable selection for Identifying
- multiple quantitative trait loci. Genetics. 164:1129--1138.
- Zhou X H, Obuchowski N A, McClish D K (2002). Statistical Methods in Diagnostic Medicine. Wiley, New York, NY, U.S.A.