

Model selection criteria for logistic and log-linear regression

Małgorzata Żak-Szatkowska

Institute of Mathematics and Computer Science,
Wrocław University of Technology, Poland

Wojnowice, 20 November, 2008

Goal

Find relation between genes and trait, not continuous:

- healthy/sick,
- number of tumors.

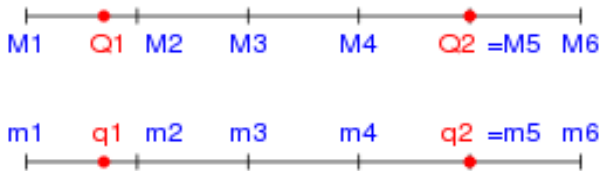
Goal

Find relation between genes and trait, not continuous:

- healthy/sick,
- number of tumors.

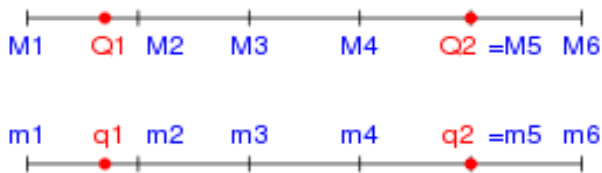
Predict probability of getting sick.

Goal



$$Q_1 q_1, Q_2 q_2 \implies Y$$

Goal



$$M_1 m_1, \dots, M_6 m_6 \implies Y$$

Notation

$Y = (Y_1, \dots, Y_n)^T$ - vector of observations.

Notation

$Y = (Y_1, \dots, Y_n)^T$ - vector of observations.

Consider a backcross population. There are only two possible genotypes: AA and Aa .

X_{ij} - variable describing the genotype of individual i at marker j .

$$X_{ij} = \begin{cases} \frac{1}{2}, & \text{if } i\text{-th individual is heterozygous at marker } j, \\ -\frac{1}{2}, & \text{if } i\text{-th individual is homozygous at marker } j. \end{cases}$$

Notation

$Y = (Y_1, \dots, Y_n)^T$ - vector of observations.

Consider a backcross population. There are only two possible genotypes: AA and Aa .

X_{ij} - variable describing the genotype of individual i at marker j .

$$X_{ij} = \begin{cases} \frac{1}{2}, & \text{if } i\text{-th individual is heterozygous at marker } j, \\ -\frac{1}{2}, & \text{if } i\text{-th individual is homozygous at marker } j. \end{cases}$$

$X_i = (1, X_{i1}, \dots, X_{iN_m})^T$ is vector of 1 and genotypes of N_m markers, for individual i , $i = 1, \dots, n$.

Logistic and log-linear regression

A linear predictor $X_i\beta$, where $\beta = (\beta_0, \beta_1, \dots, \beta_{N_m})$ - unknown parameters.

Logistic and log-linear regression

A linear predictor $X_i\beta$, where $\beta = (\beta_0, \beta_1, \dots, \beta_{N_m})$ - unknown parameters.

Let $Y_i \sim B(1, \pi_i)$ i.e. $P(Y_i = 1) = 1 - P(Y_i = 0) = \pi_i$.

$$\log \frac{\pi_i}{1 - \pi_i} = X_i\beta$$

Logistic and log-linear regression

A linear predictor $X_i\beta$, where $\beta = (\beta_0, \beta_1, \dots, \beta_{N_m})$ - unknown parameters.

Let $Y_i \sim B(1, \pi_i)$ i.e. $P(Y_i = 1) = 1 - P(Y_i = 0) = \pi_i$.

$$\log \frac{\pi_i}{1 - \pi_i} = X_i\beta$$

Let $Y_i \sim Poiss(\lambda_i)$.

$$\log \lambda_i = X_i\beta.$$

Logistic and log-linear regression

A linear predictor $X_i\beta$, where $\beta = (\beta_0, \beta_1, \dots, \beta_{N_m})$ - unknown parameters.

Let $Y_i \sim B(1, \pi_i)$ i.e. $P(Y_i = 1) = 1 - P(Y_i = 0) = \pi_i$.

$$\log \frac{\pi_i}{1 - \pi_i} = X_i\beta$$

Let $Y_i \sim \text{Poiss}(\lambda_i)$.

$$\log \lambda_i = X_i\beta.$$

General case (GLM):

$$g(E(Y_i)) = X_i\beta.$$

BIC

Goal: choosing the model that fits the data best.

BIC

Goal: choosing the model that fits the data best.

Let M_i be a model with specified k_i regressors and parameters $\beta_{(i)}$.

Let M_0 be a model without any effects, i.e. $\beta_{(0)} = (\beta_0, 0, \dots, 0)$.

BIC

Goal: choosing the model that fits the data best.

Let M_i be a model with specified k_i regressors and parameters $\beta_{(i)}$.

Let M_0 be a model without any effects, i.e. $\beta_{(0)} = (\beta_0, 0, \dots, 0)$.

$L(Y|M_i, \beta_{(i)})$ - likelihood function

$LRT_i = -2 \log \frac{L(Y|M_0, \beta_{(0)})}{L(Y|M_i, \beta_{(i)})}$ - likelihood ratio test statistic

BIC

Goal: choosing the model that fits the data best.

Let M_i be a model with specified k_i regressors and parameters $\beta_{(i)}$.

Let M_0 be a model without any effects, i.e. $\beta_{(0)} = (\beta_0, 0, \dots, 0)$.

$L(Y|M_i, \beta_{(i)})$ - likelihood function

$LRT_i = -2 \log \frac{L(Y|M_0, \beta_{(0)})}{L(Y|M_i, \beta_{(i)})}$ - likelihood ratio test statistic

Bayesian Information Criterion (Schwarz's BIC).

We choose the model that minimizes

$$BIC_i = -LRT_i + k_i \log(n).$$

mBIC

Problem of locating QTL :

- huge set of possible variables,
- moderate sample size,
- expected number of regressors is small.

mBIC

Problem of locating QTL :

- huge set of possible variables,
- moderate sample size,
- expected number of regressors is small.

BIC overestimates the predictors number. (Broman i Speed (2002)).

mBIC

Problem of locating QTL :

- huge set of possible variables,
- moderate sample size,
- expected number of regressors is small.

BIC overestimates the predictors number. (Broman i Speed (2002)).

Modification: additional penalty for the model dimension (Bogdan et al. (2004)).

We choose model that minimizes:

$$mBIC_i = -LRT_i + k \log(n) + 2k \log \left(\frac{N_m}{c_m} - 1 \right)$$

where c_m - expected number of regressors.

mBIC

Problem of locating QTL :

- huge set of possible variables,
- moderate sample size,
- expected number of regressors is small.

BIC overestimates the predictors number. (Broman i Speed (2002)).

Modification: additional penalty for the model dimension (Bogdan et al. (2004)).

We choose model that minimizes:

$$mBIC_i = -LRT_i + k \log(n) + 2k \log \left(\frac{N_m}{c_m} - 1 \right)$$

where c_m - expected number of regressors.

mBIC

Consider interactions between pairs of markers: $X_{iu}X_{iv}$, $i = 1, \dots, n$.

$N_i = \frac{N_m(N_m-1)}{2}$ - number of possible interactions.

mBIC

Consider interactions between pairs of markers: $X_{iu}X_{iv}$, $i = 1, \dots, n$.

$N_i = \frac{N_m(N_m-1)}{2}$ - number of possible interactions.

Let M_i be a model with specified k_i main effects and q_i interactions.

mBIC

Consider interactions between pairs of markers: $X_{iu}X_{iv}$, $i = 1, \dots, n$.

$N_i = \frac{N_m(N_m-1)}{2}$ - number of possible interactions.

Let M_i be a model with specified k_i main effects and q_i interactions.

We choose model that minimizes:

$$mBIC_i = -LRT_i + (k_i + q_i) \log(n) + 2k_i \log\left(\frac{N_m}{c_m} - 1\right) + 2q_i \log\left(\frac{N_{in}}{c_{in}} - 1\right)$$

where c_{in} - expected number of interactions.

mBIC

Consider interactions between pairs of markers: $X_{iu}X_{iv}$, $i = 1, \dots, n$.

$N_i = \frac{N_m(N_m-1)}{2}$ - number of possible interactions.

Let M_i be a model with specified k_i main effects and q_i interactions.

We choose model that minimizes:

$$mBIC_i = -LRT_i + (k_i + q_i) \log(n) + 2k_i \log\left(\frac{N_m}{c_m} - 1\right) + 2q_i \log\left(\frac{N_{in}}{c_{in}} - 1\right)$$

where c_{in} - expected number of interactions.

Standard $mBIC$ for linear regression:

$$c_m = c_{in} = 2.2,$$

controls the Type I error on the level of 8% when $n > 200$.

LRT and score statistic

LRT requires finding MLE for the parameters β .

LRT and score statistic

LRT requires finding MLE for the parameters β .

Numeric, time consuming methods

LRT and score statistic

LRT requires finding MLE for the parameters β .

Numeric, time consuming methods

For testing

$$H_0 : \forall i = 1, \dots, n \beta_i = 0,$$

LRT is asymptotically equivalent to Rao's score statistic, *SR*

Score statistic

$$\text{Let } U(\beta_{(k+q)}) = \left(\frac{\partial \log L(Y|M, \beta_{(k+q)})}{\partial \beta_0}, \frac{\partial \log L(Y|M, \beta_{(k+q)})}{\partial \beta_1}, \dots, \frac{\partial L(Y|M, \beta_{(k+q)})}{\partial \beta_{k+q}} \right).$$

Score statistic

$$\text{Let } U(\beta_{(k+q)}) = \left(\frac{\partial \log L(Y|M, \beta_{(k+q)})}{\partial \beta_0}, \frac{\partial \log L(Y|M, \beta_{(k+q)})}{\partial \beta_1}, \dots, \frac{\partial L(Y|M, \beta_{(k+q)})}{\partial \beta_{k+q}} \right).$$

Let $I(\beta_{(k+q)})$ be Fisher information matrix.

Score statistic

$$\text{Let } U(\beta_{(k+q)}) = \left(\frac{\partial \log L(Y|M, \beta_{(k+q)})}{\partial \beta_0}, \frac{\partial \log L(Y|M, \beta_{(k+q)})}{\partial \beta_1}, \dots, \frac{\partial L(Y|M, \beta_{(k+q)})}{\partial \beta_{k+q}} \right).$$

Let $I(\beta_{(k+q)})$ be Fisher information matrix.

Score test statistic

$$SR = [U(\hat{\beta}_0)]^T [I(\hat{\beta}_0)]^{-1} [U(\hat{\beta}_0)],$$

where $U(\beta)$ and $I(\beta)$ are evaluated at H_0 estimates: $\hat{\beta}_{(0)} = (\hat{\beta}_0, 0, \dots, 0)$.

Score statistic

$$\text{Let } U(\beta_{(k+q)}) = \left(\frac{\partial \log L(Y|M, \beta_{(k+q)})}{\partial \beta_0}, \frac{\partial \log L(Y|M, \beta_{(k+q)})}{\partial \beta_1}, \dots, \frac{\partial L(Y|M, \beta_{(k+q)})}{\partial \beta_{k+q}} \right).$$

Let $I(\beta_{(k+q)})$ be Fisher information matrix.

Score test statistic

$$SR = [U(\hat{\beta}_0)]^T [I(\hat{\beta}_0)]^{-1} [U(\hat{\beta}_0)],$$

where $U(\beta)$ and $I(\beta)$ are evaluated at H_0 estimates: $\hat{\beta}_{(0)} = (\hat{\beta}_0, 0, \dots, 0)$.

In matrix notation

$$SR = \frac{1}{\text{Var}_0(Y)} (Y - \bar{Y})^T X^T (X^T X)^{-1} X (Y - \bar{Y}),$$

where $\text{Var}_0(Y)$ is a variance of Y under H_0 and $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$.

SIC

Modification: replace *LRT* by *SR* (Inglot et al.(1997), Aerts et al. (2000), Bogdan et al. (2007)).

SIC

Modification: replace LRT by SR (Inglot et al.(1997), Aerts et al. (2000), Bogdan et al. (2007)).

Choose model that minimizes:

$$SIC_{k+q} = -SR_{k+q} + (k+q) \log(n)$$

or

$$SIC_{k+q} = -SR_{k+q} + (k+q) \log(n) + 2k \log\left(\frac{N_m}{c_m} - 1\right) + 2q \log\left(\frac{N_{in}}{c_{in}} - 1\right).$$

Score statistic and LRT for linear regression

Fit the linear model for the non-normal data:

$$Y_i = X_i\beta + \epsilon_i,$$

where ϵ_i is an environmental noise.

Score statistic and LRT for linear regression

Fit the linear model for the non-normal data:

$$Y_i = X_i\beta + \epsilon_i,$$

where ϵ_i is an environmental noise.

The *LRT* statistic is now

$$LLR = -n \log \left(\frac{RSS_{k+q}}{RSS_0} \right),$$

where RSS_{k+q} and RSS_0 are residual sum of squares from regression for M_{k+q} and M_0 .

Score statistic and LRT for linear regression

Fit the linear model for the non-normal data:

$$Y_i = X_i\beta + \epsilon_i,$$

where ϵ_i is an environmental noise.

The *LRT* statistic is now

$$LLR = -n \log \left(\frac{RSS_{k+q}}{RSS_0} \right),$$

where RSS_{k+q} and RSS_0 are residual sum of squares from regression for M_{k+q} and M_0 .

Let S^2 be a sample variance, $RSS_0 = nS^2$.

Score statistic and LRT for linear regression

It can be shown, that

$$LLR = -n \log \left(1 - \frac{\text{Var}_0(Y)}{nS^2} SR \right),$$

and as under H_0 we have $\frac{\text{Var}_0(Y)}{S^2} \rightarrow 1$

$$LLR \approx SR.$$

Score statistic and LRT for linear regression

It can be shown, that

$$LLR = -n \log \left(1 - \frac{\text{Var}_0(Y)}{nS^2} SR \right),$$

and as under H_0 we have $\frac{\text{Var}_0(Y)}{S^2} \rightarrow 1$

$$LLR \approx SR.$$

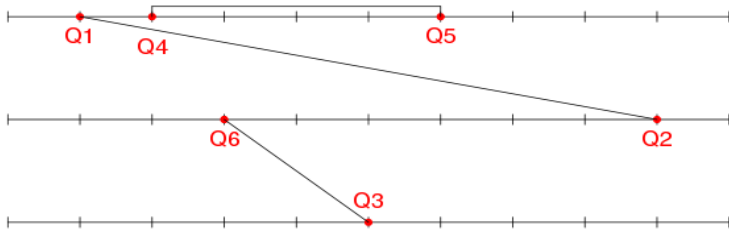
This proves that SIC and $mSIC$ are asymptotically equivalent to BIC and $mBIC$ calculated for the linear regression.

Scenario

- 3 chromosomes, 100cM each, markers every 10cM,
- $N_m = 33$, $N_i = 528$,
- $n=200$, $N = 1000$ replications,

Scenario

- 3 chromosomes, 100cM each, markers every 10cM,
- $N_m = 33$, $N_i = 528$,
- $n=200$, $N = 1000$ replications,
- genes localization:



Reported characteristics

Window used: $\pm 10cM$

Reported characteristics

Window used: $\pm 10\text{cM}$

Mean of 1000 replications:

- Power - proportion of correctly identified effects.
- $FDR = \frac{FP}{FP + \text{correct}}$ or 0.
- MR = missclassification rate.
- $MSE_{\beta} = \sum_{i=1}^{N_m + N_{in}} (\beta_i - \hat{\beta}_i)^2$.

Reported characteristics

Window used: $\pm 10cM$

Mean of 1000 replications:

- Power - proportion of correctly identified effects.
- $FDR = \frac{FP}{FP + \text{correct}}$ or 0.
- MR = missclassification rate.
- $MSE_{\beta} = \sum_{i=1}^{N_m + N_{in}} (\beta_i - \hat{\beta}_i)^2$.

Prediction on the testing set of additional 1000 observations.

- $d_{\pi} = \frac{1}{n} \sum_{i=1}^n |\pi_i - \hat{\pi}_i|$.
- $\text{correct}Y = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)$ (binomial only).

Binomial

Tablica: main effects: (1.8, 2, 2.2), interactions: (3.6, 4, 4.4)

	logistic reg. Score		linear reg. LLR		logistic reg. LRT	
	SIC	mSIC	BIC	mBIC	BIC	mBIC
Power	0.843	0.427	0.916	0.666	0.928	0.733
FDR	0.202	0.070	0.610	0.088	0.688	0.089
MR	2.306	3.607	10.675	2.383	14.588	2.018
MSE_{β}	75.367	49.238	430.377	47.221	433.323	46.893
d_p	0.125	0.194	0.176	0.151	0.154	0.134
correct Y	0.807	0.722	0.879	0.764	0.87	0.767

Percentage of correctly predicted Y on the training set: 0.894.

Poisson

Tablica: main effects: (0.6, 0.5, 0.4), interactions: (1.2, 1, 0.8)

	logistic reg. Score		linear reg. LLR		logistic reg. LRT	
	SIC	mSIC	BIC	mBIC	BIC	mBIC
Power	0.824	0.645	0.848	0.607	0.772	0.541
FDR	0.513	0.109	0.635	0.091	0.472	0.112
MR	6.838	2.65	11.80	2.766	5.938	3.188
MSE_{β}	6.65	3.452	9.785	3.335	12.694	4.046
d_p	0.428	0.382	0.483	0.397	0.450	0.433

References

1. Broman, K.W., Speed, T.P., 2002. A model selection approach for the identification of quantitative trait loci in experimental crosses. *J. Roy. Stat. Soc. B* 64, 641–656.
2. Bogdan, M., Ghosh, J.K., Doerge, R.W., 2004. Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting quantitative trait loci. *Genetics* 167, 989–999.
3. Inglot, T., Kallenberg, W.C.M., Ledwina, T., 1997. Data Driven Smooth Tests for Composite Hypotheses. *Annals of Statistics*, 25, No. 3, 1222–1250.
4. Aerts, M., Claeskens, G., Hart, J.D., 2000. Testing Lack of Fit in Multiple Regression. *Biometrika* 87, 405–424.