Construction of Disease Risk Scoring Systems using Logistic Group Lasso: Application to Porcine Reproductive and Respiratory Syndrome Survey Data

Hui Lin[a], Chong Wang[ab], Peng Liu[a] and Derald J. Holtkamp[ab]

[a]Department of Statistics, Iowa State University; bDepartment of Veterinary; [b]Diagnostic and Production Animal Medicine, Iowa State University

September 16, 2011

Background and Motivation

- Motivation: risk scoring system for PRRS
  - PRRS: Porcine Reproductive and Respiratory Syndrome
  - Major health, production and financial problem
- Aim: construct risk scoring systems for predicting diseases
- Typical approach: multivariate logistic regression + variable selection based on variable significance+risk scores are estimated coefficients
  - Low power for prediction !!

Our Work

In the study

- Propose to use the logistic group lasso algorithm to construct risk scoring systems for predicting diseases
- Apply to PRRS survey data
- Show it is superior to
  - Current scoring system based on expert opinion
  - Significance based system (logistic regression model)

Why Logistic Group Lasso?

logistic regression and Lasso

- Multivariate logistic regression:
  - Problem: quasi-complete-separation
  - Possible solution: add penalty
- Lasso: weighted l_1-norm penalty [Tibshirani 1996, Stat. Methodel.]
  - Advantage: stablize the estimation, also a variable selection tool
  - Problem: only selects individual dummy variables; the estimates are affected by the way dummy variables are encoded
  - Possible solution: add group indicator

Why Logistic Group Lasso?

Group Lasso

- Group Lasso : Yuan & Lin (2007, Journal of the Royal Statistical Society)
  - Penality: intermediate between the l_1- and l_2- type penalty
  - Variable selection on gorups instead of single variable
- Logistic Group Lasso: Meier et al. (2008, Journal of the Royal Statistical Society)

Why Logistic Group Lasso?

Quasi-Complete-Separation

$\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ binary response vector

$X = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n})^T$ design matrix in which each $\mathbf{x_i}$ is $p+1$ dimention column vector

$\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ parameter vector

The logliklihood function is as follows:

$$ln\mathcal{L}(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^{n} \left\{ y_i ln \frac{1}{1 + exp(-\mathbf{x_i}^T \boldsymbol{\beta})} + (1 - y_i) ln \left[1 - \frac{1}{1 + exp(-\mathbf{x_i}^T \boldsymbol{\beta})} \right] \right\}$$

$$D(\boldsymbol{\beta}) \equiv \frac{\partial ln\mathcal{L}(\boldsymbol{\beta}|\mathbf{y})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ y_i - \frac{1}{exp(-\mathbf{x_i}^T \boldsymbol{\beta})} \right\} \mathbf{x_i}$$

Why Logistic Group Lasso?

Quasi-Complete-Separation

- On the existence of maximum likelihood estimates in logistic regression models, A. Albert and J. A. Anderson
  - they first identified three possible mutually exclusive data patterns: i) overlap, ii) complete iii) quasi-Complete-separation

Logistic Group Lasso

Model set up

- $\mathbf{x_{i,g}}$ vector of dummy variables (i^{th} observation in group g) i = 1,...,n , g = 1,...,G
- y_i binary response for the i^{th} observation
- df_g degrees of freedom of group g

$$\mathcal{S}_\lambda(\beta) = -l(\beta) + \lambda \sum_{g=1}^{G} s(df_g) \parallel \beta_g \parallel_2$$

- $l(\beta)$ log-likelihood: $\sum_{i=1}^{n}\{y_i \eta_\beta(\mathbf{x_i}) - log[1 + exp(\eta_\beta(\mathbf{x_i}))]\}$,
- λ tuning parameter for penalty and $s(.)$ is $s(df_g) = df_g^{0.5}$

Choose tuning parameter

Leave one out cross validation

- The optimal value of λ is determined through leave-one-out cross validation
- Grid of 148 values $\{0.96\lambda_{max}, 0.96^2\lambda_{max}, ..., 0.96^{148}\lambda\}$ [2008, Journal of the Royal Statistical Society]
- Here

$$\lambda_{max} = \max_{g \in \{1,...,G\}} \{\frac{1}{s(df_g)} \parallel \mathbf{x_g^T}(\mathbf{y} - \bar{\mathbf{y}}) \parallel_2\}$$

Model Evaluation

Three criteria——Receiver Operating Characteristic analysis

- ROC curve: (False Positive Rate, True Positive Rate) as cutoff value varies
- If we use binary variable, D, to denote true outbreak status:

$$D = \begin{cases} 1 & outbreak \\ 0 & non-outbreak \end{cases}$$

The variable T is the result of the diagnostic test.

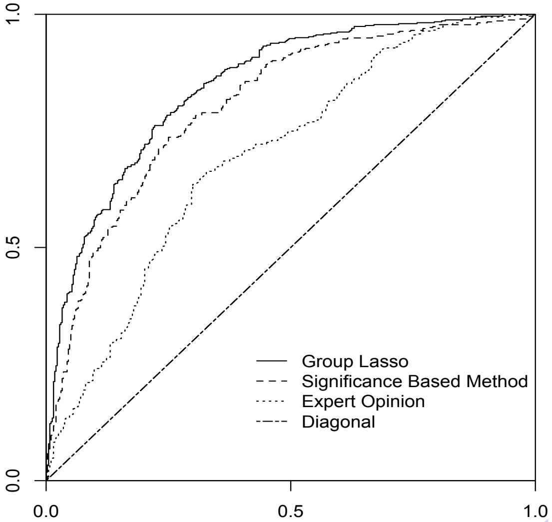
$$T = \begin{cases} 1 & test\ positive \\ 0 & test\ negative \end{cases}$$

$$1-Specificity = false\ positive\ fraction = FPF = P[T = 1|D = 0]$$

$$Sensitivity = true\ positive\ fraction = RPF = P[T = 1|D = 1]$$

Model Evaluation

Three criteria——Receiver Operating Characteristic analysis



Model Evaluation

Nonparametric Comparison (U-statistics)

Definition

Let $\mathbf{x_1}, ..., \mathbf{x_n}$ be a sample of n vectors with $\mathbf{x_\alpha} = (x_\alpha^{(1)}, ..., x_\alpha^{(r)})$, $\alpha = 1, ..., n$ and $\Phi(\mathbf{x_1}, ..., \mathbf{x_m})$ a function of $m(\leq n)$ vector arguments. Define

$$U(\mathbf{x_1}, ..., \mathbf{x_n}) = \frac{1}{n(n-1)...(n-m+1)} \sum{}'' \Phi(\mathbf{x_{\alpha_1}}, ..., \mathbf{x_{\alpha_m}})$$

where $\sum''$ stands for summation over all permutations $(\alpha_1, ..., \alpha_m)$ of m integers such that

$$1 \leq \alpha_i \leq n, \quad \alpha_i \neq \alpha_j \text{ } if \text{ } i \neq j, \quad (i, j = 1, ..., m)$$

U is the average of the values of Φ in the set of ordered subsets of m members of the sample. U is symmetric in $(x_1, ..., x_n)$. Any statistic of the form will be called a U-statistics. Function $\Phi(x_1, x_2, ..., x_m)$ is kernel of the statistics U.

Model Evaluation

Three criteria——AUC

- Assume sample of N individuals undergo a test
- C_1 —— positive group, size **m**
- C_2 —— negative group, size **N-m=n**
- X_i—— individuals in C_1, $i = 1, ..., m$
- Y_j—— individuals in C_2, $j = 1, ..., n$
- For a cut-off value $z \in \mathcal{R}$
  - $sensitivity(z) = \frac{1}{m}\sum_{i=1}^{m} I(X_i \geq z)$
  - $specificity(z) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j < z)$

Model Evaluation

More about AUC (Cont)

Assumptions:

1. All the observations from both groups are independent of each other
2. The distributions of both groups are equal

Definition

For $n \times m$ array (X_i, Y_j), Mann-Whitney test statistic U is defined as the number of (X_i, Y_j) pairs where $X_i > Y_j$.

Fact

AUC=P(sample from positive group>sample from negative group) (P78, Result 4.6 Pepe2003)

Model Evaluation

More about AUC (Cont)

$$\begin{bmatrix} (X_1, Y_1) & (X_1, Y_2) & ...... & (X_1, Y_n) \\ (X_2, Y_1) & (X_2, Y_2) & ...... & (X_2, Y_n) \\ ... & ... & ...... & ... \\ (X_m, Y_1) & (X_m, Y_2) & ...... & (X_m, Y_n) \end{bmatrix}_{m \times n}$$

Count the propotion that $(X_i > Y_j)$ ———> estimated AUC

Model Evaluation

Nonparametric Comparison – outline of the background theory

1. Construct a U statistics to estimate AUC
2. Apply a Hoeffding (genius 1) 's result to estimate the variance of U statistics (only for one)
3. Extend to a **vector** U-statistics (i.e. estimate the variance-covariance matrix) (not easy)
4. Sen (1960, genius 2) has provided consistent estimates of the elements of the variance-covariance matrix of a vector of U-statistics
5. Comparison:$g(\boldsymbol{\theta}) = \theta_1 - \theta_2$, proved that g (under some conditions) is asymptotically normally distributed

Model Evaluation

Nonparametric Comparison (U-statistics)

- Notate the AUC we are going to estimate as θ. It can be computed as the average over a kernel, ψ , as

$$\hat{\theta} = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \psi(X_i, Y_j)$$

$$\psi(X, Y) = I(Y < X) + 0.5 I(Y = X)$$

  Note: In terms of probabilities, $E(\hat{\theta}) = \theta = Pr(Y < X) + 0.5 Pr(X = Y)$. For continuous distributions, $Pr(Y = X) = 0$.

- DeLong et al. (1988) presented a nonparametric approach to compare AUC based on generalized U-statistics to generate an estimated covariance matrix.

Model Evaluation

Model comparison (nonparametric approach to compare AUC)

- Asymptotic normality and an expression for the variance can be derived from generalized U-statistics by Hoeffding (1948)(Section 5, 5.18).

Definitions

$$\xi_{10} = E[\psi(X_i, Y_j)\psi(X_i, Y_k)] - \theta^2, \quad j \neq k;$$
$$\xi_{01} = E[\psi(X_i, Y_j)\psi(X_k, Y_j)] - \theta^2, \quad i \neq k;$$

$$\xi_{11} = E[\psi(X_i, Y_j)\psi(X_i, Y_j)] - \theta^2$$

Then

$$var(\hat{\theta}) = \frac{(n-1)\xi_{10} + (m-1)\xi_{01}}{mn} + \frac{\xi_{11}}{mn}$$

Model Evaluation

Model comparison (nonparametric approach to compare AUC)

Extend Hoeffding's theory to a vector U-statistics. Let $\hat{\boldsymbol{\theta}} = (\hat{\theta}^1, \hat{\theta}^2, ..., \hat{\theta}^k)$ be a vector of statistics, representing the AUC's from the corresponding $\{X_i^r\}, \{Y_j^r\}$ $(i = 1, ..., m;\ j = 1, ..., n;\ \ 1 \le r \le k)$ of k different diagnostic measures. (Section 6)

$$\xi_{10}^{rs} = E[\psi(X_i^r, Y_j^r)\psi(X_i^s, Y_k^s)] - \theta^r\theta^s, \quad j \neq k;$$
$$\xi_{01}^{rs} = E[\psi(X_i^r, Y_j^r)\psi(X_k^s, Y_j^s)] - \theta^r\theta^s, \quad i \neq k;$$

$$\xi_{11}^{rs} = E[\psi(X_i^r, Y_j^r)\psi(X_i^s, Y_j^s)] - \theta^r\theta^s$$

Then

$$cov(\hat{\theta}^r, \hat{\theta}^s) = \frac{(n-1)\xi_{10}^{rs} + (m-1)\xi_{01}^{rs}}{mn} + \frac{\xi_{11}^{rs}}{mn}$$

Model Evaluation

Model comparison (nonparametric approach to compare AUC)

Sen (1960) has provided consistent estimates of the elements of the variance-covariance matrix of a vector of U-statistics.

$$V_{10}^r(X_i) = \frac{1}{n}\sum_{j=1}^{n} \psi(X_i^r, Y_j^r) \quad (i = 1, 2, ..., m)$$

$$V_{01}^r(Y_j) = \frac{1}{m}\sum_{i=1}^{m} \psi(X_i^r, Y_j^r) \quad (j = 1, 2, ..., n)$$

Also define the $k \times k$ matrix $\boldsymbol{S_{10}}$, which has $(r, s)^{th}$ element

$$s_{10}^{r,s} = \frac{1}{m-1}\sum_{j=1}^{n} [V_{10}^r(X_i) - \hat{\theta}^r][V_{10}^s(X_i) - \hat{\theta}^s]$$

Model Evaluation

Model comparison (nonparametric approach to compare AUC)

and similarly $\boldsymbol{S_{01}}$, which has $(r,s)^{th}$ element

$$s_{01}^{r,s} = \frac{1}{n-1}\sum_{j=1}^{n}[V_{01}^{r}(Y_j) - \hat{\theta}^r][V_{01}^{s}(Y_j) - \hat{\theta}^s]$$

The estimated covariance matrix for the vector of parameter estimates, $\hat{\boldsymbol{\theta}} = (\hat{\theta}^1, \hat{\theta}^2, ..., \hat{\theta}^k)$ is thus

$$\boldsymbol{S} = \frac{1}{m}\boldsymbol{S_{10}} + \frac{1}{n}\boldsymbol{S_{01}}$$

- Let g be a real-value function of $\hat{\boldsymbol{\theta}}$ that has **bounded second derivatives in a neihborhood of $\boldsymbol{\theta}$** .

Model Evaluation

Model comparison (nonparametric approach to compare AUC)

- Combining results from Sen(1960) and Arveson (1969, Theorem 16), it follows that if $lim_{N\to\infty}\frac{m}{n}$ is **bounded and nonzero**, then $N^{\frac{1}{2}}[g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})]$ is asympotically normally distributed with mean 0 and variance σ_g^2, where

$$\sigma_g^2 = lim_{N\to\infty} \sum_{j=1}^{k}\sum_{i=1}^{k} \frac{\partial g}{\partial \theta^i}\frac{\partial g}{\partial \theta^j}(\frac{1}{m}\xi_{10}^{i,j} + \frac{1}{n}\xi_{01}^{i,j})$$

$$s_g^2 = N \sum_{j=1}^{k}\sum_{i=1}^{k} \frac{\partial g}{\partial \theta^i}\frac{\partial g}{\partial \theta^j}(\frac{1}{m}s_{10}^{i,j} + \frac{1}{n}s_{01}^{i,j})$$

is a consistent estimate of σ_g^2.

Model Evaluation

Model comparison (nonparametric approach to compare AUC)

When g is simply a linear function, the partial derivatives are the constants that comprise the linear function.
For any contrast $\boldsymbol{L\theta'}$:

$$\frac{\boldsymbol{L\hat{\theta}'} - \boldsymbol{L\theta'}}{[\boldsymbol{L}(\frac{1}{m}\boldsymbol{S_{10}} + \frac{1}{n}\boldsymbol{S_{01}})\boldsymbol{L'}]^{\frac{1}{2}}} \sim N(0,1)$$

The test can also take the form of chi-square distribution:

$$(\hat{\theta} - \theta)\boldsymbol{L'}[\boldsymbol{L}(\frac{1}{m}\boldsymbol{S_{10}} + \frac{1}{n}\boldsymbol{S_{01}})\boldsymbol{L'}]^{-1}\boldsymbol{L(\hat{\theta} - \theta)'} \sim \chi^2_{rank(\boldsymbol{LSL'})}$$

Model Evaluation

Three Criteria

- Log-likelihood $l(\hat{\beta})$: $\sum_{i=1}^{n}\{y_i\eta_{\hat{\beta}}(\mathbf{x_i}) - log[1 + exp(\eta_{\hat{\beta}}(\mathbf{x_i}))]\}$
- Maximum correlation coefficient [Yeo and Burge 2004, J. Computnl Biol]

$$\rho_{max} = max\{\rho_\tau | \tau \in (0,1)\}$$

  - $\tau \in (0,1)$ threshold to classify predicted probability into binary disease status
  - ρ_τ Pearson correlation coefficient between the true binary disease status and the preditive disease status with threshold τ.

Application to PRRS Data

- American Association of Swine Veterinarians (AASV) Production Animal Disease Risk Assessment Program (PADRAP)
- Surveys completed between March 2005 and March 2009
- Responses obtained from the most recently completed survey for each site
  - **Explanatory variables:** 127 questions
  - **Response variable:** whether a breeding herd site reported a clinical PRRS outbreak in the past 3 years
  - **Number of farms:** 896 (499, 56% positive)

Application to PRRS Data

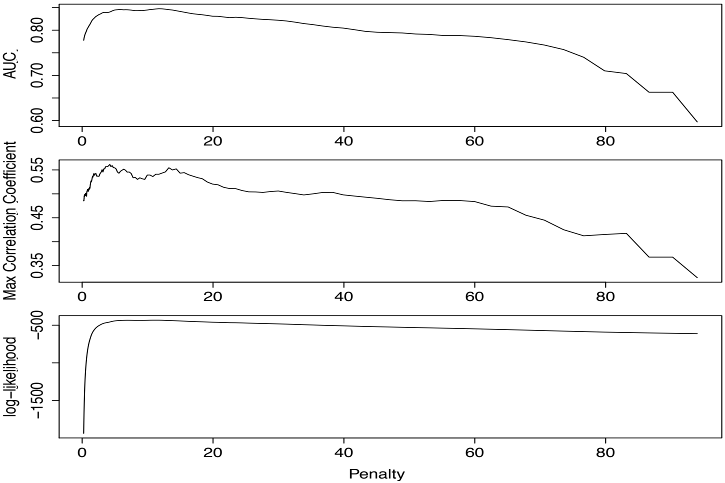
Three Criteria

- Leave-one-out cross-validation: one of the 896 farms is excluded and the other 895 farms are used as a training data set
  - AUC
  - Log- likelihood
  - Maximum correlation coefficient

Application to PRRS Data

Results for three criteria

- The optimal values of λ are : 11.72, 4.22 and 11.72



Application to PRRS Data

Distribution of estimated probabilities

Chosen $\lambda = 11.72$

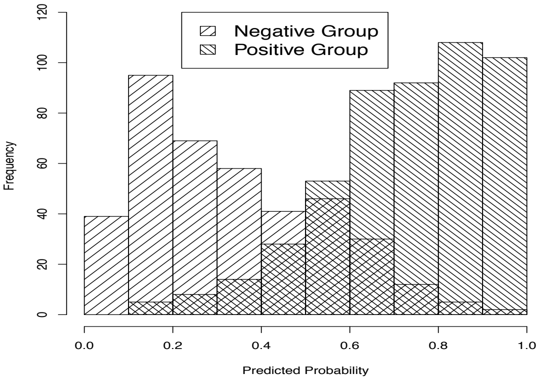


Figure: Distributions of estimated probabilities for both negative and positive groups

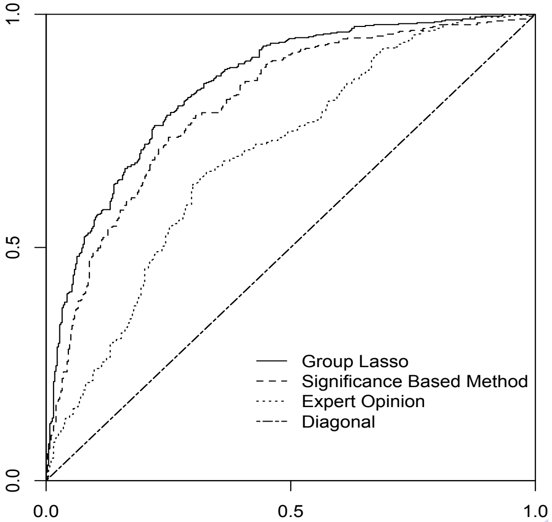
Comparison among risk scoring systems

Three Systems

- Plug in the chosen λ then apply logistic group lasso to PRRS data
- Compare with two other systems:
    1. The current system based on expert opinion
    2. Significance based system

Comparison among risk scoring systems

ROC Curves



Comparison among risk scoring systems

AUC Comparison

Table: AUC estimations for three risk scoring systmes

Model Names	AUC	95% CI
Group Lasso	0.848	(0.822, 0.873)
Significance Based Method	0.807	(0.773, 0.841)
Expert Opinion	0.696	(0.661, 0.731)

- These AUCs are compared by using the nonparametric approach of DeLong (1988, Biometrics)

Discussion

- What we have done?
  - Introduce the logistic group lasso algorithm for development of risk scoring systems for diseases.
  - Choose tuning parameter: leave-one-out cross validation with criterion of AUC
  - Apply to PRRS data
    - Our system is better than the other two systems
    - 74 of the 127 questions analyzed are excluded

Discussion

- Set scores to explanatory variables
- Identify questions that could be removed without affecting predictive power
- Demonstrate how a program can be used Decrease the reliance on expert opinion

Simulation Study

```
> count1
 [1] 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[75] 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0
> count2
 [1] 7 2 2 3 4 0 2 4 2 2 2 0 3 0 6 3 2 0 0 2 4 4 0 1 5 2 2 1 6 6 3 1 0 0 2 0 3 2 1 0 2 0
2 2 3 0 0 1 3 1 3 9 2 0 2 3 4 1 1 0 1 0 2 1 1 0 0 4 4 3 3 0 8 7
[75] 1 1 3 0 0 0 4 0 2 0 2 4 0 0 4 0 1 1 6
> count3
 [1] 18  8 19 28 31 17 22 25 10 32 23  6 19 11 42 14 23 21 20 11 42 15 30  7 32  6 17 20
38 32 23 24 30  6 34 13 33 24  8 18  5 17 16 18  8  1 18 18 29
[50] 22  6 32  2 15 23  8  9  1 20 14 25 11 17 16 11  2  8 32 25 26 22 12 40 27 22 12 14
15 28 27 11  9 16 13 27 31 12 28 11  3 30 18 28
> summary(count1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.00000 0.03226 0.00000 1.00000
> summary(count2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   2.000   2.054   3.000   9.000
> summary(count3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   11.00   18.00   18.96   27.00   42.00
```

Thank you!