Selecting variables in multivariate linear models with Subselect

A. PEDRO DUARTE SILVA[(1)(*)]
JORGE CADIMA[(2)]
MANUEL MINHOTO[(3)]
JORGE ORESTES CERDEIRA[(2)]

(1) FEG/CEGE – UNIV. CATÓLICA PORTUGUESA – C.R. PORTO
(2) I.S. AGRONOMIA – UNIV. TÉCNICA DE LISBOA
(3) DEP. MATEMÁTICA – UNIVERSIDADE DE ÉVORA

(*) Supported by: FEDER / POCI 2010

Ciência.Inovação 2010 Programa Operacional Ciência e Inovação 2010
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA E DA TECNOLOGIA

The *Subselect* Package

OVERVIEW

1. GOALS AND ISSUES OF VARIABLE SELECTION
2. VARIABLE SELECTION IN LINEAR REGRESSION
3. VARIABLE SELECTION IN LINEAR MODELS WITH MULTIPLE RESPONSES
4. VARIABLE SELECTION IN GENERALIZED LINEAR MODELS
5. The *Subselect* PACKAGE
6. FINAL REMARKS

The *Subselect* Package

GOALS OF VARIABLE SELECTION

- PARCIMONY / INTERPETABILITY
- BETTER PREDICTIONS
- EFFICIENT ESTIMATION / INFERENCE

ISSUES IN VARIABLE SELECTION

- CHOICE OF SELECTION CRITERIA
- EFFICIENCY / EFECTIVNESS OF SELECTION ALGORITHMS
- EFFECTS OF SELECTION BIAS

The *Subselect* Package

VARIABLE SELECTION IN LINEAR REGRESSION

$y = X b + e \quad RSS = e^T e \qquad R^2 = 1 - \frac{RSS}{(n-1)s_y^2}$

- RSS (OR R^2) - BASED SELECTION CRITERIA :

ADJUSTED $R^2 \rightarrow$ $R_a^2 = 1 - \frac{RSS}{(n-k-1)s_y^2}$ (DESCRIP.INDEX)

MALLOWS $C_P \rightarrow$ $C_k = \frac{RSS}{\hat{\sigma}^2} - (n - 2(k+1))$ (PRED.CRITERION)

AKAIKE AND SCHWARZ $\rightarrow$

$$AIC = n \ln\left(\frac{2\pi RSS}{n-k-1}\right) + n + 2k$$

$$BIC = -\frac{1}{2}\left(n \ln\left(\frac{2\pi RSS}{n-k-1}\right) + n + k \ln(n)\right)$$

(INFORMATION CRITERIA)

The *Subselect* Package

VARIABLE SELECTION IN LINEAR REGRESSION

- PROBLEMS OF SELECTION BIAS:
  1. PARAMETER ESTIMATES MAY BE OVERBLOWN
  2. CLASSICAL INFERENCE METHODS ARE NOT VALID
  3. GOODNESS OF FIT MEASURES ARE TOO OPTIMISTIC
- REMEDIES
  1. SHRINKAGE METHODS
  2. MULTIPLE-TESTS INFERENCE
  3. RESAMPLING AND/OR CROSS-VALIDATION

The *Subselect* Package

VARIABLE SELECTION IN LINEAR REGRESSION

OTHER APPROACHES :

- BAYES FACTORS AND BAYES AVERAGING
- GARROTE, LASSO, ELASTIC NET AND RELATIVES
- GENERAL TO SPECIFIC MODELLING

The R *Subselect* Package

LINEAR MODELS WITH MULTIPLE RESPONSES

$Y = X B + U$ $\qquad r = \min(\dim(X),\dim(Y))$

$T = S_{XX}$ $\quad H = S_{XX} - S_{XY} S_{YY}^{-1} S_{YX}$ $\qquad ccr_i^2 = \text{Eigval}_i(T^{-1}H)$

Comparison Criteria:

$$ccr_1^2$$

(*max* ccr_1^2 $\Leftrightarrow$ *max* Roy λ_1)

$$\tau^2 = 1 - \left(\prod_{i=1}^{r} (1 - ccr_i^2) \right)^{1/r}$$

(*max* τ^2 $\Leftrightarrow$ *min* Wilks Λ)

Multivariate Indices

$$\varsigma^2 = 1 - \frac{r}{\sum_{i=1}^{r} \left(1 - ccr_i^2\right)^{-1}}$$

(*max* ς^2 $\Leftrightarrow$ *max* Lawley-Hotelling trace)

$$\xi^2 = \frac{\sum_{i=1}^{r} ccr_i^2}{r}$$

(*max* ξ^2 $\Leftrightarrow$ *max* Bartlett-Pillai trace)

The *Subselect* Package

A LINEAR HYPOTHESIS FRAMEWORK:

$X = A\,\Psi + U \qquad H0:\ C\,\Psi = 0$

→ SELECT COLUMNS OF X IN ORDER TO EXPLAIN H1

PARTICULAR CASES:

- LINEAR DISCRIMINANT ANALYSIS

$A = [1_g] \qquad \Psi = [\mu_g] \qquad C = \begin{bmatrix} 1 & -1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 1 & 0 & \dots & -1 \end{bmatrix}$

- MULTI-WAY MANOVA/MANCOVA EFFECTS

$\Omega = \mathcal{R}(A) \qquad \omega = \mathcal{R}(A) \cap \mathcal{N}(C\,(A^TA)^-\,A^T) \qquad r = \dim(\Omega) - \dim(\omega)$

$ccr_i^2 = Eigval_i(T^{-1}H) \qquad T = X'\,(I - P_\omega)\,X \qquad H = X'\,(P_\Omega - P_\omega)\,X$

The *Subselect* Package

GENERALIZED LINEAR MODELS

$$g(y) = X B + u = [X_1\ X_2]\,[B_1^T\ B_2^T]^T + u$$

Comparison Criteria:

$$\lambda = 2\left[\ln L(\hat{B}) - \ln L(\hat{B}_{1R}, 0)\right]$$ **(Likelihood ratio statistic)**

$$\lambda^* = \hat{B}_2^T (\hat{I}_{22} - \hat{I}_{21}\hat{I}_{11}^{-1}\hat{I}_{12})\hat{B}_2 \qquad \hat{I} = -\left.\frac{\partial^2 \ln L(B)}{\partial(B)\partial(B^T)}\right|_{B=\hat{B}} = \begin{bmatrix} \hat{I}_{11} & \hat{I}_{12} \\ \hat{I}_{21} & \hat{I}_{22} \end{bmatrix}$$ **(Wald statistic)**

$$\lambda^{**} = 2\left[\ln L(\hat{B}) - \ln L(\tilde{B}_1, 0)\right] \qquad \tilde{B}_1 = \hat{B}_1 + \hat{I}_{11}^{-1}\hat{I}_{12}\hat{B}_2$$

AIC $= \lambda^{\#} + 2(k + m(k))$ $\qquad \lambda^{\#} = \lambda$, λ^* OR λ^{**}

The *Subselect* Package

The Subselect Package

Search routines for (combinatorial) criteria optimization

Exact Algorithm:

leaps - based on Furnival and Wilson´s leaps and bounds algorithm for linear regression

- viable with up to 30 - 35 original variables

Heuristics:

anneal - simulated annealing

genetic - genetic algorithm

improve - restricted local improvement

The *Subselect* Package

Comparison criteria → **functions of T^{-1} H** **eigenvalues**
T, H symmetric **rank(H) = *r***

Linear Regression **(r=1)** → ***max* R^2**

Linear Models with Multiple Responses → ***max***
"ccr12", "tau2", "xi2" or "zeta2"

Generalized Linear Models **(r=1)** → ***min***
Wald statistic (λ^*)

The *Subselect* Package

Search functions arguments :

- **Matrices T and H**
- **Comparison criterion (models with multiple responses)**
- **Tuning parameters for heuristics**
- **Maximum time allowed for exact search**
- **Variables forcibly included or excluded in the selected subsets**
- **Number of solutions by subset dimensionality**
- **Numerical tolerance for detecting singular or non-symmetrical matrices**

The *Subselect* Package

Auxiliary functions:

lmHmat - **creates H and T matrices for linear regression/canonical correlation analysis**

ldaHmat - **creates H and T matrices for linear discriminant analysis**

glhHmat - **creates H and T matrices for an analysis based on a linear hypothesis specified by the user**

The *Subselect* Package

Expanding a Subselect Analysis

Other issues in variable selection can be tackled by combining subselect with the wider system:

- **Searches across different dimensionalities can be handled by post-processing the results of subselect**

- **Selection biases can be estimated by resampling or cross-validation procedures that include the basic searches within larger cycles**

- **Relative importance of individual variables can be accessed by collecting statistics on groups of "good" subsets of different dimensionalities**

The *Subselect* Package

Final Remarks

- Subselect implements effective searches for variable subsets in a wide range of models

- Exact searches are viable if the number of original variables is not much larger than 30

- Random searches are usually good and much better than greedy ("stepwise") algorithms

- Many subsets lead to similar criterion values. Finding the exact optimum may not be so important

- Subselect is most useful when used in combination with other utilities of the system

References

Cadima J, Cerdeira JO and Minhoto M (2004). Computational Aspects of Algorithms for Variable Selection in the Context of Principal Components. *Computational Statistics and Data Analysis* **47**: 225-236.

Duarte Silva, A.P. (2001). Efficient Variable Screening for Multivariate Analysis. *Journal of Multivariate Analysis* **76**, 35-62.

Furnival, G.M. & Wilson, R.W. (1974). Regressions by Leaps and Bounds. *Technometrics* **16**: 499-511.

Lawless, J.F and Singhal, K. (1978). Efficient Screening of NonNormal Regression Models. *Biometrics. **34:** 318-327.*

Lawless, J.F and Singhal, K. (1987). ISMOD: An All-Subsets Regression Program for Generalized Linear Models. I. Statistical and Computational Background. *Computing Methods and Programs in Biomedecine* **24:** 125-134