Journal of Data Science **11**(2013), 371-387

Variable Selection in the *Chlamydia Pneumoniae* Lung Infection Study

Yuan Kang and Nedret Billor*
Auburn University

Abstract: In this study, the data based on nucleic acid amplification techniques (Polymerase chain reaction) consisting of 23 different transcript variables which are involved to investigate genetic mechanism regulating chlamydial infection disease by measuring two different outcomes of muring *C. pneumonia* lung infection (disease expressed as lung weight increase and *C. pneumonia* load in the lung), have been analyzed. A model with fewer reduced transcript variables of interests at early infection stage has been obtained by using some of the traditional (stepwise regression, partial least squares regression (PLS)) and modern variable selection methods (least absolute shrinkage and selection operator (LASSO), forward stagewise regression and least angle regression (LARS)). Through these variable selection methods, the variables of interest are selected to investigate the genetic mechanisms that determine the outcomes of chlamydial lung infection. The transcript variables Tim3, GATA3, Lacf, Arg2 (X_4, X_5, X_8 and X_{13}) are being detected as the main variables of interest to study the *C. pneumonia* disease (lung weight increase) or *C. pneumonia* lung load outcomes. Models including these key variables may provide possible answers to the problem of molecular mechanisms of chlamydial pathogenesis.

Key words: LASSO, multicollinearity, partial least squares regression, stepwise regression, variable selection.

1. Introduction

Chlamydophila pneumonia (*C. pneumonia*), an obligate intracellular bacterial pathogen, is the most common chlamydial pathogen that causes community-acquired respiratory infections. Although the infection is typically mild acute, it is strongly associated with atherosclerotic coronary heart diseases. *C. pneumonia* infection may also lead to chronic, persistent infections with several possible disease outcomes such as chronic inflammatory diseases of presumably noninfectious etiology (Cannon *et al.*, 2005; Wang, 2005; Frikha-Gargouri *et al.*, 2008).

*Corresponding author.

The complex interaction between chlamydial replication and host response determines the outcome of the infection not by a single all-influencing factor but rather by a series of accumulated host- and pathogen-associated factors. Also, human studies indicated a major influence of host genetics on disease severity following chlamydial infection (Wang *et al.*, 2008).

Since *C. pneumonia* can lead to severe clinical disease, correct diagnosis and therapy are important issues. However, conventional assays for the detection of *C. pneumonia* have limitations, and there is a need for more accurate, convenient and rapid diagnostic methods. Nucleic acid amplification techniques (such as polymerase chain reaction) have such a potential to offer clinical laboratories a convenient means to detect *C. pneumonia* and ensure optimal, timely and appropriate clinical decisions and patient care (Boman *et al.*, 1999). In this study, the researchers choose 23 different transcript variables measured at a molecular level based on the modified polymerase chain reaction technique. These transcript variables are the key markers of the immune and inflammatory response and play pivotal roles in the regulation of the immune response to chlamydial infection or play a key role in a protective host response to chlamydia infection (Wang *et al.*, 2008). Usually, the evaluations of protection from this disease are determined by survival of mice and lung weight increase, as well as elimination of *C. pneumonia* organisms by determination of total chlamydial lung loads. The lung weight increase is a reliable measure of disease intensity, and high increases reflect severe disease (Li *et al.*, 2010). Therefore, in our study these two indexes, percent lung weight increase (based on naïve lung weights of 138.4 mg for adult female A/J mice) and the logarithm of total *C. pneumonia* lung loads are treated as the dependent variables which are the observed results of these 23 transcript variables being manipulated.

In this study, disease intensity and chlamydial lung loads were determined early after rechallenge, i.e., on day 3 and later on day 10 at peak disease, and then mice were sacrificed by CO_2 inhalation 3 days or 10 days after inoculation, and lungs were weighed, snap frozen in liquid nitrogen, and stored at -80 °C until further processing to get the value of the 23 parameters. Therefore, the whole experiment consisted of 16 different groups of each two time points (day 3 and day 10), comprising a total of 320 female mice, i.e., each group contains 10 female mice. The dependent variables are Y_1 lung disease (lung weight increase %) and Y_2 lung *C. pneumonia* load (i.e., chlamydial lung burden), defined as average percentage of lung weight increase and the $\log_{10}$ *C. pneumonia* / lung, respectively. The independent variables are the 23 different transcript variables, which were $\log_2$ transformed.

Variable selection in multivariate analysis is a very relevant step, because the removal of non-informative variables will produce better predicting and sim-

pler models and at the same time maintain almost essentially all the information provided by the original set of regressors. The technique can be used to gain a better understanding of the regression relationship through a simplified description of it, or to reduce the number of regressors required for effective predication of the regressand. In either case, further studies involving such regression relationship will be easier and less expensive to carry out if fewer variables are involved. Goals in variable selection methods include: accurate predictions, interpret models−determining which predictors are meaningful, stability−small changes in the data should not result in large changes in the subset of predictors used, the associated coefficients, or the predictions, and avoiding bias in hypothesis tests during or after variable selection. Traditional methods, such as stepwise regression, all-subsets regression, ridge regression, principal component and partial least squares based methods fall short in one or more of these criteria. Modern procedures such as boosting (Freund and Schapire, 1997) forward stagewise regression (Hastie *et al.*, 2007), and LASSO, least absolute shrinkage and selection operator (Tibshirani, 1996), LARS, least angle regression (Efron *et al.*, 2004) improve stability and predictions.

The objectives of this study are to select the multivariate model candidates based on a few well-known selection methods and criteria, and to construct a simple model which can efficiently predict the late disease outcomes with high accuracy.

Descriptions of the variable selection methods employed in this study are given in Section 2. Results of the analysis and conclusion are provided in Sections 3 and 4.

2. Variable Selection Methods

2.1 Stepwise Regression Based Variable Selection Methods

One common approach to select a subset of variables from a complex model is stepwise regression. A stepwise regression is a procedure to examine the impacts of each variable on the outcomes step by step. The variables that do not contribute much to the variance explained would be removed. There are several versions of stepwise regression such as forward selection, backward elimination, and stepwise (Al-Subaihi, 2002).

The stepwise regression has its own limitations. When the number of variables is large compared to the number of observations in the data set or a multicollinearity problem is detected among the variables, the stepwise algorithm may not function or end up with throwing nearly all the variables into or out from the model, especially at a low F-to-enter or F-to-remove threshold.

2.2 PLS Based Variable Selection Method

The standard multiple regression model defined by the equation

$$y = X\beta + \varepsilon, \tag{2.1}$$

where X is a $n \times p$ matrix of explanatory variables (predictors), y is a $n \times 1$ vector of response variable, β is a $p \times 1$ vector of unknown parameters, and ε is a $n \times 1$ vector of error terms whose rows are identically and independently distributed. The ordinary least squares (OLS) estimator of β, b_{OLS}, in the model given by (2.1) is the solution of the following optimization problem:

$$b_{\text{OLS}} = \arg\max_{b} \text{corr}(Xb, y)^2. \tag{2.2}$$

In many applications of multiple regression, multicollinearity is inevitable as a result of large number of variables collected by modern technologies of computers, networks, and sensors. Despite having desirable properties, the OLS estimator can have an extremely large variance and results in imprecise prediction when the data are multicollinear. Moreover, solution of (2.2) is not unique when $n < p$.

Recently, partial least squares (PLS) has become an important statistical tool for modeling relations between sets of observed variables by means of latent variables, which have the "best" predictive power, especially for statistical problems dealing with high dimensional data sets. PLS is a member of nonlinear iterative least squares (NILES) procedures developed by Wold (1966, 1975). In order to deal with multicollinearity and/or dimensionality problem, we regress the response variable y on a subset of the k orthogonal (latent) vectors stored in a score matrix of size $n \times k$ by which important features of X have been retained. Score matrix is formed by taking linear combinations of columns of X.

PLS regression constructs the columns of score matrix, $T = [t_1, t_2, \cdots, t_k]$, by solving the following optimization problem for $h = 1, 2, \cdots, k$ $(k \leq p)$:

$$\begin{aligned} &r_h = \arg\max_{||r||=1} \text{Cov}(X_r, y)^2 \\ &\text{subject to } t_h' t_j = 0 \text{ for } 1 \leq j < h. \end{aligned}$$

So, PLSR balances the maximal correlation criteria for OLS given in (2.1) with the requirement of explaining as much as variability in both X and y -space.

PLS is a distribution free approach to regression and path modeling, robust against other data structural problems such as skew distributions and omission of regressors. It is a useful tool to reveal a few underlying predictive factors that account for most of the variation in the response (Geladi, 2005; Cassel *et al.*, 1999).

Usually to explore further which predictors can be eliminated from the analysis, the regression coefficients for the standardized data are evaluated. Predictors with small absolute coefficients making a small contribution to the response can be eliminated from the model. Another statistic summarizing the contribution of a variable to the model is the variable importance for projection (VIP), which estimates the relative importance of each independent variable X in fitting both predictors and responses and thus is often used for variable selection (Abudu *et al.*, 2010).

The VIP scores are defined as

$$\frac{p\sum_{i=1}^{k} SS_i W_{ij}^2}{\sum_{i=1}^{k} SS_i}, \quad i = 1, 2, \cdots, k, \quad j = 1, 2, \cdots, p. \tag{2.3}$$

It is assumed that there are k latent variables selected from p predictor variables (X_j). A regression model based on PLS was built using the latent vectors obtained from X to predict Y. As shown in (2.3), W_{ij} represents the loading vector between the ith latent variable and the independent variable X_j. SS_i implies the response variance explained by the ith latent variable when a PLS model is developed. The ratio of the variance explained by X_j to the total variance implies the relative influence of each predictor variable on the total variance $\sum_{i=1}^{k} SS_i W_{ij}^2 / \sum_{i=1}^{k} SS_i$. The VIP score is obtained when the number of the predictor variables is multiplied by the influence of each predictor variable (Han and Kim, 2003). Therefore the VIP coefficients reflect the relative importance of each predictor variable in fitting both predictors and responses. The larger a VIP score, the greater the contribution of the associated predictor variable provides to the PLS model (Han and Kim, 2003).

Predictors with relatively small VIP coefficients (in absolute value), less than 1 (one), are considered to have small contribution to the prediction and might be excluded from the model; predictors with VIP scores close to or greater than 1 (one) can be considered the most relevant for explaining Y. This is called "greater than one rule", and is used as the criterion for variable selection (Han and Kim, 2003; Chong and Jun, 2005).

2.3 The LASSO, Forward Stagewise and LARS Methods

There are two main reasons why the data analyst is not satisfied with the OLS estimates based variable selection methods: 1. prediction accuracy, the OLS estimator often is best linear unbiased but with high variance. Better prediction performance can be achieved by sacrificing a little bias. Because the small increase in bias can be traded by a larger reduction in variance, resulting in a more desirable estimator overall. 2. interpretation; analysts often wish to determine

smaller subset of large number of predictors, which has the strongest effect. Subset selection can be extremely variable, important repressors may drop in any step, small changes in data produce different models; therefore it destroys the prediction accuracy. Shrinking and setting some coefficients to zero, help us to have prediction accuracy or small subset of predictors. Although ridge regression, principal component and PLS based methods provide more stable models, they do not set any coefficients to zero, and does not provide us interpretable models.

LASSO proposed by Tibshirani (1996) is a popular technique for model selection and estimation in linear regression models. It employs an L1-type penalty on the regression coefficients which tends to produce sparse models, and thus is often used as a variable selection tool as in Tibshirani (1997) and Osborne *et al.* (2000). Knight and Fu (2000) studied the asymptotic properties of LASSO-type estimators. They showed that under appropriate conditions, the LASSO estimators are consistent for estimating the regression coefficients. It has been demonstrated in Tibshirani (1996) that the LASSO is more stable and accurate than traditional variable selection methods such as best subset selection. Efron *et al.* (2004) proposed the LARS, and showed that there is a close connection between the LARS, the LASSO, and the Forward Stagewise regression. Each of these procedures involves a tuning parameter that is chosen to minimize the prediction error.

Consider the common Gaussian linear regression model. LASSO estimate is the solution to

$$\min_{\beta}(Y - X\beta)^T(Y - X\beta), \quad \text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t,$$

where $t \geq 0$ is a tuning parameter. An alternative formulation of the LASSO is to solve the penalized likelihood problem

$$\min_{\beta} \frac{1}{n}(Y - X\beta)^T(Y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Both formulations are equivalent in the sense that, for any given $\lambda \in [0; \infty)$, there exists a $t \geq 0$ such that the two problems have the same solution, and vice versa. Tuning parameter can be chosen based on Mallows' C_p, AIC and BIC. The LASSO is a constraint version of the OLS estimates. It shrinks some coefficients and sets rest of them to zero, also continues with the useful properties of both subset selection and ridge regression.

The forward stagewise regression is an iterative procedure, where successive estimates are built via a series of small steps. Letting $\eta_0 = X\beta$, and beginning with $\hat{\eta}_0 = 0$ is the current estimate, the next step is taken in the direction of the greatest correlation between covariate X_j and the current residual.

The LARS is a new model selection algorithm which is a useful and less greedy version of traditional forward selection or forward stepwise regression. It uses mathematical formula to speed up the computations. The number of the covariates is as the same as the number of steps that are required for the full set of solutions. The LARS algorithm starts with all coefficients equal to zero, then finds the predictor that has the largest correlation with the response variable, say X_{j1} and increases the coefficient in the direction of the sign of its correlation with the response. Next it finds residuals, stops when any other predictor, say X_{j2} has as much correlation with the residual computed for the first coefficient as does X_{j1}. The LARS proceeds in a direction equiangular between these coefficients X_{j1} and X_{j2} until the third variable X_{j3}, has the most correlation with the residual. Then LARS proceeds equiangular between these three coefficients, which is the "least angle direction" until the fourth variable enters the model. LARS procedure continues until all variables are in the model. Efron *et al.* (2004) showed that there is a close relationship among these procedures in that they give almost identical solution paths.

3. Analysis of Results

As we mentioned earlier, there is a need for more efficient and timely methods to diagnose such disease. Therefore it would be beneficial to take an action at early stage of the disease if model could be constructed to predict the late disease outcomes on day 10 (i.e., lung weight increase and *C. pneumonia* load) by using transcript variables from the average measurements in 16 groups in day 3.

There are two groups in day 3 and day 10 and two responses of different disease outcomes, lung weight increase % (Y_1) and lung *C. pneumonia* load (i.e., chlamydial lung burden) (Y_2) defined as average percentage of lung weight increase (Y_1) and the $\log_{10}$ *C. pneumonia* / lung (Y_2), respectively. It is worth noting that these two responses measured on day 10 have no significant correlation to each other (p-value = 0.8238).

According to the result of the Hotelling's T-square test to test if there was a difference between the mean vector of two responses in day 3 and day 10 (Wilk's $\Lambda = 0.83280946$, p-value = 0.0705), the late disease outcomes can be predicted from the early transcript variables, which means a model can be constructed by using day 3 transcript variables to predict the late disease outcomes on day 10. Therefore, in our study, the 23 independent variables are chosen from day 3 and both dependent variables stand for the values on day 10. As a result of this, we are interested in determining which early transcript variables on day 3 would be contributing to the prediction of the average late disease outcomes (i.e., the means of lung weight increase on Day 10 and lung *C. pneumonia* load on Day 10, respectively). We used PROC GLMSELECT in SAS 9.2 for the results of

stepwise, LASSO and LARS methods.

3.1 Variable Selection Results Based on Stepwise Regression Methods

The stepwise regression was attempted in this study, transcript variables X_4, X_5 and X_{13} were selected to construct a model to predict Y_1 on day 10 and variables X_5, X_8, X_{14} were selected to build model to predict Y_2 on day 10 (Table 1a and 1b).

In a forward selection, with specified the number of candidate effect that was entered in the model sequentially, specific subset was determined. For example, if the number of selection steps was set at 7 or 8, variables X_1, X_2, X_3, X_4, X_5, X_{13} and X_{21} were selected to construct model to predict Y_1 on day 10. Similarly, variables X_2, X_5, X_8, X_{14}, X_{17} and X_{19} were selected to build model to predict Y_2 on day 10. More variable selection combinations are shown in Table 1a and 1b. However, if the number was not explicitly specified, a "best" subset model will be determined by the Schwarz Bayesian Information Criterion (SBC). Add effects that give the lowest value of the SBC statistic at each step and stop at the step where adding any effect would increase the SBC statistic. Therefore, variables X_4, X_5 and X_{13} were selected to construct model to predict Y_1 on day 10 and variables X_5, X_8, X_{14} were selected to build model to predict Y_2 on day 10.

Table 1a: Summary of variable selection from multiple statistical methods in model Y_1

(Y_1)	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Stepwise				X^*	X^*			
Forward	*#	*#	#	X^*#	X^*#			
LASSO & LARS	*#		#	*#		*#		X^*#
PLS							S	S
(Y_1)	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
Stepwise					X^*#			
Forward					X^*#			
LASSO & LARS					X^*#			#
PLS		S	S		S			
(Y_1)	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	
Stepwise								
Forward					#			
LASSO & LARS				*#				
PLS								

Symbols: X: Include all and find the best, *: The best model with 5 or 6 variables, #: The best model with 7 or 8 variables, S: PLS.

Table 1b: Summary of variable selection from multiple statistical methods in model Y_2

(Y_2)	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Stepwise		X			$X^*\#$			$X^*\#$
Forward		$X^*\#$			$X^*\#$	#		$X^*\#$
LASSO & LARS		#			$^*\#$		$^*\#$	#
PLS				S	S		S	
(Y_2)	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}
Stepwise						$X^*\#$		
Forward						$X^*\#$		
LASSO & LARS	$^*\#$			$^*\#$		$X^*\#$		$^*\#$
PLS	S			S		S		
(Y_2)	X_{17}	X_{18}	X_{19}	X_{20}	X_{21}	X_{22}	X_{23}	
Stepwise								
Forward	#		$^*\#$					
LASSO & LARS								
PLS			S				S	

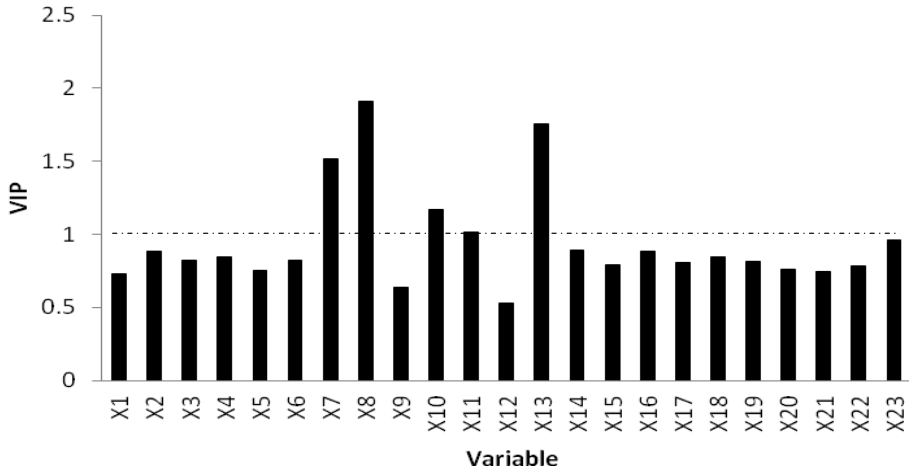
Symbols: X: Include all and find the best, *: The best model with 5 or 6 variables, #: The best model with 7 or 8 variables, S: PLS.

3.2 Variable Selection Results Based on PLS Method

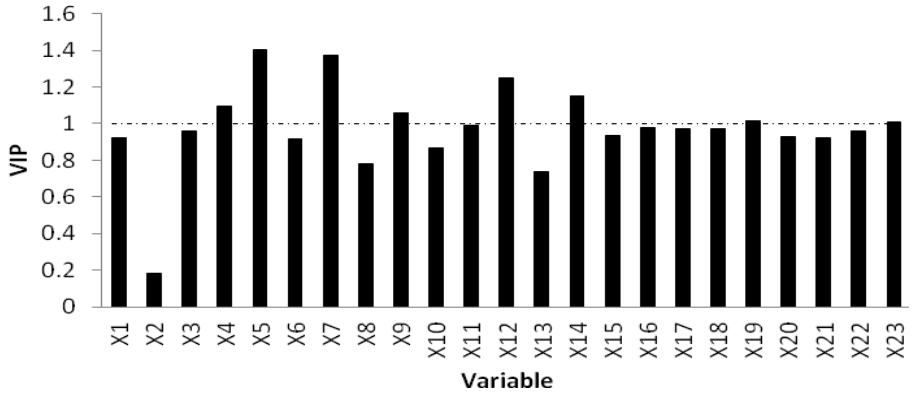
According to "greater than 1 rule", the variables X_1-X_6, X_9 and X_{12}-X_{23} with small VIP scores in the PLS model 1 are excluded and the variables X_7, X_8, X_{10}, X_{11}, and X_{13} are selected to build model 1 to predict Y_1 (Figure 1(a)). According to the VIP scores plot (Figure 1(b)), variables X_1-X_3, X_6, X_8, X_{10}, X_{11}, X_{13}, X_{15}-X_{18} and X_{20}-X_{22} have small VIP values in the PLS model 2 and variables X_4, X_5, X_7, X_9, X_{12}, X_{14}, X_{19} and X_{23} are selected to construct model 2 to predict Y_2.

3.3 Variable Selection Results Based on LASSO Method

Since we obtained identical models for the forward stagewise and LARS, we will provide the results from the LASSO method. If the number of selection steps was set at 7 or 8, variables X_1, X_3, X_4, X_6, X_8, X_{13}, X_{16} and X_{20} were selected to construct model to predict Y_1 on day 10. Similarly, variables X_2, X_5, X_7, X_8, X_9, X_{12}, X_{14} and X_{16} were selected to build a model to predict Y_2 on day 10. Without the specified the number of selection steps, X_8 and X_{13} were chosen in model Y_1 and X_{14} was selected in model Y_2. More variable selection combinations are shown in Table 1a and 1b.



(a)



(b)

Figure 1: VIP scores plot of PLS model for Y_1 (a) and Y_2 (b)

3.4 Summary Results of the Variable Selection Methods

Utilizing statistical methods like stepwise regression, LASSO and PLS-VIP, several variables have been selected with specified different selection and stopping criteria. Table 1a and 1b summarizes different combination of variables selected from stepwise regression, PLS method in both model 1 and model 2 with specified number of candidate effects or selection steps. The four variables X_4, X_5, X_8, X_{13} are the "most" selected variables from these methods to construct model Y_1, and the four variables X_2, X_5, X_8, X_{14} are the "most" selected variables from these methods to construct model Y_2. We have also detected one outlier (7^{th}

observation) with moderately high residual (studentized residual = 2.985) based on the fitted model of Y_1 vs. X_4, X_5, X_8, X_{13}. However this observation did not affect the results of the statistical analysis, therefore we included this observation in our data.

The value of various all-possible-regression selection criteria are shown in the Table 2 with these variables and the "Criterion Panel" in Figure 2 provides a graphical view of the evolution of these fit criteria as the "most" selected variables entered in the model sequentially. Good linear models are achieved with small values of AIC (Akaike's criterion, Akaike, 1969, Darlington, 1968), SBC (Schwarz's Bayesian criterion, Schwarz, 1978), BIC (Sawa Bayesian information criterion, Sawa, 1978), AICC (a small sample bias corrected version of AIC, Hurvich and Tsai, 1989) and C_p (Mallows, 1973) and a high adjusted R^2 value (Darlington, 1968) close to 1.

Table 2: Fit summary statistics of the selected model

Model	Variables	Adj-R^2	R^2	AIC	AICC	BIC	SBC	C_p
Y_1	X_4, X_5, X_8, X_{13}	0.8008	0.8539	96.776	106.11	84.909	82.639	5
Y_2	X_2, X_5, X_8, X_{14}	0.8407	0.8831	−10.16	−0.824	−22.02	−24.29	5

4. Conclusions

In this article, various variable selection methods such as stepwise regression, forward selection, PLS and LASSO were employed to fulfill the preselection step. At this step, for example, 23 descriptors were reduced to 3 by stepwise regression and forward selection, 2 by LASSO without specified number of variable effects, and 5 by PLS-VIP in model 1. These preselected variables served as a starting pool for the comparison of variable selection methods. After the step of preselection, the "most" variables are selected from these methods to construct models and good linear models are achieved with small AIC, SBC, BIC, AICC and C_p and a high adjusted R^2 value close to 1.

Although stepwise regression methods are often used for variable selection due to their simplicity, it cannot overcome the over-fitting in our case because of the variables outnumbering the observations and high correlation among the predictor variables. Besides, forward regression method yielded unsatisfactory results due to rank deficiency problem. To overcome the multicollinearity problem, a possible solution is to use only a subset of the predictor variables, where the subset is chosen so that it does not contain multicollinearity problem. Numerous subset selection methods are available. In this paper, PLS-VIP method, LASSO and LARS were employed to circumvent such problem.

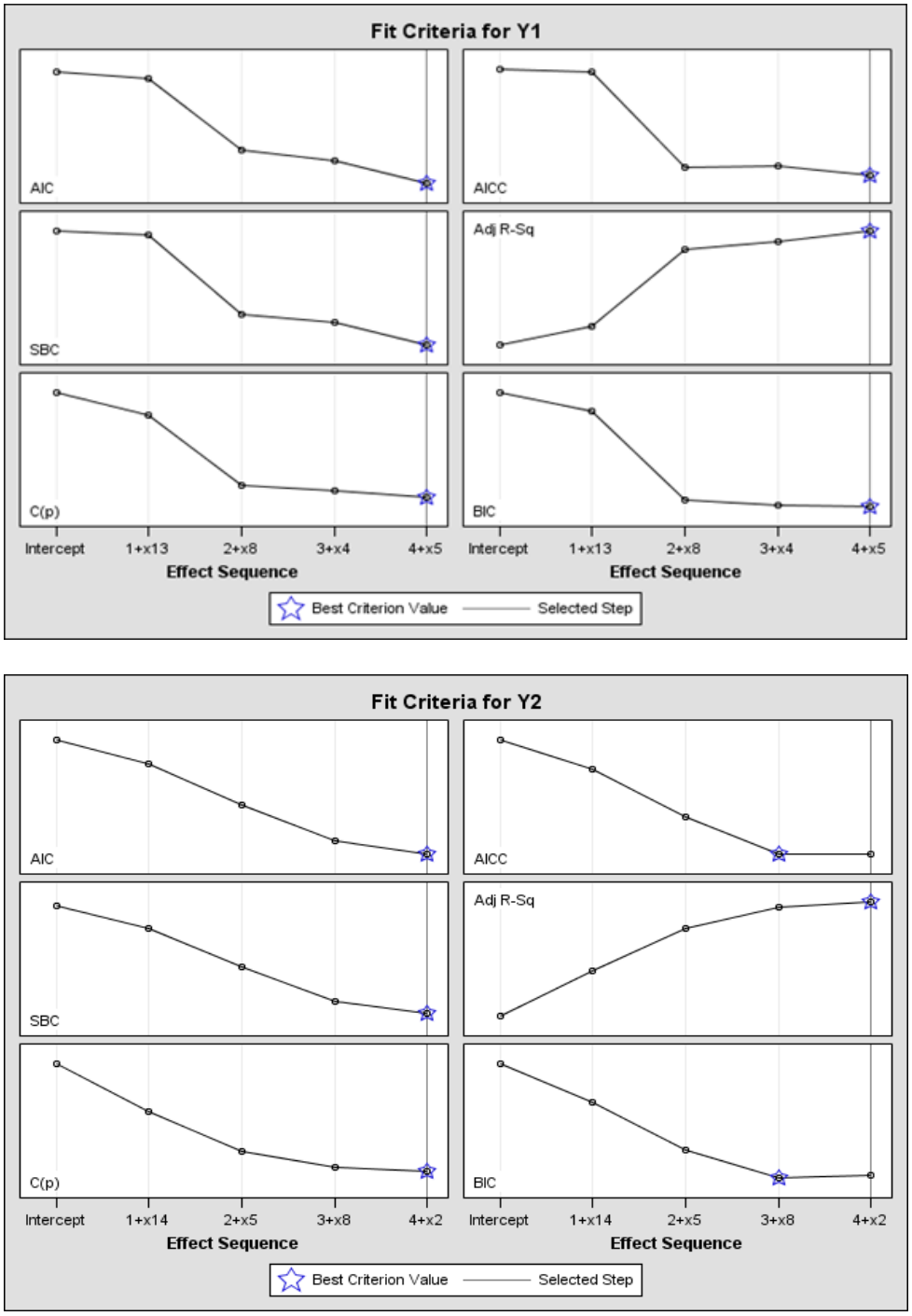


Figure 2: Criterion Panel for model Y_1 (upper) and Y_2 (lower)

Utilizing the statistical methods like stepwise regression, PLS-VIP, several variables were preselected and served as a starting pool. Simple model candidates were generated based on the preselected variable pool by employing several well-known model selection criteria in multiple regressions. The researchers of this study are concerned about the primary role of cellular markers and inflammatory regulators in *C. pneumonia* disease regulation. Transcript variables Tim3, GATA3, Lacf, Arg2 (X_4, X_5, X_8 and X_{13}) are the main variables of interest to study the *C. pneumonia* disease (lung weight increase) or *C. pneumonia* lung load outcomes. These excessive $CD4^+$ T helper cells immunity (Tim3, GATA3), macrophage activation (Arg2) and an exaggerated contribution of innate immunity of the early response of *C. pneumonia* (Lacf) are the key determinants in precipitation of late disease. IL-6 (X_{14}) as a critical regulator for amplifying inflammation had an enhancing effect on chlamydial growth in vitro. Studies also showed that blockade of IL-6 trans-signaling corrects hyperinflammation and increases chlamydial load in $Dusp1^{-/-}$ mice (Rodriguez *et al.*, 2005, 2010). As a hallmark of B-cells, CD19 (X_2) promotes the proliferation and survival of mature B cells. The studies show that B cells play an important role in the initiation of T cell responses to *Chlamydia trachomatis* (Mouse Pneumonitis) lung infection and pulmonary chlamydial infection and related to impaired cytokine production (Ramsey *et al.*, 1988; Yang *et al.*, 1998). Mutations in CD19 are associated with severe immunodeficiency syndromes characterized by diminished antibody production. CD19 has been used to diagnose cancers that arise from this type of cell notably B-cell lymphomas and also been implicated in autoimmune diseases and may be a useful treatment target (van Zelm *et al.*, 2006, 1995; Fujimoto *et al.*, 2007). However, there is no direct evidence shown that CD19 has a strong impact on chlamydial infection. The role of CD19 in the chlamydial growth is less well understood.

Models including these key variables may provide possible answers to the problem of molecular mechanisms of chlamydial pathogenesis. It is worth noting that variable GATA3 (X_5), which is necessary and sufficient for Th2 cytokine gene expression in CD4 T cells, is the most frequent variable selected from multiple statistical methods, suggested a primary role in immune responses to regulate the *C. pneumonia* disease. This result is consistent with the studies that Th2 immunity is a required component to control inflammation elicited by the Th1 component of the anti-chlamydial immune response (Wang, 2005).

Therefore, a better understanding of the regression relationship can be reached through a reduction of the number of regressors required for effective predication of the regressand. Since fewer variables are involved, further studies involving such regression relationship may be easier and less expensive to perform and researches can better focus on the variables of interest.

Acknowledgements

The authors thank Drs. Wang and Kaltenboeck from Department of Pathobiology, Auburn University, for the data provided and their helpful suggestions.

References

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* **21**, 243-247.

Abudu, S., King, J. and Pagano, T. (2010). Application of partial least-squares regression in seasonal streamflow forecasting. *Journal of Hydrologic Engineering* **15**, 612-623.

Boman, J., Gaydos, C. A. and Quinn, T. C. (1999). Molecular diagnosis of *Chlamydia pneumoniae* infection. *Journal of Clinical Microbiology* **37**, 3791-3799.

Cannon, C. P., Braunwald, E., McCabe, C. H., Grayston, J. T., Muhlestein, B., Giugliano, R. P., Cairns, R. and Skene, A. M. (2005). Antibiotic treatment of *Chlamydia pneumoniae* after acute coronary syndrome. *New England Journal of Medicine* **352**, 1646-1654.

Cassel, C. M., Hackl, P. and Westlund, A. H. (1999). Robustness of partial least squares method for estimating latent variable quality structures. *Journal of Applied Statistics* **26**, 435-446.

Chong, I. G. and Jun, C. H. (2005). Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems* **78**, 103-112.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression (with discussion). *Annals of Statistics* **32**, 407-499.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin* **69**, 161-182.

Frikha-Gargouri, O., Gdoura, R., Znazen, A., Arab, N. B., Gargouri, J., Jemaa, M. B. and Hammami, A. (2008). Evaluation and optimization of a commercial enzyme linked immunosorbent assay for detection of Chlamydophila pneumoniae IgA antibodies. *BMC Infectious Diseases* **8**, 98.

Freund, Y. and Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55**, 119-139.

Fujimoto, M. and Sato, S. (2007). B cell signaling and autoimmune diseases: CD19/CD22 loop as a B cell signaling device to regulate the balance of autoimmunity. *Journal of Dermatological Science* **46**, 1-9.

Geladi, P. (2005). Notes on the history and nature of partial least squares (PLS) modelling. *Journal of Chemometrics* **2**, 231-246.

Han, S. H. and Kim, J. (2003). A comparison of screening methods: selecting important design variables for modeling product usability. *International Journal of Industrial Ergonomics* **32**, 189-198.

Hastie, T., Taylor, J., Tibshirani, R. and Walther, G. (2007). Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics* **1**, 1-29.

Hurvich, C. M. and Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297-307.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics* **28**, 1356-1378.

Li, Y., Ahluwalia, S. K., Borovkov, A., Loskutov, A., Wang, C., Gao, D., Poudel. A., Sykes, K. F. and Kaltenboeck, B. (2010). Novel *Chlamydia pneumoniae* vaccine candidates confirmed by Th1-enhanced genetic immunization. *Vaccine* **28**, 1598-1605.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

Osborne, M. R., Presnell, B. and Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis* **20**, 389-404.

Ramsey, K. H., Soderberg, L. S. and Rank, R. G. (1988). Resolution of chlamydial genital infection in B-cell-deficient mice and immunity to reinfection. *Infection and Immunity* **56**, 1320-1325.

Rodriguez, N., Dietrich, H., Mossbrugger, I., Weintz, G., Scheller, J., Hammer, M., Quintanilla-Martinez, L., Rose-John, S., Miethke, T. and Lang, R. (2010). Increased inflammation and impaired resistance to *Chlamydia pneumoniae* infection in Dusp1$^{-/-}$ mice: critical role of IL-6. *Journal of leukocyte biology* **88**, 579-587.

Rodriguez, N., Fend, F., Jennen, L., Schiemann, M., Wantia, N., Prazeres da Costa, C. U., Dürr, S., Heinzmann, U., Wagner, H. and Miethke, T. (2005). Polymorphonuclear neutrophils improve replication of *Chlamydia pneumoniae* in vivo upon MyD88-dependent attraction. *Journal of Immunology* **174**, 4836-4844.

Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica* **46**, 1273-1291.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.

Scheuermann, R. H. and Racila, E. (1995). CD19 antigen in leukemia and lymphoma diagnosis and immunotherapy. *Leukemia and Lymphoma* **18**, 385-397.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267-288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine* **16**, 385-395.

van Zelm, M. C., Reisli, I., van der Burg, M., Castaño, D., van Noesel, C. J., van Tol, M. J., Woellner, C., Grimbacher, B., Patiño, P. J., van Dongen, J. J. and Franco, J. L. (2006). An antibody-deficiency syndrome due to mutations in the CD19 gene. *New England Journal of Medicine* **354**, 1901-1912.

Wang, C. M. (2005). Multivariate analysis of *Chlamydia pneumoniae* lung infection in two inbred mouse strains. Ph.D. Dissertation, Auburn University, Auburn, Alabama.

Wang, C. M., van Ginkel, F. W., Kim, T., Li, D., Li, Y., Dennis, J. C. and Kaltenboeck, B. (2008). Temporal delay of peak T-cell immunity determines *Chlamydia pulmonary* disease in mice. *Infection and Immunity* **76**, 4913-4923.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis* (Edited by P. R. Krishnaiaah), 391-420. Academic Press, New York.

Wold, H. (1975). Path models with latent variables: the NIPALS approach. In *Quantitative Sociology*: *International Perspectives on Mathematical and Statistical Modeling* (Edited by H. M. Blalock *et al.*), 307-357. Academic Press, New York.

Yang, X. and Brunham, R. C. (1998). Gene knockout B cell-deficient mice demonstrate that B cells play an important role in the initiation of T cell responses to *Chlamydia trachomatis* (mouse pneumonitis) lung infection. *Journal of Immunology* **161**, 1439-1446.

Received December 6, 2011; accepted January 1, 2013.

Yuan Kang
Department of Mathematics and Statistics Auburn
University Auburn
Auburn, AL. 36849-5168, USA
kangyua@tigermail.auburn.edu

Nedret Billor
Department of Mathematics and Statistics Auburn
University Auburn
Auburn, AL. 36849-5168, USA
billone@auburn.edu