IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 26, 2021, accepted July 6, 2021, date of publication July 9, 2021, date of current version July 15, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3095914

Linear Model Selection and Regularization for Serum Prostate-Specific Antigen Prediction of Patients with Prostate Cancer Using R

GONGLI LI AND HAN LI
The Australian National University, Canberra, ACT 2601, Australia
Corresponding author: Gongli Li (u6141461@anu.edu.au)

ABSTRACT Prostate cancer is the commonly diagnosed cancer worldwide, and there were 1,276 thousand new prostate cancer cases and 359 thousand deaths in 2018. Prostate-specific antigen (PSA) blood level is often elevated in men with prostate cancer, so PSA testing can detect prostate tumours when they are small, low-grade, and localized. The PSA testing is hard to apply on the less developed and poor areas without sufficient medical funds, so the early accurate PSA level prediction by statistical machine learning models is significant to avoid later stages of prostate cancer that spread outside the Prostate. In this literature, we compare three linear model selection and regularization methods (shrinkage, subset selection, dimension reduction) and nine candidate models (OLS regression, Ridge regression, Lasso regression, Elastic net, best subset selection, forward subset selection, backward subset selection, PCR, PLS) based on leave-one-out-cross-validation (LOOCV) prediction error. As the selection criteria leave-one-out cross-validation is sensitive to outliers, Mahalanobis distance is used for outlier detection and deletion before running each model. The shrinkage method (only lasso and elastic net models) and subset selection method (based on adjusted R^2, BIC, Cp, and cross-validation prediction error) can select the variables out. The feature selection results show that prostate weight, cancer volume, amount of benign prostatic hyperplasia, and whether seminal vesicle invasion is necessary variables must include predicting PSA. Age and capsular penetration are the least important variables. The variables of Gleason score, a percent of Gleason scores 4 or 5 are essential sometimes. All the diagnostic figures and results are coded by R, open access, and published on IEEE Xplore Code Ocean.

INDEX TERMS Machine learning, linear model selection and regularization, prostate-specific antigen prediction, prostate cancer screening, R programming.

I. INTRODUCTION

An adenocarcinoma is a type of cancer that arises in the cells of glands. Most prostate gland cells are of the glandular type, so adenocarcinoma is the most common cancer type in the prostate [1]. Prostate cancer is a commonly diagnosed cancer worldwide and crucial challenges in developed and developing countries [2]. Statistical results show that there were 1,276 thousand new prostate cancer cases and 359 thousand deaths in 2018 [3]. However, the prognosis for prostate cancer is relatively good if it is detected early. Prostate-specific antigen (PSA) is a protein that is produced in the glandular epithelium of the prostate. It is secreted into the prostatic acini lumen and has an important physiological role in prostatic fluid [4]. As The blood level of Prostate-specific antigen is often elevated in men with prostate cancer [5], PSA has profoundly affected the diagnosis and treatment of prostate cancer, and prostate tumours can be detected by PSA testing when they are small, low-grade, and localized [6]. Some doctors and professional organizations suggested that a man has to take PSA examinations every year from age 50 [7], but most medical testing, including PSA testing, is hard to apply on the less developed and poor areas without sufficient medical funds [8], so the early accurate PSA prediction by statistical learning models is meaningful to avoid later stages of prostate cancer that spread outside the prostate. In recent years, plenty of machine learning classification models have been proposed and published for Prostate

The associate editor coordinating the review of this manuscript and approving it for publication was Gina Tourassi.

TABLE 1. Description and statistics summary of all nine attributes for PSA prostate cancer dataset.

Attributes	Description	Mean±SD / Set
lcavol	log of cancer volume (cc).	1.350±1.179
lweight	log of prostate weight (g)	3.629±0.428
age	age of a patient	68.866±7.445
lbph	log of the amount of benign prostatic hyperplasia (cm^2)	0.100±1.451
svi	Seminal vesicle invasion; 1=Yes, 0=No	{'0'=76, '1'=21}
lcp	log of capsular penetration (cm)	-0.179±1.398
gleason	Gleason score	{'6'=35, '7'=56, '8'=1, '9'=5}
pgg45	percent of Gleason scores 4 or 5	24.381±28.204
lpsa	log Prostate specific antigen (ng/mL)	2.478±1.154

cancer monitoring as well as detection, such as decision tree (DT) [9], Logistic Regression (LG) [10], and Random Forest [11]. Besides, some literature has been proposed computer-vision models to predict and detect Prostate cancer, such as Fully Convolutional Neutral Network (FCNN) [12], Near-infrared (NIRF) [13], and Artificial Neural Networks System (ANNS) [14]. However, few papers predict Prostate-specific antigen for Patients with Adenocarcinoma of the Prostate.

We compare three linear model selection and regularization methods (subset selection, dimension reduction, and shrinkage) and eight models (forward selection, backward selection, exhaustive selection, PLS, PCR, ridge, lasso, and elastic net) to find their optimal tuning parameters and leave-one-out cross-validation prediction error from the pre-process prostate cancer dataset. Besides, lasso, elastic net, and subset selection did variable selection to choose the necessary three predictors and not important predictors for PSA prediction. All the diagnostic figures and results are coded by R, open access, and IEEE Xplore Code Ocean. The framework of the remaining paper is as follows: Section II describes the dataset and proposed methodology. Finally, the experimental results are reported with the interpretation and concluded with a discussion in section III.

II. DATA AND METHODS

This section focuses on the data and methodology used for the literature. Subsections II-A, II-B, and II-C respectively explain the dataset, background of a PSA test and proposed framework.

A. DATA DESCRIPTION

The linear Machine Learning models were trained and tested on a pubic source prostate-specific antigen dataset [15] of 97 male patients before radical prostatectomy. The surgical operation that removes the entire prostate gland along with some surrounding tissue.

Table 1 shows the descriptions and brief statistical summary of the attributes, where the set of svi and gleason are discrete numerical variables only with two and four values. The Gleason grading system divides the two largest tumour areas in a tissue sample into (1-5) levels, where 1 is the least aggressive and 5 is the most aggressive, then add these two levels together to get a Gleason score. The BPH and capsular penetration variables originally contained zeros, and a small number was substituted before the log transform was taken. The original paper did not declare why the log transform was taken though PSA varies over a wide range, probably aiming for the variable's linearity. Besides, it is also not clear why the variable pgg45 was constructed.

Fig. 1 shows the bar plots for gleason and svi and density-histogram plots for other variables. Fig. 2 shows the density correlation plots and Pearson correlation, calculated as (1), for the other six variables.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{1}$$

The left-bottom correlation plot between a log of cancer volume and a prostate-specific antigen log shows a high positive Pearson correlation of 0.734. Besides, a log of prostate weight, a capsular penetration log, and percent of Gleason scores 4 or 5 are also positively correlated with a log of prostate-specific antigen with Pearson correlation equal to 0.433, 0.549, and 0.422. Fig. 3 shows the box plots of variables of svi and gleason with outliers. The outliers are data points x that do not fall within the distances as (2).

$$x \notin [Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR] \tag{2}$$

B. PSA TESTING

Prostate-specific antigen is a protein generated by normal and malignant prostate cells. In the PSA test, the laboratory analysis blood samples usually reported nanograms of PSA per mL of blood. In the past, most doctors thought that PSA levels of 4.0 ng/mL or lower were normal [16]. However, recent studies have shown that some men with PSA levels below 4.0 ng/mL have prostate cancer, while many men with higher PSA levels do not have prostate cancer [17]. For instance, if a person has prostatitis or a urinary tract infection, his PSA level will usually rise. In contrast, some drugs used to treat benign prostatic hyperplasia can reduce men's PSA levels, such as finasteride and dutasteride [18]. However, generally speaking, the higher a person's PSA level, the more likely he is to develop prostate cancer.

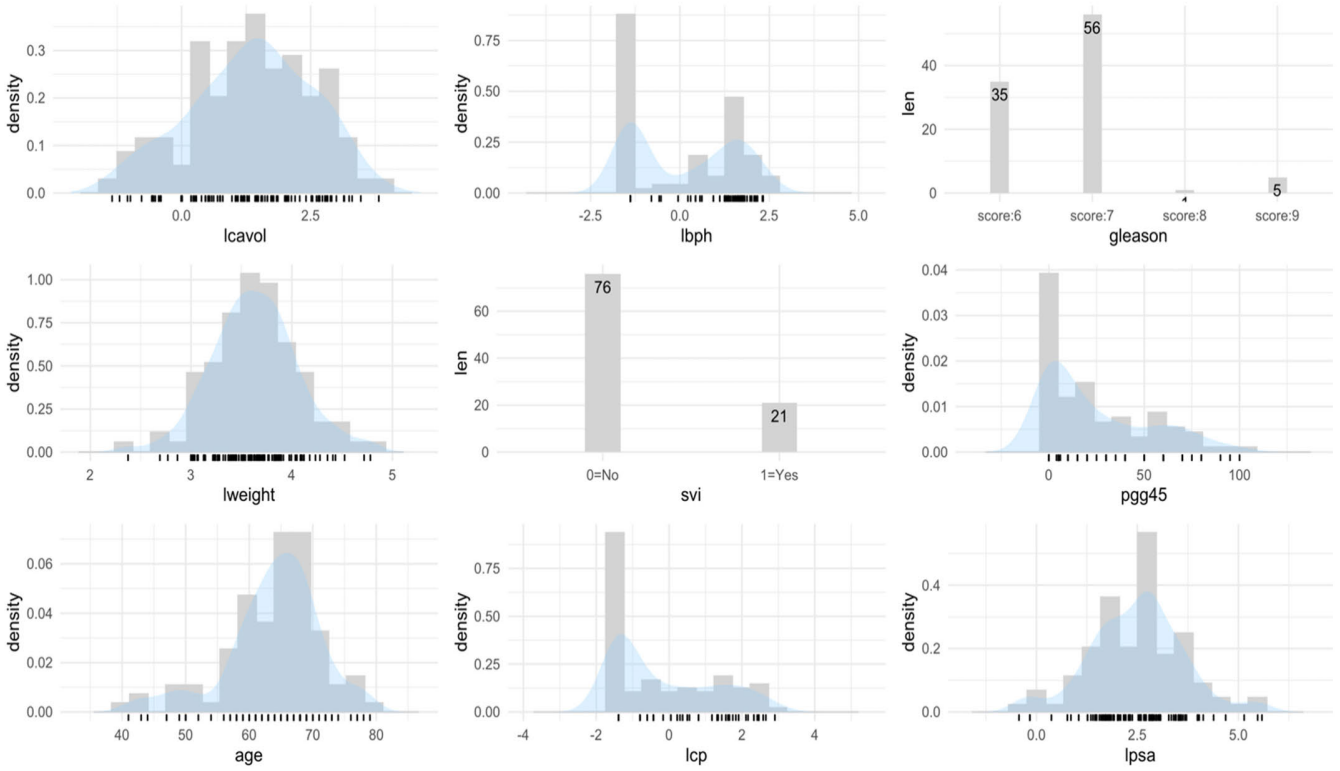


FIGURE 1. The population distribution of all attributes in the PSA prostate cancer dataset.

C. PROPOSED FRAMEWORK

In this literature, the proposed framework has been illustrated in Fig. 4. Transfer numerical variables to Factor variables, data standardization, and outlier detection are applied in data pre-processing. There are ten candidate models (OLS regression, Ridge regression, Lasso regression, Elastic net, best subset selection, forward subset selection, backward subset selection, PCR, and PLS) using to find the lowest LOOCV prediction error. Subset selection methods using adjusted R^2, BIC, and Cp to eliminate unnecessary variables do not follow this framework.

1) FACTOR VARIABLES AND DATA STANDARDIZATION

In Fig. 3, the box plots show prostate-specific antigen levels are significantly different for patients whether seminal vesicle invasion (p-value < 0.001) and a Gleason score greater than 6 (p-value < 0.001). Therefore, the Gleason score is classified by the cut-off level of a Gleason score greater than 6. These two variables are treated as factor variables. Standardization is the concept and step of scaling and transforming to equal each feature's equal contribution. As formulation (3), Z-score normalization is one of the standardization techniques for achieving standard normal distribution with zero mean and unit variance.

$$S(x) = \frac{x - \bar{x}}{\sigma} \tag{3}$$

The variables of lcavol, lweight, age, lbph, lcp, and pgg45 are treated as numerical variables and did the Z-score normalization.

2) MAHALANOBIS DISTANCE TO DETECT OUTLIER

Mahalanobis distance [19], introduced by Prof. P. C Mahalanobis in 1936, is a reasonable multivariate distance metric that measures the distance between a point and a distribution. Mahalanobis distance calculates the distance between two points by considering a covariance factor that is a difference between that and Euclidean distance. The Mahalanobis distance between two points p_1 and p_2 is presented as (4), where S is the covariance of multivariate data X.

$$D^2 = \left(x_{p_1} - x_{p_2}\right)^T S^{-1} \left(x_{p_1} - x_{p_2}\right) \tag{4}$$

The Mahalanobis distance is an effective way to find outliers for multivariate data [20]. The idea is to calculate the Mahalanobis distance between each point and centre that can be chosen as a mean value of multivariate data as (5). where

$$x_i = \left[x_{i,1}, x_{i,2}, \ldots, x_{i,p}\right]', \bar{x} = \left[\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p\right]'.$$
$$D_i^2 = (x_i - \bar{x})^T \cdot S^{-1} \cdot (x_i - \bar{x}) \tag{5}$$

When n is relatively large and X is a k-dimensional Gaussian random vector with mean vector μ and rank k covariance matrix C, D_i^2 follows χ_k^2, proved in [21]. Based on

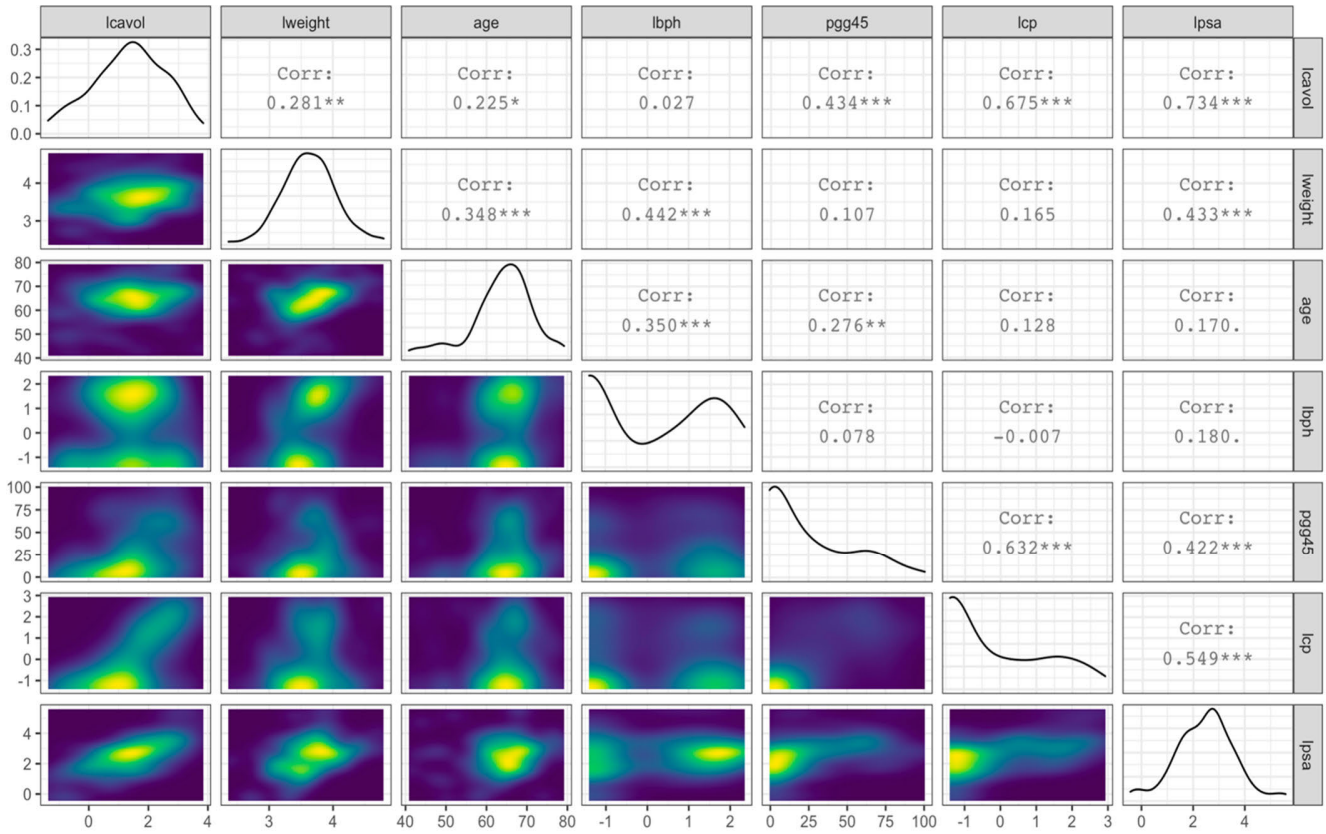


FIGURE 2. The density correlation plots and pearson correlation for numerical variables of lcavol, lweight, age, lbph, pgg45, lcp, and lpsa.

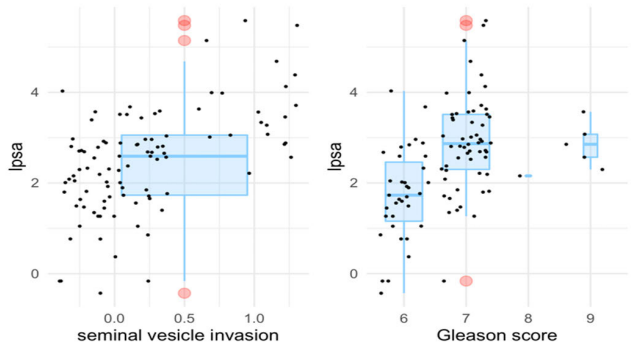


FIGURE 3. The box plots of variables of svi, gleason, and capsular penetration.

a chosen critical p-value $\alpha$with its critical value, it is $1-\alpha$ confident that the i-th observation is an outlier if $D_i^2 \geq \chi_k^2(1-\alpha)$.

3) OLS REGRESSION

The loss function of ordinary least squares regression is finding the plane that minimizes the sum of squared errors (SSE) between the observed and predicted response (6).

$$minimize\left\{SSE=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2\right\} \tag{6}$$

The OLS performance depends on the key assumptions of OLS regression:

- No or little multicollinearity
- There are more observations (n) than features (p)
- Homoscedastic (constant variance in residuals)
- No autocorrelation
- Multivariate normality
- Linear relationship

However, the number of features (p) is large for many real-life data sets. The OLS assumptions are easy to be violated for large p; there are three classical methods (shrinkage, dimension reduction, subset selection) to solve the large features.

4) SUBSET SELECTION METHOD

The first step of best subset selection is to fit a separate least squares regression for each possible combination of the p predictors, and then identifying the best subset of each number of predictors from $1,\ldots,p$. The last step is to find the optimal number of predictors among the best subset using adjusted R^2, BIC, C_p, or cross-validation prediction error. The detail is shown in Algorithm 1.

As Algorithm 2, stepwise forward selection starts with an empty set of attributes as the minimal set. The most relevant attribute is chosen (having minimum p-value or maximum SSR) and added to the minimal set. As Algorithm 3, stepwise backward elimination is initial with all the attributes; in each

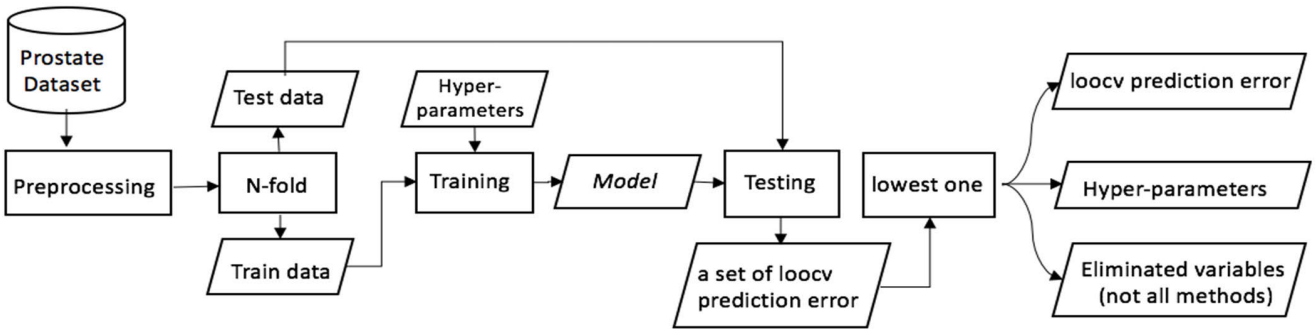


FIGURE 4. The proposed block diagram of PSA prediction by LOOCV prediction error.

iteration, one attribute is eliminated from the set of attributes whose p-value is highest or SSR is lowest.

Best subset selection and stepwise selection, each of which contains a subset of the p-predictors, are used to create a set of models. Adjusted R^2, BIC, Cp, and cross-validation prediction error are ways to determine which of these models is best. Choosing minimum test error estimation using the validation set or cross-validation is a direct method. Adjusted R^2, BIC, and Cp indirectly estimate test error methods by adjusting to the training error to avoid overfitting. For a fitted least-squares model containing d predictors, the Cp [22], invented by C.L. Mallows, estimates test MSE (7).

$$Cp = \frac{1}{n}(RSS + 2d\hat{\sigma}^2) \tag{7}$$

The penalty of $2d\hat{\sigma}^2$ increases as the number of predictors d in the model increases, where $\hat{\sigma}^2$ estimates the variance of the error ϵ using the full model containing all predictors. The technique compares the full model with a smaller model with "d" parameters and determines how much error is left unexplained by the partial model. Various suggestions have been made about exactly how the statistic should be interpreted, but the general consensus is that smaller Cp values are better as they indicate smaller amounts of unexplained error. BIC [23] is derived from a Bayesian perspective, but the resulting formula, as (8), is similar to Cp, where n is a number of observations.

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)\, d\hat{\sigma}^2) \tag{8}$$

The R^2 is defined as $1 - RSS/TSS$, whee $TSS = \sum (y_i - \bar{y})^2$. The R^2 increases when more variables are added because RSS always decreases as more variables including in the model. Adjusted R^2 [24] statistic (9), adds penalty on R^2 when increasing number of variables.

$$Adjusted\ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \tag{9}$$

5) SHRINKAGE (REGULARIZATION) METHOD

Regularized regression methods' loss function is very similar to OLS regression; however, a penalty parameter (P) is added (10).

$$minimize\ \{SSE + P\} \tag{10}$$

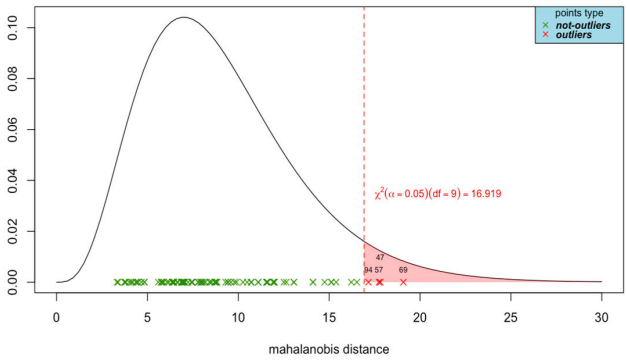


FIGURE 5. Mahalanobis distance method to identify outliers.

Regularized regression puts constraints on the magnitude of the coefficients and will progressively shrink the coefficients towards zero. This constraint reduces the magnitude and fluctuations of the coefficients and will reduce the variance of our model.

Ridge regression [25] constraints the coefficients by adding $\lambda \sum_{j=1}^{p} \beta_j^2$ to the loss function (11).

$$minimize\left\{SSE + \lambda \sum_{j=1}^{p} \beta_j^2\right\} \tag{11}$$

The penalty parameter is called L_2 because it means a second-order penalty is used on the coefficients. Tuning parameter λ controls the penalty parameter that can take on a wide range of values.

The full name of the Lasso model [26] is called as least absolute shrinkage and selection operator. L_1 penalty $\lambda \sum_{j=1}^{p} |\beta_j|$ in the loss function is used as (12).

$$minimize\left\{SSE + \lambda \sum_{j=1}^{p} |\beta_j|\right\} \tag{12}$$

Rather than ridge regression pushing variables to approximately but not equal to zero, the lasso penalty will actually shrinkage coefficients to zero, so the lasso model can improve the model with regularization and conduct automated variable selection.

The elastic net [27], which combines the L_1 and L_2 penalties, is a generalized ridge and lasso model (13).

$$minimize \left\{ SSE + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 \right\} \tag{13}$$

The elastic net model is proposed for effective regularization via the ridge penalty and the lasso penalty's feature selection characteristics.

6) DIMENSION REDUCTION METHOD

Both subset selection and shrinkage methods are defined using the original predictors $X_1, \ldots, X_p$. The dimension reduction method [28] transforms the predictors and then fit a least-squares model using the transformed variables. Let $Z_1, \ldots, Z_m$ represent linear combinations of the original p predictors, where $M < p, Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$ for some constants $\phi_{1m}, \ldots, \phi_{pm} (m = 1, \ldots, M)$. Therefore, the linear regression model is fitted using the least-squares as (14) for $i = 1, \ldots, n$.

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i \tag{14}$$

Principal component analysis [29] is a popular unsupervised learning method for deriving a low-dimensional set of features from a large set of variables. The principal component regression (PCR) [30] involves constructing the first M principle components $Z_1, \ldots Z_M$ using PCA to reduce dimension from p to M, and then using these M components as the predictors in a linear regression model to fit optimal least square.

Partial least square regression (PLSR) [31] is a dimension reduction method like PCR. However, PLSR identifies the new features $Z_1, \ldots, Z_M$ in a supervised way. It uses response Y to identify new features that approximate the old features well and related to the response. In PLSR computing the first direction $Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$, PLSR places the highest weight on the variables that are most strongly related to the response. The residuals from regressing each variable on Z_1 are the remaining information that has not been explained by the first PLSR direction. And then, Z_2 was computed using the orthogonalized data in the same fashion of Z_1 based on the original data. $Z_1, \ldots, Z_M$ can be computed after M times repeating. The last step is to use optimal least squares to fit a linear model to predict Y using the new features $Z_1, \ldots, Z_M$.

7) LEAVE-ONE-OUT CROSS VALIDATION

Leave-one-out cross-validation [32] is K-fold cross-validation taken to its logical extreme, with K equal to N (the number of observations). It means that N separate times, the function approximator is trained on all the data except for one point, and a prediction is made for that point. Then the average error is computed and used to evaluate the model. The evaluation given by leave-one-out cross-validation error is great for a small sample dataset. Still, it is not friendly for a large sample dataset because it seems very expensive to compute.

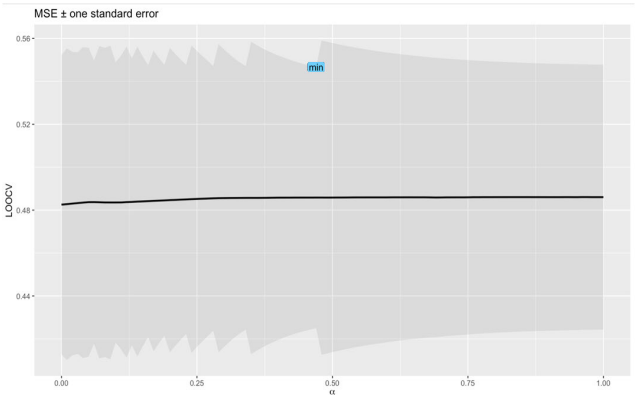


FIGURE 6. LOOCV prediction error ± one standard error of elastic net α from 0 to 1 by 0.01.

III. RESULTS AND DISCUSSION

This section illustrated the results of Fig. 4. Mahalanobis distance result identifies outlier. The results of three linear model selection and regularization methods are detail explained with plots. In the end, we compare ten candidate models (OLS regression, Ridge regression, Lasso regression, Elastic net, best subset selection, forward subset selection, backward subset selection, PCR, PLS) and choose the best one based on leave-one-out-cross-validation (LOOCV) prediction error.

A. OUTLIER DETECTION BY MAHALANOBIS DISTANCE RESULT

For all 97 observations, the Mahalanobis distance D_i^2, i from 1 to 97, can be calculated as (5) by inputting mean $\bar{x}$ and covariance matrix S. As the S is rank 9 covariance matrix, D_i^2 follows $\chi^2_{df=9}$. Set critical p-value αequals to 0.05 such as $\chi^2_{df=9}(1-\alpha) \approx 16.919$, the outlier regions are $[16.919, \infty)$. Fig. 5 shows all the 97 observations' Mahalanobis distance on x-lab. The outlier region is filled in light red on the right tail area. The D_{94}^2, D_{57}^2,D_{47}^2, and D_{69}^2 Mahalanobis distance are greater than 16.989, so it is 95% confident that the observations (index: 94, 57, 47, 69) are outliers.

B. SHRINKAGE (REGULARIZATION) METHOD RESULT

Fig. 7 shows ridge, lasso, and elastic net with α equals 0.47 coefficient paths and their leave one out cross-validation mean squared error across the λ values. From the coefficient path plots, the penalty parameters are controlled by the tuning parameter λ. The coefficients equal OLS regression

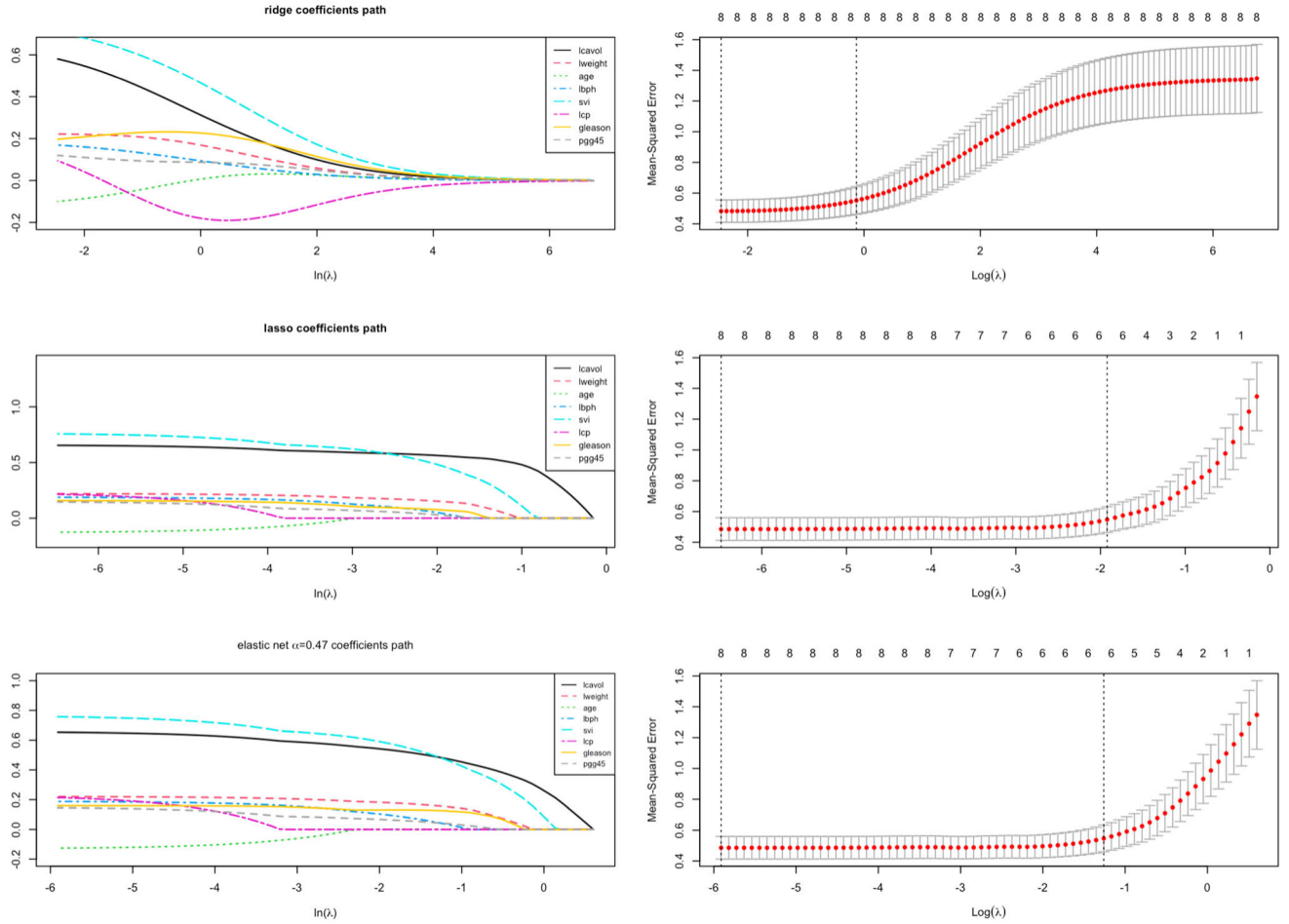


FIGURE 7. Coefficients path and grid LOOCV prediction error plot of lasso, ridge, and elastic net where $\alpha = 0.47$.

coefficients when $\lambda = 0$ for the coefficient paths plots because there is no effect and the objective functions equals the normal OLS regression objective function of simply minimizing SSE. As $\lambda \to \infty$, the penalty parameter becomes large and force coefficients of lasso and elastic net to exact zero but coefficients of the ridge to approximately but not equal to zero. The coefficients path plots illustrate the detail of how the largest λ values have pushed these coefficients to nearly 0.

In LOOCV MSE plots across the λ values for ridge and lasso, the plots show the MSE rise considerably when λ cross over the second vertical dashed lines for ridge and lasso. At the top of each LOOCV MSE plot, the numbers represent the number of variables in the model. As ridge regression does not force any variables to exact zero, all variables will remain in the model, and the top numbers are 8 across all the λ values. The upper and lower bar around the MSE results for each λ denote the MSE plus/minus its standard error. The first and second vertical dashed lines refer to the λ value with the minimum MSE and largest λvalue within one standard error minimum MSE. Adding one standard error to the minimum MSE value can get a more regularized model, a largest λvalue within one standard error minimum MSE is used to evaluate model performance.

TABLE 2. LOOCV prediction error for lasso, ridge, and elastic net where $\alpha = 0.47$.

		λ	PMSE	SE	Nonzero
Ridge	Min	0.0857	0.4826	0.07328	8
	1se	0.8773	0.5522	0.08996	8
Lasso	Min	0.00153	0.4860	0.07375	8
	1se	0.14634	0.5478	0.08347	6
Elastic net $\alpha = 0.47$	Min	0.00271	0.4858	0.07374	8
	1se	0.28370	0.5467	0.08671	6

The elastic net penalty has two tuning parameters: λ for the complexity and α for the compromise between LASSO and ridge. Fig. 6 is LOOCV prediction error $\pm$ one standard error of Elastic net α from 0 to 1 by 0.01. To find the best tuning parameter α, a tuning grid that searches across a range from 0 to 1 by 0.01 is created. Then, iterate over each α value and extract the minimum and one standard error MSE values and their respective λ. The blue label point shows that the

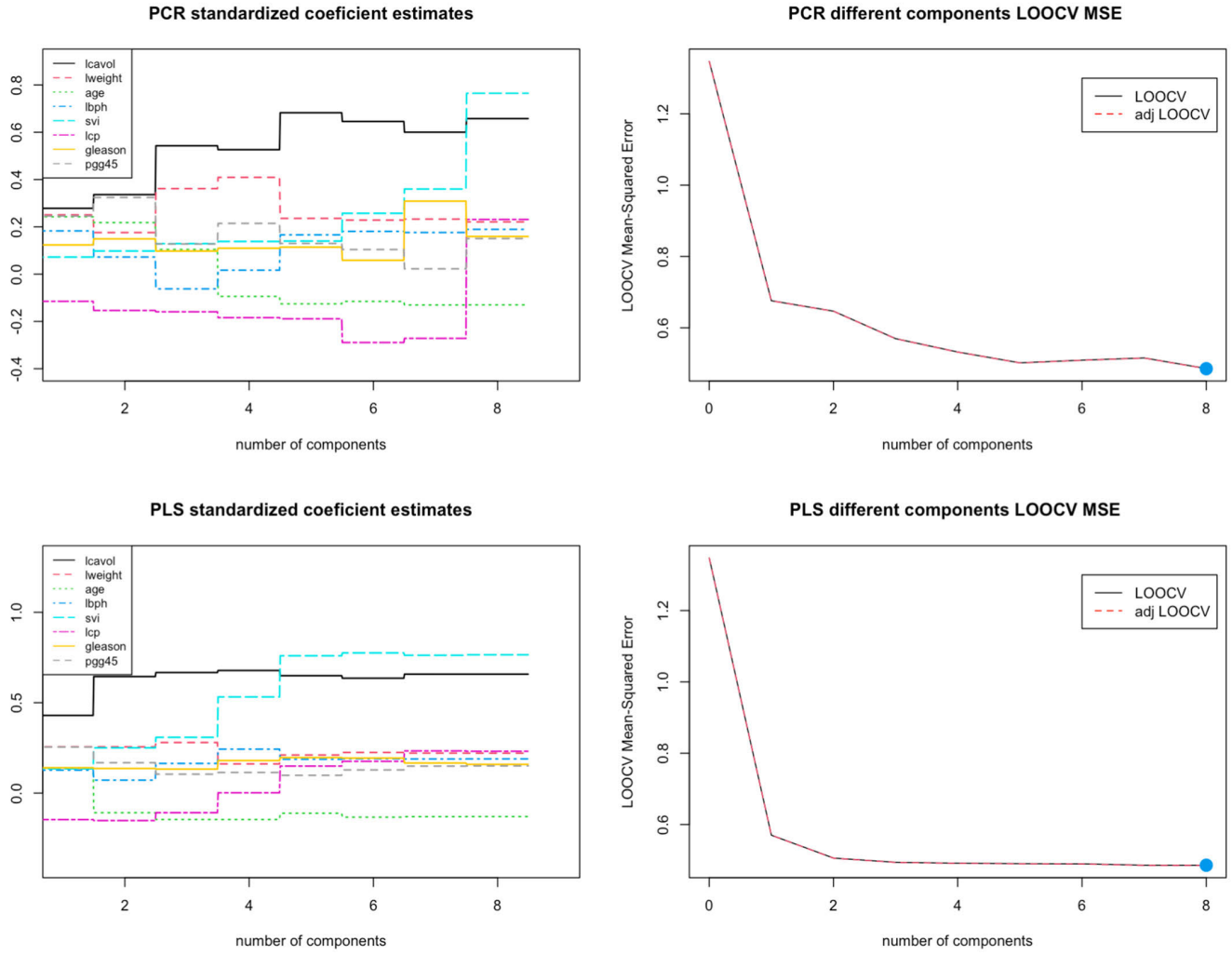


FIGURE 8. Standardized coefficients path and LOOCV and adjusted LOOCV prediction error for different components using PLS and PCR methods.

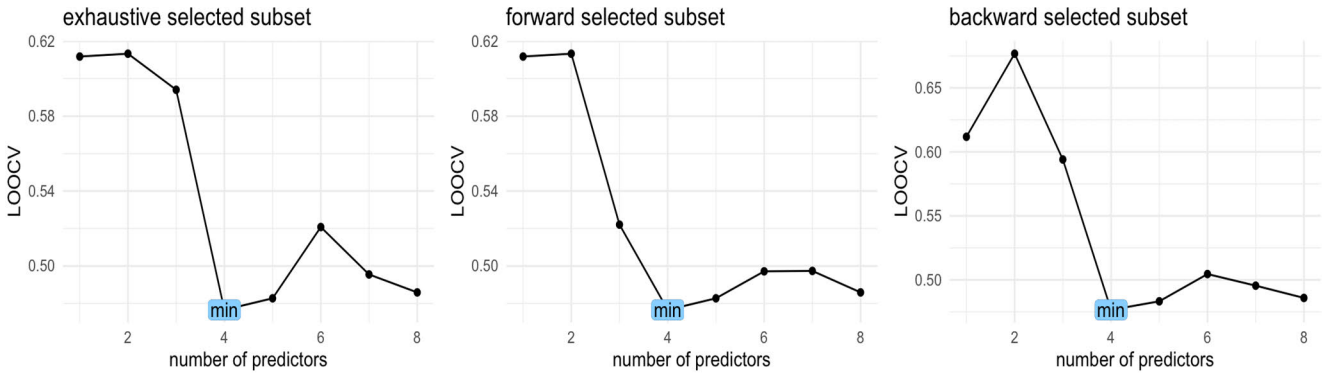


FIGURE 9. LOOCV prediction error with a different number of predictors for exhaustive, forward, and backward selection.

largest λvalue within one standard error minimum MSE for all α from 0 to 1 by 0.01 has minimum MSE when α equals to 0.47.

Table 2 shows the detail of λ values with the minimum MSE and largest λ value within one standard error of the minimum MSE, LOOCV-MSE, standard error, and some variables shrinkage toward zero. Elastic net with α equals 0.47 > Lasso > Ridge based on minimum 1se PMSE compare. Besides, Lasso and Elastic net with α equals 0.47 models eliminated lcp and age variables.

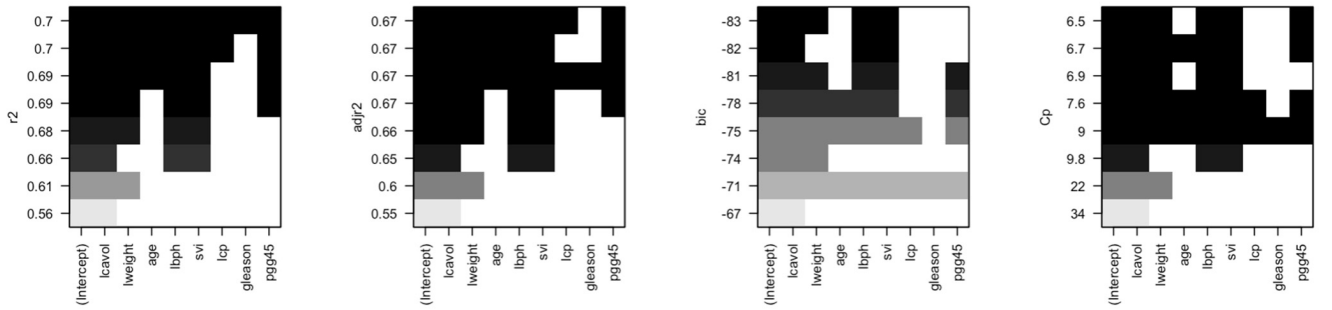


FIGURE 10. Exhaustive subset selection plots for methods of R^2 adjusted R^2, BIC, and C_p.

TABLE 3. Exhaustive subset selection result for 94 data points excluded four outliers.

One	lcavol
Two	lcavol, lweight
Three	lcavol, lbph, svi
Four	lcavol, lweight, svi, lbph
Five	lcavol, lweight, lbph, svi, pgg45
Six	lcavol, lweight, age, lbph, svi, pgg45
Seven	lcavol, lweight, svi, age, lbph, pgg45, lcp
Eight	lcavol, lweight, svi, gleason, age, lbph, pgg45, lcp

TABLE 4. Forward subset selection result for 94 data points excluded four outliers.

One	lcavol
Two	lcavol, lweight
Three	lcavol, lweight, svi
Four	lcavol, lweight, svi, lbph
Five	lcavol, lweight, svi, lbph, pgg45
Six	lcavol, lweight, svi, lbph, pgg45, age
Seven	lcavol, lweight, svi, lbph, pgg45, age, lcp
Eight	lcavol, lweight, svi, lbph, pgg45, age, lcp, gleason

TABLE 5. Backward subset selection result for 94 data points excluded four outliers.

One	lcavol
Two	lcavol, svi
Three	lcavol, svi, lbph
Four	lcavol, svi, lbph, lweight
Five	lcavol, svi, lbph, lweight, pgg45
Six	lcavol, svi, lbph, lweight, pgg45, age
Seven	lcavol, svi, lbph, lweight, pgg45, age, lcp
Eight	Lcavol, svi, lbph, lweight, pgg45, age, lcp, gleason

C. DIMENSION REDUCTION METHOD RESULT

Fig. 8 shows standardized coefficient estimates for different components and leave-one-out cross-validation MSE on preprocess Prostate data set using PCR and PLSR. The dimension reduction method (PCR and PLSR) can make feature selection but cannot select the variables out; the coefficients are toward zero but not equal to zero when the number of components decreases to one.

To find the optimal tuning parameter M (number of components), ordinary CV estimate, and adjusted CV (bias-corrected CV estimate) are applied to find the minimum MSE. From the figures, it shows there is virtually no difference for LOOCV and adjusted LOOCV. The blue dots are the location of the minimum LOOCV MSE for PLSR and PCR models. Eight components minimize LOOCV MSE for PCR and PLSR methods to predict log of prostate-specific antigen. LOOCV MSE of PCR model with eight components (0.4859) is larger than that of seven components (0.5155) and six components (0.5094), and the second lowest LOOCV MSE is PCR model with five components (0.50197). LOOCV MSE of PLSR model with eight components (0.4859) is almost equal to that of seven (0.4860), six (0.4897), five (0.4904), four (0.4914), and three (0.4940). As mentioned in the dimension reduction methods introduction, the main practical difference between PCR and PLSR is that PCR often needs more components than PLSR to achieve the same prediction error. In this literature, LOOCV MSE of PLSR model with two variables (0.5061) is smaller than LOOCV MSE of PCR model with the variables that is less than eight.

Therefore, an optimal tuning parameter for PCR linear is eight-components that did not reduce the dimension. It is the same as OLS regression, but the independent variables did orthogonal transformation. It is hard to determine an optimal tuning parameter (number of components) for PLSR model. PLSR model with eight-components has minimum LOOCV MSE. However, choosing seven, six, or five components are reasonable as well.

D. SUBSET SELECTION METHOD RESULT

Use all 94 excluded 4 outlier data to build the regression subset model; the selected subsets are not the same for exhaustive, backward, and forward subset selection methods shown in Table 3, 4, and 5, but the selected four variables are the same in these three subset selection methods. They contain lcavol, lweight, svi, and lbph variables.

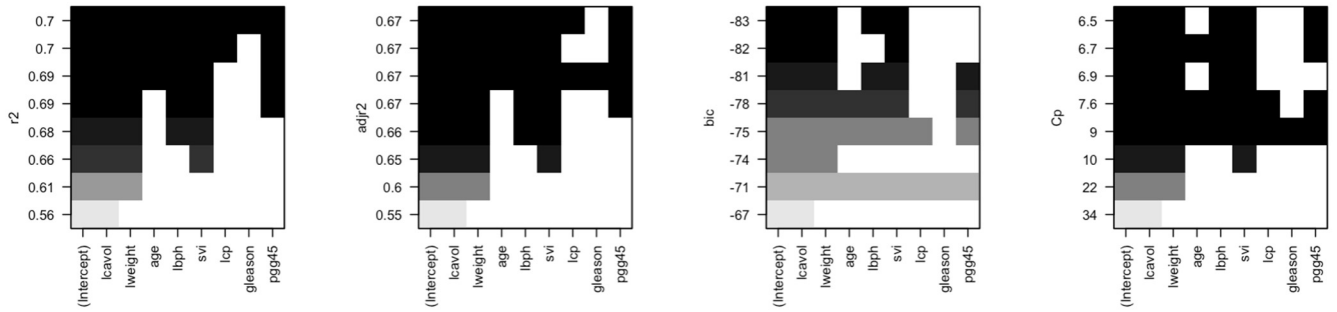


FIGURE 11. Forward subset selection plots for methods of R^2, adjusted R^2, BIC, and C_p.

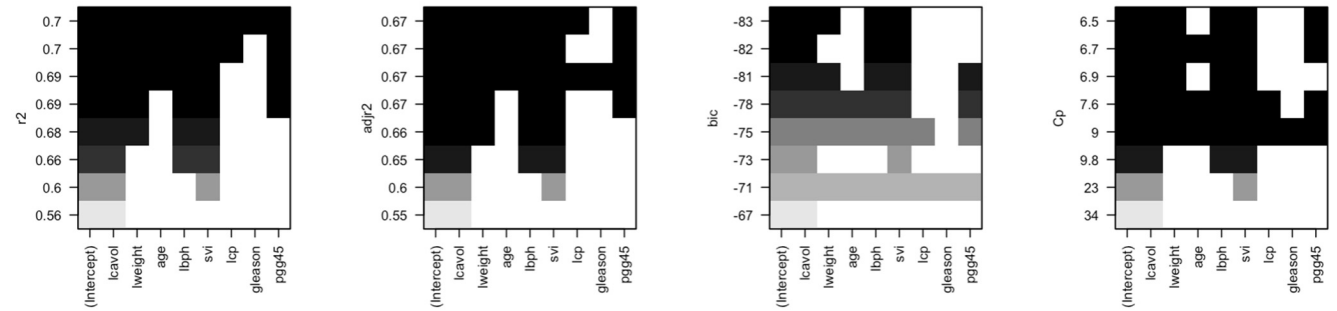


FIGURE 12. Backward subset selection plots for methods of R^2, adjusted R^2, BIC, and C_p.

TABLE 6. Excluded variables result for 8 variable selection methods.

	Subset selection method						Shrinkage method	
				min LOOCV MSE				
Exclude variables	adj R^2	BIC	Cp	exhaustive	forward	backward	Elastic net 1st error LOOCV MSE	Lasso 1st error LOOCV MSE
	gleason	pgg45, age, lcp, gleason	age, lcp, gleason	pgg45, age, lcp, gleason			lcp, age	lcp, age

Fig. 9 shows the LOOCV MSE for different predictors among the exhaustive, forward, and backward selected subset. If the selected variables in Table 3, 4, and 5 are not same, they will have different LOOCV MSE. The number of predictors with the lowest LOOCV MSE (0.47656) is equal to four among these three subset selection methods. The selected variables are lcavol, lweight, svi, and lbph.

Fig. 10, 11, and 12 show the R^2, adjusted R^2, BIC, and Cp statistics of regression subset model for a different number of predictors among exhaustive, forward, and backward subset selection methods.R^2 statistic increases from 0.56 when only one variable is included in the model to 0.70 when all variables are included. As expected, the R^2 statistic increases monotonically as more variables are included, so R^2 statistic cannot evaluate the machine learning model's performance. The number of predictors in subsets is a tuning parameter. The best selected model for these three selection methods by BIC, Cp and adjusted R^2 are the same. The best performer using BIC method is four variables including in the linear regression model that are lcavol, lweight, lbph, and svi. Adjusted R^2 method suggests seven variables that only exclude gleason variable. Cp method selects five variables out that are lcavol,lweight, lbph, svi, and pgg45.

E. DISCUSSION

Table 6 shows eight feature selection results by directly or indirectly calculating prediction error to eliminate variables

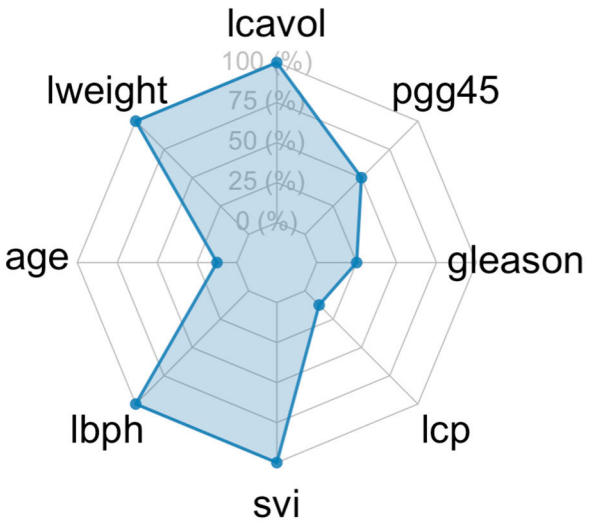


FIGURE 13. Radar plot of variable selection results of by subset selection methods (adjusted R^2, C_p, *BIC*, LOOCV prediction error) and shrinkage methods (elastic net, and lasso LOOCV prediction error).

based on the above experimental results, and Fig. 13 visualizes the variable selection results. The variables of lweight, lcaval, lbph, and svi are included all the times, and the variable of lcp and age are included only once by adjusted R^2 method. Therefore, prostate weight, cancer volume, amount of benign prostatic hyperplasia, and whether seminal vesicle invasion are necessary variables for prostate-specific antigen prediction. Age and capsular penetration are not important variables.

IV. CONCLUSION

Although the sample size is small in this experimental study, the data is collected precious and accurate. The Prostate specimens were subjected to detailed histological and morphometric analysis [15], and all 97 observations are effective without obvious outliers and extreme points. Besides, leave-one-out cross-validation is only efficient using a small sample size of data by standard computers, which can help reviewers and readers with different versions and R-language environments to reproduce the same figure and table results.

The paper discusses some important considerations for feature selection with a detailed R code. Besides, it proves prostate weight, cancer volume, amount of benign prostatic hyperplasia, and whether seminal vesicle invasion are necessary variables that must include predicting PSA. Age and capsular penetration are least important variables. The variables of Gleason score, a percent of Gleason scores 4 or 5 are important sometimes. Lastly, the less developed and poor areas are hard to apply PSA testing, so PSA prediction by statistical models is meaningful for them to detect prostate cancer early.

APPENDIX
ALGORITHMS FOR SUBSET SELECTION

A. EXHAUSTIVE

Algorithm 1 Best Subset Selection

1) Let M_0 represents the null model, which contains no predictors. The model predicts the sample mean for each observation.
2) For $k = 1, \ldots, p$
   (a) For all $\binom{p}{k}$ models that contain k predictors
   (b) Pick the best among these $\binom{p}{k}$ models as M_k,based on the smallest RSS or largest R^2.
3) Select the best model among $M_0, \ldots, M_p$ by Adjusted R^2, C_p, *BIC* or cross-validation prediction error.

B. FORWARD

Algorithm 2 Forward Stepwise Selection

1) Let M_0 represents the null model, which contains no predictors.
2) For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in M_k with one additional predictor.
   (b) Choose the best among these $p-k$ models and call it M_{k+1}.
3) Select the best model among $M_0, \ldots, M_p$ by Adjusted R^2, C_p, *BIC* or cross-validation prediction error.

C. BACKWARD

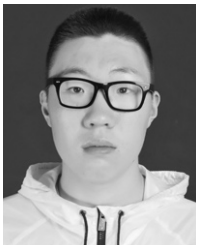
Algorithm 3 Backward Stepwise Selection

1) Let M_p represents the full model, which contains all p predictors.
2) For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in M_k, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it M_{k-1}.
3) Select the best model among $M_0, \ldots, M_p$ by Adjusted R^2, C_p, *BIC* or cross-validation prediction error.

REFERENCES

[1] J. Mohler, R. R. Bahnson, B. Boston, J. E. Busby, A. D'Amico, J. A. Eastham, C. A. Enke, D. George, E. M. Horwitz, R. P. Huben, and P. Kantoff, "Prostate cancer," *J. Nat. Comprehensive Cancer Netw.*, vol. 8, no. 2, pp. 162–200, 2010.
[2] P. D. Baade, D. R. Youlden, S. M. Cramb, J. Dunn, and R. A. Gardiner, "Epidemiology of prostate cancer in the Asia–Pacific region," *Prostate Int.*, vol. 1, no. 2, pp. 47–58, Jun. 2013.

[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018.
[4] J. Constantinou and M. R. Feneley, "PSA testing: An evolving relationship with prostate cancer screening," *Prostate Cancer Prostatic Diseases*, vol. 9, no. 1, pp. 6–13, Mar. 2006.
[5] I. M. Thompson and D. P. Ankerst, "Prostate-specific antigen in the early detection of prostate cancer," *CMAJ*, vol. 176, no. 13, pp. 1853–1858, 2007.
[6] H. Lilja, D. Ulmert, and A. J. Vickers, "Prostate-specific antigen and prostate cancer: Prediction, detection and monitoring," *Nature Rev. Cancer*, vol. 8, no. 4, pp. 268–278, Apr. 2008.
[7] K. Lyzun and A. McMullen, "Prostate man' the ageing superhero: A unique approach to encouraging prostate health awareness among men over 50," *J. Commun. Healthcare*, vol. 2, no. 1, pp. 7–19, 2009.
[8] A. Cassels, "Health sector reform: Key issues in less developed countries," *J. Int. Develop.*, vol. 7, no. 3, pp. 329–347, May 1995.
[9] C.-H. Wu, K. Fang, and T.-C. Chen, "Applying data mining for prostate cancer," in *Proc. Int. Conf. New Trends Inf. Service Sci.*, Jun. 2009, pp. 1063–1065.
[10] N. A. R. Perez, E. G. Vargas, and O. M. F. Cuellar, "Supervised classifiers of prostate cancer from magnetic resonance images in T2 sequences," in *Proc. 14th Iberian Conf. Inf. Syst. Technol. (CISTI)*, Jun. 2019, pp. 1–4.
[11] J. Jung, H. Hong, H. Lee, S. I. Hwang, and H. J. Lee, "Uni- and multimodal radiomic features for the predicting prostate cancer aggressiveness," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1–4.
[12] Y. Wang, B. Zheng, D. Gao, and J. Wang, "Fully convolutional neural networks for prostate cancer detection using multi-parametric magnetic resonance images: An initial investigation," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3814–3819.
[13] R. Sammouda, H. Aboalsamh, and F. Saeed, "Comparison between K mean and fuzzy C-mean methods for segmentation of near infrared fluorescent image for diagnosing prostate cancer," in *Proc. Int. Conf. Comput. Vis. Image Anal. Appl.*, Jan. 2015, pp. 1–6.
[14] J. B. Janney, J. J. Christilda, S. S. Mary, and D. Haritha, "Early diagnosis of prostate cancer using image processing techniques," in *Proc. IEEE Int. Conf. Power, Control, Signals Instrum. Eng. (ICPCSI)*, Sep. 2017, pp. 2868–2871.
[15] T. A. Stamey, J. N. Kabalin, J. E. Mcneal, I. M. Johnstone, F. Freiha, E. A. Redwine, and N. Yang, "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. II. Radical prostatectomy treated patients," *J. Urol.*, vol. 141, no. 5, pp. 1076–1083, May 1989.
[16] W. J. Catalona, A. W. Partin, J. A. Finlay, D. W. Chan, H. G. Rittenhouse, R. L. Wolfert, and D. L. Woodrum, "Use of percentage of free prostate-specific antigen to identify men at high risk of prostate cancer when psa levels are 2.51 to 4 ng/mL and digital rectal examination is not suspicious for prostate cancer: An alternative model," *Urology*, vol. 54, no. 2, pp. 220–224, Aug. 1999.
[17] J. H. Fowke, L. B. Signorello, S. S. Chang, C. E. Matthews, M. S. Buchowski, M. S. Cookson, F. M. Ukoli, and W. J. Blot, "Effects of obesity and height on prostate-specific antigen (PSA) and percentage of free PSA levels among African-American and Caucasian men," *Cancer*, vol. 107, no. 10, pp. 2361–2367, Nov. 2006.
[18] R. B. Nadler, P. A. Humphrey, D. S. Smith, W. J. Catalona, and T. L. Ratliff, "Effect of inflammation and benign prostatic hyperplasia on elevated serum prostate specific antigen levels," *J. Urol.*, pp. 407–413, Aug. 1995.
[19] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The Mahalanobis distance," *Chemometrics Intell. Lab. Syst.*, vol. 50, no. 1, pp. 1–18, 2000.
[20] K. I. Penny, "Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance," *J. Roy. Stat. Soc. Ser. C, Appl. Statist.*, vol. 45, no. 1, pp. 73–81, 1996.
[21] C. Park, "A note on the chi-square test for multivariate normality based on the sample Mahalanobis distances," *J. Korean Stat. Soc.*, vol. 28, no. 4, pp. 479–488, 1999.
[22] C. L. Mallows, "Some comments on CP," *Technometrics*, vol. 42, no. 1, pp. 87–94, 2000.
[23] K. P. Burnham and D. R. Anderson, "Multimodel inference: Understanding AIC and BIC in model selection," *Sociol. Methods Res.*, vol. 33, no. 2, pp. 261–304, Nov. 2004.
[24] J. Miles, "R squared, adjusted R squared," in*Wiley StatsRef: Statistics Reference Online*. Hoboken, NJ, USA: Wiley, 2014.
[25] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970.
[26] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc. B, Methodol.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.
[27] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005.
[28] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro, "Dimension reduction and coefficient estimation in multivariate linear regression," *J. Roy. Stat. Soc. B, Stat. Methodol.*, vol. 69, no. 3, pp. 329–346, Jun. 2007.
[29] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.
[30] I. T. Jolliffe, "A note on the use of principal components in regression," *J. Roy. Stat. Soc. C, Appl. Statist.*, vol. 31, no. 3, pp. 300–303, 1982.
[31] P. Geladi and B. R. Kowalski, "Partial least-squares regression: A tutorial," *Analytica Chim. Acta*, vol. 185, pp. 1–17, Jun. 1986.
[32] N. K. Chlis, "Comparison of statistical methods for genomic signature extraction," Tech. Univ. Crete, Chania, Greece, Tech. Rep., 2013.



GONGLI LI was born in Zhengzhou, Henan, China, in 1997. He received the bachelor's degree in statistics (data analysis major and mathematics minor) from The Australian National University, in June 2021, where he is currently pursuing the master's degree in statistics.

HAN LI was born in Gongyi, Zhengzhou, Henan, China, in 1998. She received the bachelor's degree (Hons.) in accounting from Victoria University and Henan University, in 2020. She is currently pursuing the master's degree in management finance with the University of Melbourne.

Since 2020, she has been a Research Assistant with the Board of Directors Office, Strategic Research Department, Lian Chu Securities Company Ltd. Her research interests include asset pricing, business analysis, corporate finance, financial engineering, and data analysis.

Mrs. Li was a recipient of the Golden Key International Honor Society from Victoria University, in 2020.