Prostrate_example

Jack Tubbs

November 2021

Contents

1 Introduction

In this document, I have reproduced the discussion material found in the lecture notes for the topic, "Model Selection Methods". My intent is to illustrate some of the methods with a single data set found in Hastie et. al. book entitled "The Elements of Statistical Learning" using the prostrate cancer data and example as discussed in their book.

Model Selection is a very rich topic in regression and has become a widely use method when doing machine learning and data science. In this document I have elected just to illustrate a few ways of using this theory and corresponding methods.

I have restricted my software choice to using SAS. It should be noted that Hastie's book has R code with illustrative example.

2 Model Selection

The methods presented in this chapter are used to determine appropriate subset models in the multiple regression problem. Various statistics can be used, including; R^2, the adjusted R^2, s^2 the residual mean square (s^2), and Mallow's C_p statistic.

Mallow's statistics – C_p

The statistics is

$$C_p = RSS_p/s^2 - (n - 2p),$$

where RSS_p is the residual sum of squares from the model containing $p =$ rank(X) parameters (including β_0), and $s^2 = MS_E(r)$ is the residual mean square from the full model containing r predictor variables [assumed to be the most reliable estimate of σ^2]. Note, when $p = r + 1$, $C_p = p$. The idea of using this statistics is to find the smallest value of p such that $C_p \approx p$.

3 Subset Selection

Statistical selection procedures include;

Forward Selection

The forward procedure starts with a single variable model (often selected with the highest R^2) and then adds additional variables that satisfy an entry criteria. The process continues until no other variables satisfy the entry criteria.

Backward Selection

The backward procedure starts with the complete (full) model and then eliminates variables that satisfy an exit criteria. The procedure is as follows;

1. Compute the regression model using all the predictor variables.
2. The partial F-value is calculated for every predictor variable using the type II sum of squares.
3. The lowest partial F-value, F_L, is compared with a preselected significance level, F_0.
   (a) If $F_L < F_0$, remove the variable corresponding to F_L, say X_L then recompute the model using the reduced model.
   (b) If $F_L > F_0$, adopt the regression model as calculated.

Stepwise Selection

Stepwise is a forward selection method that selects the best single predictor (say, the variable with the largest R^2), X_1 and fit the equation $\hat{y} = f(X_1)$. If this model is not significant, stop and conclude that $\hat{y} = \bar{y}$. If the model is significant, select the next predictor variable, say X_2 based upon the one with the largest partial F-value and the equation is given by $\hat{y} = f(X_1, X_2)$. This model checked for improvement in the R^2 and partial F-values for both variables in the equation. (This differs from the forward procedure in that a first variable may be excluded from the model at this step whereas in forward selection once a variable enters the model it remains). These partial F-values are used to determine whether or not a variable remains in the model or is excluded. The procedure continues until no new variables satisfy the entry criteria.

3.1 SAS Model-Selection Methods

The PROC REG selection methods include;

- **Full Model Fitted (NONE)**
- **Forward Selection (FORWARD)**

  The forward-selection technique begins with no variables in the model. For each of the independent variables, the FORWARD method calculates F statistics that reflect the variable's contribution to the model if it is included. The p-values for these F statistics are compared to the SLENTRY= value that is specified in the MODEL statement. Once a variable is in the model, it stays.

- **Backward Elimination (BACKWARD)**

  The backward elimination technique begins by calculating F statistics for a model, including all of the independent variables. Then the variables are deleted from the model one by one until all the variables remaining in the model produce F statistics significant at the SLSTAY= level specified in the MODEL statement.

- **Stepwise (STEPWISE)**

  The stepwise method is a modification of the forward-selection technique and differs in that variables already in the model do not necessarily stay there. As in the forward-selection method, variables are added one by one to the model, and the F statistic for a variable to be added must be significant at the SLENTRY= level. After a variable is added, however, the stepwise method looks at all the variables already included in the model and deletes any variable that does not produce an F statistic significant at the SLSTAY= level.

- **Maximum R^2 Improvement (MAXR)**

  The maximum R^2 improvement technique does not settle on a single model. Instead, it tries to find the "best" one-variable model, the "best" two-variable model, and so forth, although it is not guaranteed to find the model with the largest R^2 for each size.

- **Minimum R^2 (MINR) Improvement**

  The MINR method closely resembles the MAXR method, but the switch chosen is the one that produces the smallest increase in R^2. For a given number of variables in the model, the MAXR and MINR methods usually produce the same "best" model, but the MINR method considers more models of each size.

- **R^2 Selection (RSQUARE)**

  The RSQUARE method finds subsets of independent variables that best predict a dependent variable by linear regression in the given sample.

- **Adjusted R^2 Selection (ADJRSQ)**

  This method is similar to the RSQUARE method, except that the adjusted R^2 statistic is used as the criterion for selecting models, and the method finds the models with the highest adjusted R^2 within the range of sizes.

- **Mallows' C_p Selection (CP)**

  This method is similar to the ADJRSQ method, except that Mallows' C_p statistic is used as the criterion for model selection. Models are listed in ascending order of C_p.

The PROC GLMSELECT selection procedures include;

- **Least Angle Regression (LAR)** Least angle regression was introduced by Efron et al. (2004). As in the forward selection method, the LAR algorithm produces a sequence of regression models where one parameter is added at each step, terminating at the full least squares solution when all parameters have entered the model.

  The algorithm starts by centering the covariates and response (independent and dependent variables), and scaling the covariates (independent variables) so that they all have the same corrected sum of squares. Initially all coefficients are zero, as is the predicted response. The predictor that is most correlated with the current residual is determined and a step is taken in the direction of this predictor. The length of this step determines the coefficient of this predictor and is chosen so that some other predictor and the current predicted response have the same correlation with the current residual. At this point, the predicted response moves in the direction that is equiangular between these two predictors. Moving in this direction ensures that these two predictors continue to have a common correlation with the current residual. The predicted response moves in this direction until a third predictor has the same correlation with the current residual as the two predictors already in the model.

A new direction is determined that is equiangular between these three predictors and the predicted response moves in this direction until a fourth predictor joins the set having the same correlation with the current residual. This process continues until all predictors are in the model.

- **Lasso Selection (LASSO)** LASSO (least absolute shrinkage and selection operator) selection uses constrained ordinary least squares regression where the sum of the absolute values of the regression coefficients is constrained to be smaller than a specified parameter. More precisely let $X = (x_1, x_2, \ldots, x_m)$ denote the matrix of covariates and let y denote the response, where the x_i has been centered and scaled, and y has been centered (as in the LARS procedure). Let the parameter t be specified then the LASSO regression coefficients $\beta = (\beta_1, \beta_2, \ldots, \beta_m)$ are the solution to

$$\text{minimize}||y - X\beta||^2 \qquad \text{subject to} \quad \sum_{j=1}^{m} |\beta_j| \leq t$$

  When t is small enough, some of the LASSO regression coefficients will be zero. Hence, LASSO selects a subset of the regression coefficients for each LASSO parameter, t. By increasing the LASSO parameter in discrete steps, you obtain a sequence of regression coefficients where the nonzero coefficients at each step correspond to selected parameters.

  Early implementations (Tibshirani 1996) of LASSO selection used quadratic programming techniques to solve the constrained least squares problem for each LASSO parameter t. Osborne, Presnell, and Turlach (2000) developed a "homotopy method" that generates the LASSO solutions for all values of t. Efron et al. (2004) derived a variant of their algorithm for least angle regression that can be used to obtain a sequence of LASSO solutions from which all other LASSO solutions can be obtained by linear interpolation.

- **Elastic Nets** JMP version 11 uses a new selection called *elastic net*, that is a convex combination of the lasso and ridge regression method. This procedure adjusts for limitations in the lasso method, that include:
  - In the $p > n$ case, the lasso selects at most n variables before it saturates, because of the nature of the convex optimization problem. This seems to be a limiting feature for a variable selection method. Moreover, the lasso is not well defined unless the bound on the L_1 -norm of the coefficients is smaller than a certain value.
  - If there is a group of variables among which the pairwise correlations are very high, then the lasso tends to select only one variable from the group and does not care which one is selected.
  - For usual $n > p$ situations, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression.

Selection Criteria

PROC GLMSELECT supports a variety of fit statistics that you can specify as options in the MODEL statement. The statistics are:

ADJRSQ - adjusted R-square statistic (Darlington 1968; Judge et al. 1985)

AIC - Akaike information criterion (Darlington 1968; Judge et al. 1985)

AICC - corrected Akaike information criterion (Hurvich and Tsai 1989)

BIC - Sawa Bayesian information criterion (Sawa 1978; Judge et al. 1985)

CP - Mallows C_p statistic (Mallows 1973; Hocking 1976)

PRESS - predicted residual sum of squares statistic

SBC - Schwarz Bayesian information criterion (Schwarz 1978; Judge et al. 1985)

SL - significance level of the F statistic used to assess an effectÕs contribution to the fit when it is added to or removed from a model

VALIDATE - average square error over the validation data

PROC GLMSELECT supports a variety of fit statistics that you can specify as options in the MODEL statement. The statistics are:

ADJRSQ - adjusted R-square statistic (Darlington 1968; Judge et al. 1985)

AIC - Akaike information criterion (Darlington 1968; Judge et al. 1985)

AICC - corrected Akaike information criterion (Hurvich and Tsai 1989)

BIC - Sawa Bayesian information criterion (Sawa 1978; Judge et al. 1985)

CP - Mallows C_p statistic (Mallows 1973; Hocking 1976)

PRESS - predicted residual sum of squares statistic

SBC - Schwarz Bayesian information criterion (Schwarz 1978; Judge et al. 1985)

SL - significance level of the F statistic used to assess an effectÕs contribution to the fit when it is added to or removed from a model

VALIDATE - average square error over the validation data

Table 1: Formulas and Definitions for Model Fit Summary Statistics

Statistic	Definition or Formula
n	number of observations
p	number of parameters including the intercept
$\hat{\sigma}^2$	estimate of pure error variance from fitting the full model
SST	total sum of squares corrected for the mean for the dependent variable
SSE	error sum of squares
ASE	$\frac{SSE}{n}$
MSE	$\frac{SSE}{n-p}$
R^2	$1-\frac{SSE}{SST}$
$ADJRSQ$	$1-\frac{(n-1)(1-R^2)}{n-p}$
AIC	$n\ln\left(\frac{SSE}{n}\right)+2p$
$AICC$	$1+\ln\left(\frac{SSE}{n}\right)+\frac{2(p+1)}{n-p-2}$
BIC	$n\ln\left(\frac{SSE}{n}\right)+2(p+2)q-2q^2$ where $q=\frac{n\hat{\sigma}^2}{SSE}$
$CP(C_p)$	$\frac{SSE}{\hat{\sigma}^2}+2p-n$
$PRESS$	$\sum_{i=1}^{n}\frac{r_i^2}{(1-h_i)^2}$ where r_i= residual at observation i and h_i = leverage of observation $i = x_i(X'X)^{-}x_i$
$RMSE$	$\sqrt{MSE}$
SBC	$n\ln\left(\frac{SSE}{n}\right)+p\ln(n)$

3.2 Shrinkage Methods

The following material is given in Seber and Lee in chapter 12. Shrinkage methods were first proposed by James and Stein (1961). Suppose that $Z \sim N_p(\mu, \sigma^2\ I_p)$ when $p > 2$. The obvious estimate of μ is Z, which is the minimum variance unbiased estimate. However, this estimate is unsatisfactory since the expected length of Z, given by $||\ Z\ ||^2$ tends to be too large since,

$$\begin{aligned} E(||\ Z\ ||^2) &= \sum_{i-1}^{p} E[Z_i]^2 \\ &= \sum_{i-1}^{p} E[\sigma^2 + \mu_i^2] \\ &= p\sigma^2 + ||\ \mu\ ||^2 \\ &> ||\ \mu\ ||^2\ . \end{aligned}$$

Thus, some of the elements of the estimate are too large. This suggests "shrinking" some of the elements of Z, and considering an estimate of the form $\tilde{\mu} = cZ$, where $0 < c < 1$.

This estimate is biased, but it is possible to choose c so that $\tilde{\mu}$ has a smaller mean-square error than Z as an estimate of μ. Consider

$$\begin{aligned} E(||\ \tilde{\mu} - \mu\ ||^2) &= \sum_{i=1}^{p} E[(cZ_i - \mu_i)^2] \\ &= \sum_{i=1}^{p} E[(cZ_i - c\mu + c\mu - \mu_i)^2] \\ &= \sum_{i=1}^{p} E[(c(Z_i - \mu_i) - (1-c)\mu_i)^2] \\ &= \sum_{i=1}^{p} [c^2\sigma^2 + (1-c)^2\mu_i^2] \\ &= c^2 p\sigma^2 + (1-c)^2\ ||\ \mu\ ||^2\ . \end{aligned}$$

which is minimized by choosing $c = ||\ \mu\ ||^2\ /(p\sigma^2 + ||\ \mu\ ||^2)$. Thus the optimal estimate can be written as

$$\tilde{\mu} = (1 - \frac{p\sigma^2}{p\sigma^2 + ||\ \mu\ ||^2})Z.$$

However, this estimate is not a practical estimate since it requires knowledge of $||\ \mu\ ||$. Which in turns suggest the estimate

$$\tilde{\mu} = (1 - \frac{p\sigma^2}{||\ Z\ ||^2})Z.$$

One can do better than this as James and Stein showed that of all estimates of the form

$$\tilde{\mu} = (1 - \frac{b}{||\ Z\ ||^2})Z,$$

are best, in the sense of minimum mean-square error. The choice for b then becomes $b = (p-2)\sigma^2$ provided that $p > 2$.

In the regression case with orthonormal predictor variables (i.e., $X'X = I_p$) and known variance σ^2 and $\beta_0 = 0$, one can apply the James-Stein estimate procedure directly by setting $Z = \hat{\beta}$ and $\mu = \beta$ to obtain the "shrinkage" estimator

$$\tilde{\beta} = (1 - \frac{(p-2)\sigma^2}{||\ \hat{\beta}\ ||^2})\hat{\beta}$$

which has the smallest MSE (mean-square error) of any estimator of the form $(1-c)\hat{\beta}$. If σ^2 is unknown then it can be replaced by the usual estimate s^2 of σ^2 where the optimality property no longer holds.

3.3 Ridge Regression

Ridge regression is a popular method for detecting multicollinearity within a regression model. It was first proposed by Hoerl and Kennard (1970) and it was one of the first biased estimation procedures. The idea is fairly simple. Since the matrix $X'X$ is ill-condition or nearly singular one can add positive constants to the diagonal matrix and insure that the resulting matrix is not ill-conditioned. That is, consider the biased normal equations given by

$$(X'X + kI_n)\beta = X'y.$$

With a resulting biased estimate for β given by

$$\tilde{\beta}(k) = (X'X + kI_n)^{-1}X'y,$$

where k is called the shrinkage parameter. Since, $E(\tilde{\beta}) \neq \beta$ some do not want to use such a procedure. However in spite of the fact that $\tilde{\beta}$ is biased, it does have the effect of reducing the variance in the estimator. It can be shown that,

$$var(\hat{\beta}_j) = \sigma^2 1/\lambda_j,$$

where λ_j is the j^{th} eigenvalue of $X'X$. So when $X'X$ is ill-conditioned some of the $\lambda_j's$ are very small, hence $var(\hat{\beta}_j)$ is very large. However,

$$var(\tilde{\beta}_j) = \sigma^2\lambda_j/(\lambda_j + k)^2.$$

Consider the example where $\sigma^2 = 1$, $\lambda_1 = 2.985, \lambda_2 = 0.01$, and $\lambda_3 = 0.005$, the usual least squares estimation gives,

$$\sum_{j=1}^{3} var(\hat{\beta}_j) = \sigma^2 \sum_{i=1}^{3} 1/\lambda_j = .3350 + 100 + 200 = 300.3350.$$

However, if $k = 0.10$ we have,

$$\sum_{j=1}^{3} var(\hat{\beta}_j) = \sigma^2 \sum_{i=1}^{3} \lambda_j/(\lambda_j + k)^2 \approx 2.3.$$

This process of reducing the total variance is very desirable and has led people to proposed similar estimation procedures called shrinkage estimators. In this class, we are interested using this procedure as a way of identifying multicollinearity and the variables which may contribute to this problem.

3.4 LASSO

The LASSO procedure is similar to ridge regression in that the estimator is a "shrinkage estimate" whereby one trades off the unbiased property for an estimate that is much more precise (smaller dispersion or mean square error). That is, the least squares estimate is

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\{\|y - X\beta\|^2\}.$$

Whereas the ridge estimate is

$$\hat{\beta}_{ridge} = \underset{\beta}{\text{argmin}}\{\|y - X\beta\|^2 + \lambda \|\beta\|^2\}$$

where $\|\beta\| = \sum_{i=1}^{p} \beta_i^2$ is the usual l_2 or Euclidean norm. The LASSO estimate is

$$\hat{\beta}_{lasso} = \underset{\beta}{\text{argmin}}\{\|y - X\beta\|^2 + \lambda \|\beta\|\}$$

where $\|\beta\| = \sum_{i=1}^{p} |\beta_i|$ is the l_1 norm.

Hastie, Tibshirani and Friedman book entitled "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" have a very understandable discussion of the above methods. I have reproduced some of their material here.

In this section we discuss and compare the three approaches discussed so far for restricting the linear regression model: subset selection, ridge regression and the lasso. In the case of an orthonormal input matrix X the three procedures have explicit solutions. Each method applies a simple transformation to the least squares estimate $\hat{\beta}_j$. Ridge regression does a "proportional shrinkage". Lasso translates each coefficient by a constant factor λ, truncating at zero. This is called "soft thresholding," and is used in the context of wavelet-based smoothing in Section 5.9. Best-subset selection drops all variables with coefficients smaller than the Mth largest; this is a form of "hard thresholding."

3.5 Elastic Net Selection

The METHOD=ELASTICNET option specifies the elastic net method proposed by Zou and Hastie (2005), which bridges the LASSO method and ridge regression. The elastic net method strikes a balance between having a parsimonious model and borrowing strength from correlated regressors, by solving the least squares regression problem with constraints on both the sum of the absolute coefficients and the sum of the squared coefficients.

More specifically, the elastic net coefficients $\beta = (\beta_1, \beta_2, \ldots, \beta_m)$ are the solution to the constrained optimization problem

$$\min\|\mathbf{y} - \mathbf{X}\beta\|^2 \quad \textit{subject to} \quad \sum_{j=1}^{m} \mid \beta_j \mid \; \leq t_1, \sum_{j=1}^{m} \beta_j^2 \leq t_2$$

This can be written as the equivalent Lagrangian form

$$\min\|\mathbf{y} - \mathbf{X}\beta\|^2 + \lambda_1 \sum_{j=1}^{m} \mid \beta_j \mid + \lambda_2 \sum_{j=1}^{m} \beta_j^2$$

Elastic net can be treated as a convex combination of LASSO and ridge penalty; pure LASSO and pure ridge are two limiting cases. If t_1 is set to a very large value or, equivalently, if λ_1 is set to 0, then the elastic net method reduces to ridge regression. If t_2 is set to a very large value or, equivalently, if λ_2 is set to 0, then the elastic net method reduces to LASSO. If t_1 and t_2 are both large or, equivalently, if λ_1 and λ_2 are both set to 0, then the elastic net method reduces to ordinary least squares regression.

The elastic net method can overcome the limitations of LASSO in the following three scenarios:

- If you have more parameters than observations ($m > n$), the LASSO method selects at most n variables before it saturates, because of the nature of the convex optimization problem. This can be a defect for a variable selection method. By contrast, the elastic net method can select more than n variables in this case because of the ridge regression regularization.

- If there is a group of variables that have high pairwise correlations, then whereas LASSO tends to select only one variable from that group, the elastic net method can select more than one variable.

- If you have more observations than parameters ($n > m$), and there are high correlations between predictors, then it has been empirically observed that the prediction performance of LASSO is dominated by ridge regression. In this case, the elastic net method can achieve better prediction performance by using ridge regression regularization.

3.6 Example: Prostrate Cancer Data

One of the data sets in their book came from a study by Stamey et al. (1989). They examined the correlation between the level of prostate-specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), **age**, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**). The correlation matrix of the predictors given in Table 3.1 (not shown) many strong correlations. Figure 1.1 (not shown)

of Chapter 1 is a scatterplot matrix showing every pairwise plot between the variables. We see that svi is a binary variable, and gleason is an ordered categorical variable. We see, for example, that both lcavol and lcp show a strong relationship with the response lpsa, and with each other. We need to fit the effects jointly to untangle the relationships between the predictors and the response.

We fit a linear model to the log of prostate-specific antigen, lpsa, after first standardizing the predictors to have unit variance[1]. We randomly split the dataset into a training set of size 67 and a test set of size 30. We applied least squares estimation to the training set, producing the estimates, standard errors and Z-scores shown in Table 3.2. The Z-scores measure the effect of dropping that variable from the model. A Z-score greater than 2 in absolute value is approximately significant at the 5% level. (For our example, we have nine parameters, and the 0.025 tail quantiles of the t_{67-9} distribution are ± 2.002) The predictor lcavol shows the strongest effect, with lweight and svi also strong. Notice that lcp is not significant, once lcavol is in the model (when used in a model without lcavol, lcp is strongly significant). We can also test for the exclusion of a number of terms at once, using the F-statistic (3.13). For example, we consider dropping all the non-significant terms in Table 3, namely age, lcp, gleason, and pgg45.

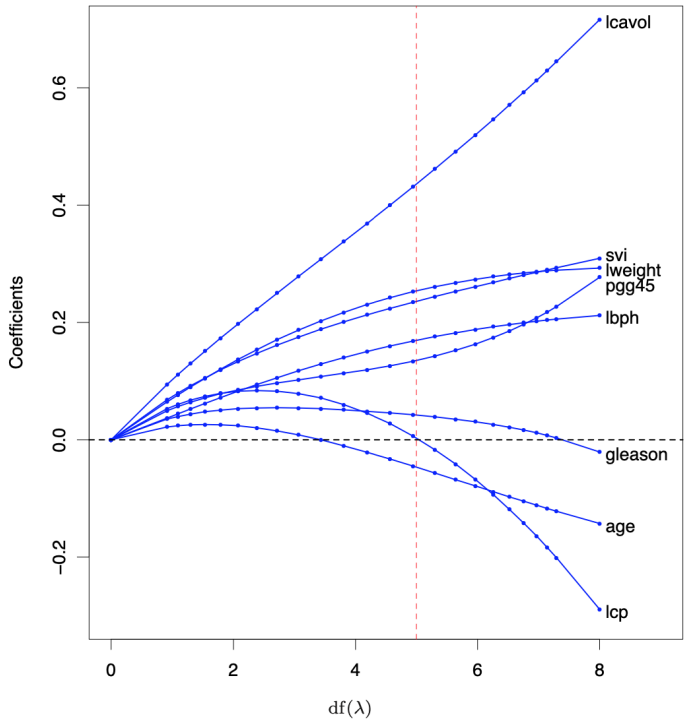


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter* λ *is varied. Coefficients are plotted versus* $\mathrm{df}(\lambda)$*, the effective degrees of freedom. A vertical line is drawn at* $\mathrm{df} = 5.0$*, the value chosen by cross-validation.*

Figure 1: Ridge Plots for Prostrate Model

4 SAS

4.1 Code - Descriptive Statistics

[1]This step is absolutely necessary when using the newer machine learning procedures.

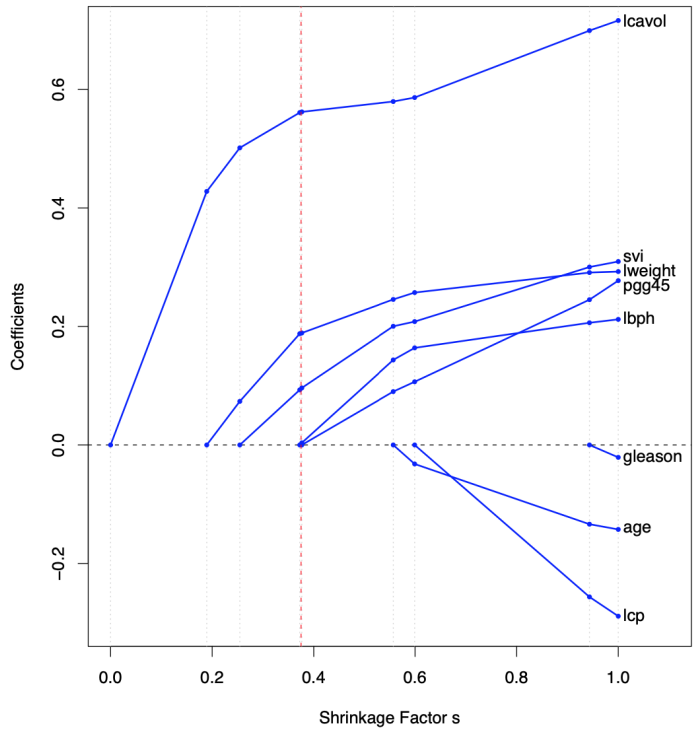


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat{\beta}_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

Figure 2: Lasso Plots for Prostrate Model

```
options center nodate pagesize=80 ls=70;
libname LDATA '/home/jacktubbs/my_shared_file_links/jacktubbs/myfolders/SAS Data Sets/';

/* Simplified LaTeX output that uses plain LaTeX tables  */
ods latex path='/home/jacktubbs/my_shared_file_links/jacktubbs/LaTeX/Class'
 file='prostrate_sel.tex' style=journal
stylesheet="sas.sty"(url="sas");

/*
http://support.sas.com/rnd/base/ods/odsmarkup/latex.html
*/
ods graphics / reset width=5in  outputfmt=png
  antialias=on;
*/;

title1 'HTpostrate Data';
data prostrate; set ldata.HTpostrate;
high_lpsa = (lpsa > 3);        /*your choice for cut off  */
run;

proc means data=prostrate q1 median q3; where train = 'T';
var age gleason lbph lcavol lcp lpsa lweight pgg45;
run;
```

TABLE 3.3. *Estimated coefficients and test error results, for different subset and shrinkage methods applied to the prostate data. The blank entries correspond to variables omitted.*

Term	LS	Best Subset	Ridge	Lasso	PCR	PLS
Intercept	2.465	2.477	2.452	2.468	2.497	2.452
lcavol	0.680	0.740	0.420	0.533	0.543	0.419
lweight	0.263	0.316	0.238	0.169	0.289	0.344
age	−0.141		−0.046		−0.152	−0.026
lbph	0.210		0.162	0.002	0.214	0.220
svi	0.305		0.227	0.094	0.315	0.243
lcp	−0.288		0.000		−0.051	0.079
gleason	−0.021		0.040		0.232	0.011
pgg45	0.267		0.133		−0.056	0.084
Test Error	0.521	0.492	0.492	0.479	0.449	0.528
Std Error	0.179	0.143	0.165	0.164	0.105	0.152

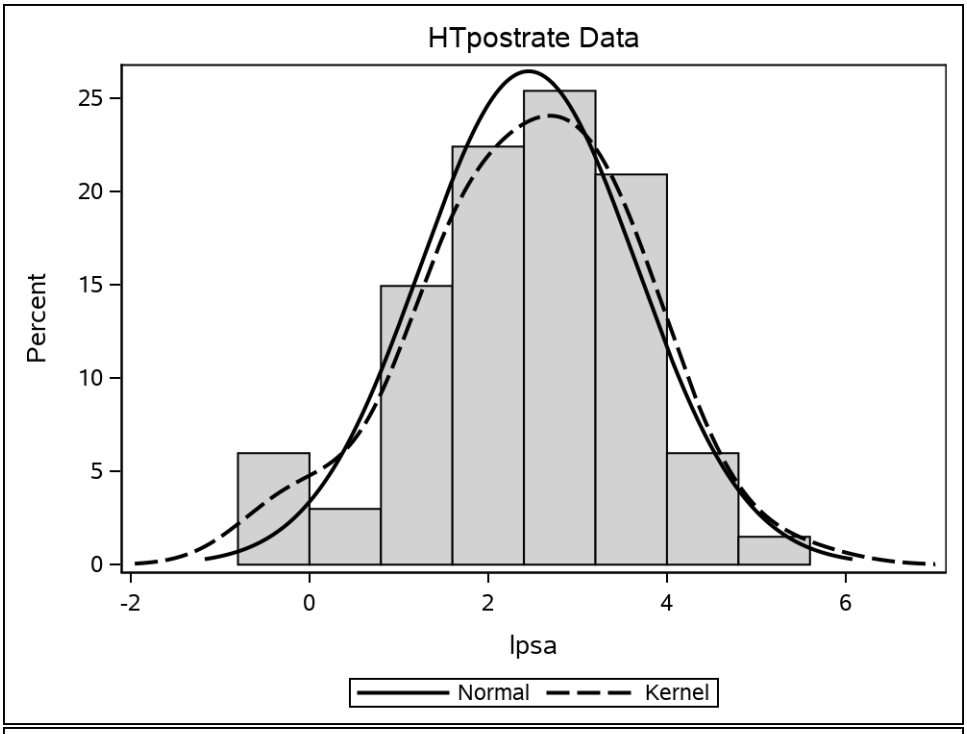
Figure 3: Coefficient Estimates for Prostrate Models

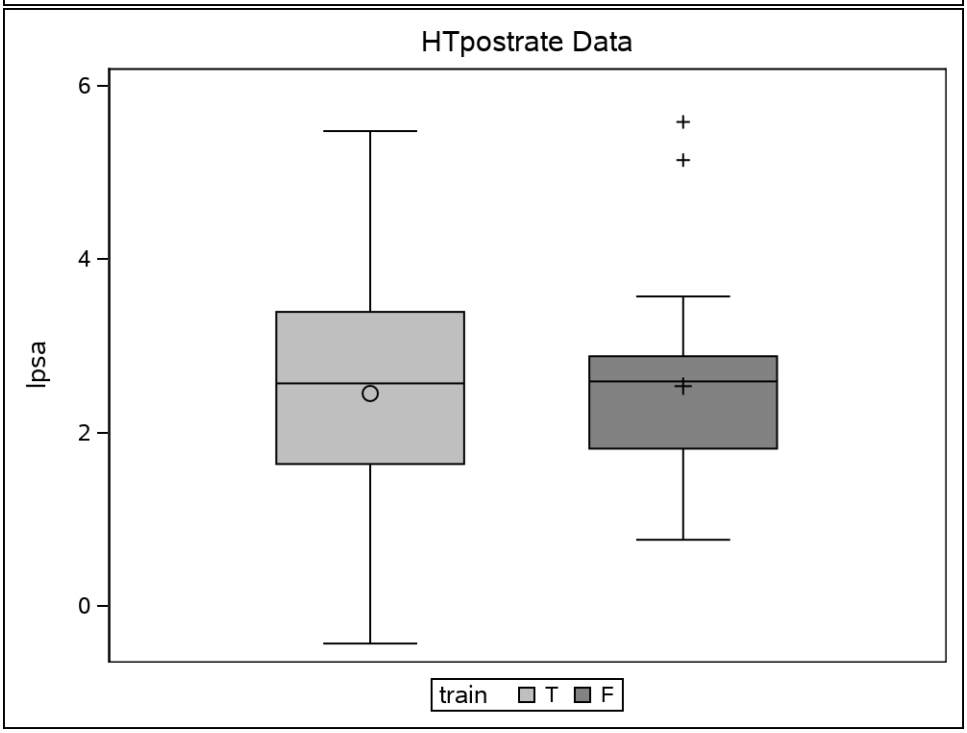
```
proc freq data=prostrate; table train;
run;

proc sgplot data=prostrate; where train = 'T';
histogram lpsa;
density lpsa;
density lpsa/ type= kernel;
run;

proc sgplot data=prostrate;
vbox lpsa/group=train;
run;
```

HTpostrate Data

The FREQ Procedure

train				
train	*Frequency*	*Percent*	*Cumulative Frequency*	*Cumulative Percent*
F	30	30.93	30	30.93
T	67	69.07	97	100.00


HTpostrate Data
Percent
25
20
15
10
5
0
-2
0
2
4
6
lpsa
Normal
Kernel


HTpostrate Data
lpsa
6
4
2
0
train
T
F

4.2 Code 2 - The GLMSELECT Procedure

```
proc glmselect data=prostrate
plot=CriterionPanel; where train = 'T';
model lpsa = age gleason lbph lcavol lcp lweight pgg45
            / selection=stepwise(select=SL stop=PRESS) ;
run;
```

HTpostrate Data

The GLMSELECT Procedure

Data Set	WORK.PROSTRATE
Dependent Variable	lpsa
Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	PRESS
Entry Significance Level (SLE)	0.15
Stay Significance Level (SLS)	0.15
Effect Hierarchy Enforced	None

Number of Observations Read	67
Number of Observations Used	67

Dimensions	
Number of Effects	8
Number of Parameters	8

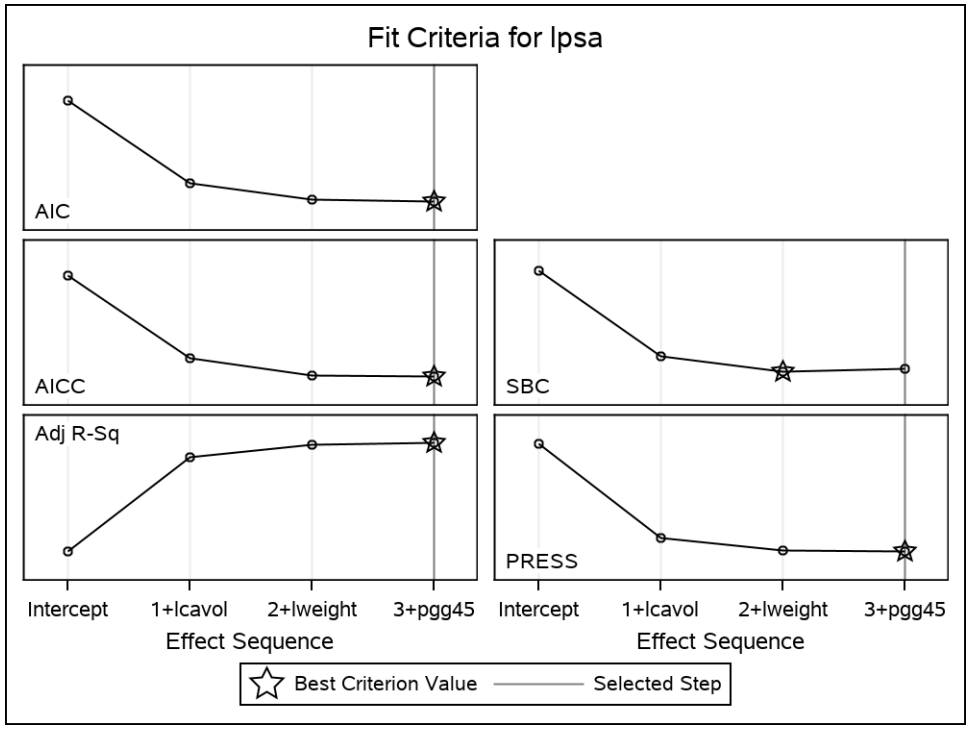
HTpostrate Data

The GLMSELECT Procedure

Stepwise Selection Summary						
Step	*Effect Entered*	*Effect Removed*	*Number Effects In*	*PRESS*	*F Value*	*Pr > F*
0	*Intercept*		1	99.2212	0.00	1.0000
1	*lcavol*		2	47.3773	75.55	<.0001
2	*lweight*		3	40.5033	12.83	0.0007
3	*pgg45*		4	40.2461*	2.95	0.0909
** Optimal Value of Criterion*						

Selection stopped at a local minimum of the PRESS criterion.

Stop Details				
Candidate For	*Effect*	*Candidate PRESS*		*Compare PRESS*
Entry	lbph	40.5011	>	40.2461
Removal	pgg45	40.5033	>	40.2461



HTpostrate Data

The GLMSELECT Procedure

Selected Model

Note	The selected model is the model at the last step (Step 3).

Effects:	Intercept lcavol lweight pgg45

Analysis of Variance				
Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*
Model	3	60.84741	20.28247	36.06
Error	63	35.43403	0.56244	
Corrected Total	66	96.28145		

Root MSE	0.74996
Dependent Mean	2.45235
R-Square	0.6320
Adj R-Sq	0.6144
AIC	34.31967
AICC	35.30328
PRESS	40.24608
SBC	−25.86156

Parameter Estimates				
Parameter	*DF*	*Estimate*	*Standard Error*	*t Value*
Intercept	1	−1.222729	0.725248	−1.69
lcavol	1	0.553546	0.089034	6.22
lweight	1	0.768076	0.203796	3.77
pgg45	1	0.006200	0.003611	1.72

4.3 Code 3 – The GLMSELECT Procedure

```
proc glmselect
data=prostrate plot=CriterionPanel; where train = 'T';
model lpsa = age gleason lbph lcavol lcp lweight pgg45
             / selection=lasso (choose=CP steps=10) ;
run;
```

HTpostrate Data

The GLMSELECT Procedure

Data Set	WORK.PROSTRATE
Dependent Variable	lpsa
Selection Method	LASSO
Stop at Specified Number of Steps	10
Choose Criterion	C(p)
Effect Hierarchy Enforced	None

Number of Observations Read	67
Number of Observations Used	67

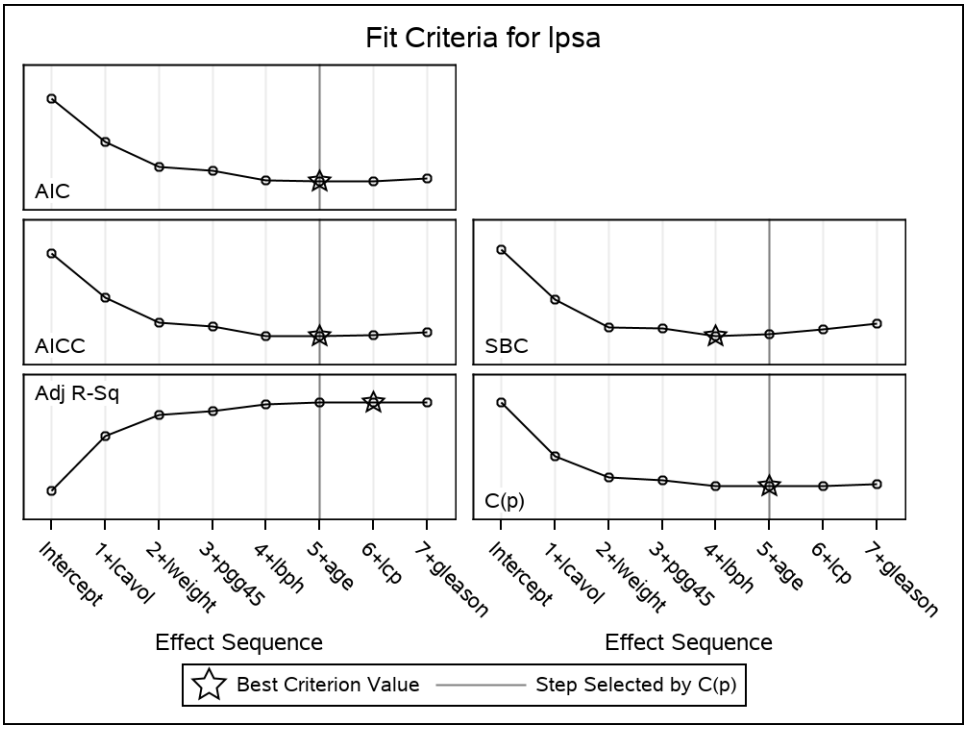
Dimensions	
Number of Effects	8
Number of Parameters	8

HTpostrate Data

The GLMSELECT Procedure

LASSO Selection Summary				
Step	*Effect Entered*	*Effect Removed*	*Number Effects In*	*CP*
0	*Intercept*		1	109.6813
1	*lcavol*		2	42.8572
2	*lweight*		3	17.0801
3	*pgg45*		4	13.6114
4	*lbph*		5	6.4424
5	*age*		6	5.8444*
6	*lcp*		7	6.4881
7	*gleason*		8	8.0000
** Optimal Value of Criterion*				

Selection stopped because all effects are in the final model.



HTpostrate Data

The GLMSELECT Procedure

Selected Model

Note	The selected model, based on C(p), is the model at Step 5.

Effects:	Intercept age lbph lcavol lweight pgg45

Analysis of Variance				
Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*
Model	5	62.74503	12.54901	22.83
Error	61	33.53642	0.54978	
Corrected Total	66	96.28145		

Root MSE	0.74147
Dependent Mean	2.45235
R-Square	0.6517
Adj R-Sq	0.6231
AIC	34.63194
AICC	36.53024
BIC	−31.17450
C(p)	5.84440
SBC	−21.13991

Parameter Estimates		
Parameter	*DF*	*Estimate*
Intercept	1	−0.030600
age	1	−0.010630
lbph	1	0.097489
lcavol	1	0.557679
lweight	1	0.625772
pgg45	1	0.006189

4.4 Code 4 – ELASTICNET

```
proc glmselect data=prostrate plot=CriterionPanel; where train = 'T';
model lpsa = age gleason lbph lcavol lcp lweight pgg45
             / selection=ELASTICNET (choose=CP steps=10) ;
run;
```

HTpostrate Data

The GLMSELECT Procedure

Data Set	WORK.PROSTRATE
Dependent Variable	lpsa
Selection Method	ELASTICNET
Stop at Specified Number of Steps	10
Choose Criterion	C(p)
Effect Hierarchy Enforced	None

Number of Observations Read	67
Number of Observations Used	67

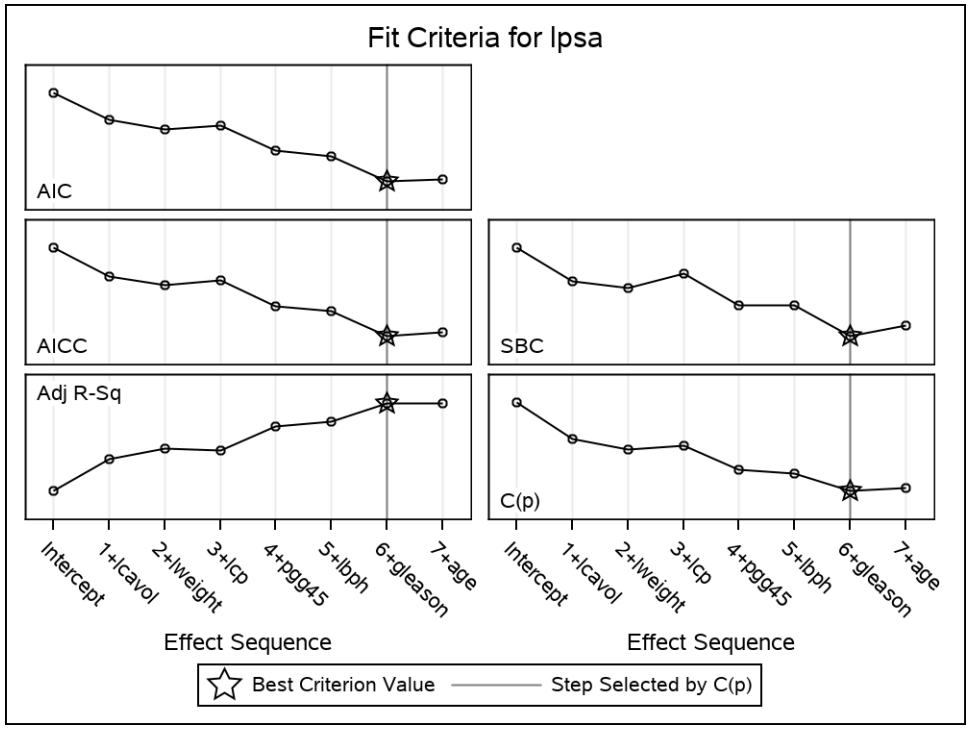
Dimensions	
Number of Effects	8
Number of Parameters	8

HTpostrate Data

The GLMSELECT Procedure

Elastic Net Selection Summary				
Step	*Effect Entered*	*Effect Removed*	*Number Effects In*	*CP*
0	*Intercept*		1	35.7175
1	*lcavol*		2	18.3184
2	*lweight*		3	13.0793
3	*lcp*		4	15.0782
4	*pgg45*		5	3.4497
5	*lbph*		6	1.6752
6	*gleason*		7	−6.5939*
7	*age*		8	−5.2690
** Optimal Value of Criterion*				

Selection stopped because all effects are in the final model.



HTpostrate Data

The GLMSELECT Procedure

Selected Model

Note	The selected model, based on C(p), is the model at Step 6.

Effects:	Intercept gleason lbph lcavol lcp lweight pgg45

Analysis of Variance				
Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*
Model	6	51.91925	8.65321	11.70
Error	60	44.36219	0.73937	
Corrected Total	66	96.28145		

Root MSE	0.85987
Dependent Mean	2.45235
R-Square	0.5392
Adj R-Sq	0.4932
AIC	55.37556

AICC	57.85832
BIC	−5.80546
C(p)	−6.59388
SBC	1.80841

Parameter Estimates		
Parameter	*DF*	*Estimate*
Intercept	1	−0.000315
gleason	1	0.074212
lbph	1	0.069400
lcavol	1	0.258607
lcp	1	0.079320
lweight	1	0.418353
pgg45	1	0.004129

4.5 Code 5 – Ridge Regression

```
proc reg data=prostrate outvif outest=b ridge=0 to .4 by .02;
            where train = 'T';
model lpsa = age gleason lbph lcavol lcp lweight pgg45 ;
run;
```

HTpostrate Data

The REG Procedure

Model: MODEL1

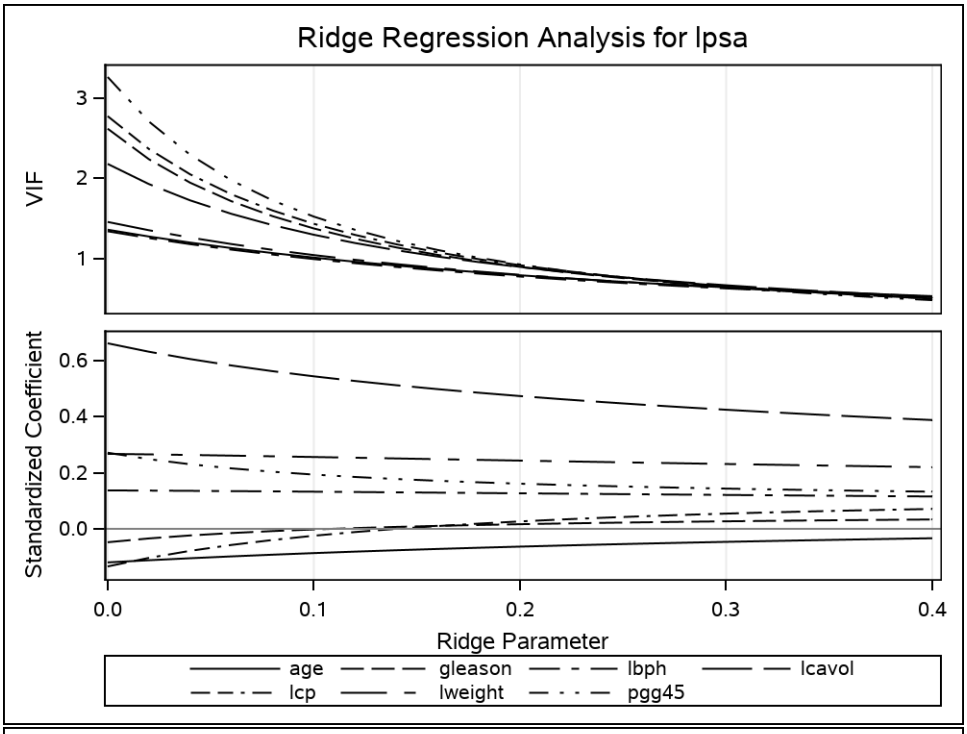
Dependent Variable: lpsa lpsa

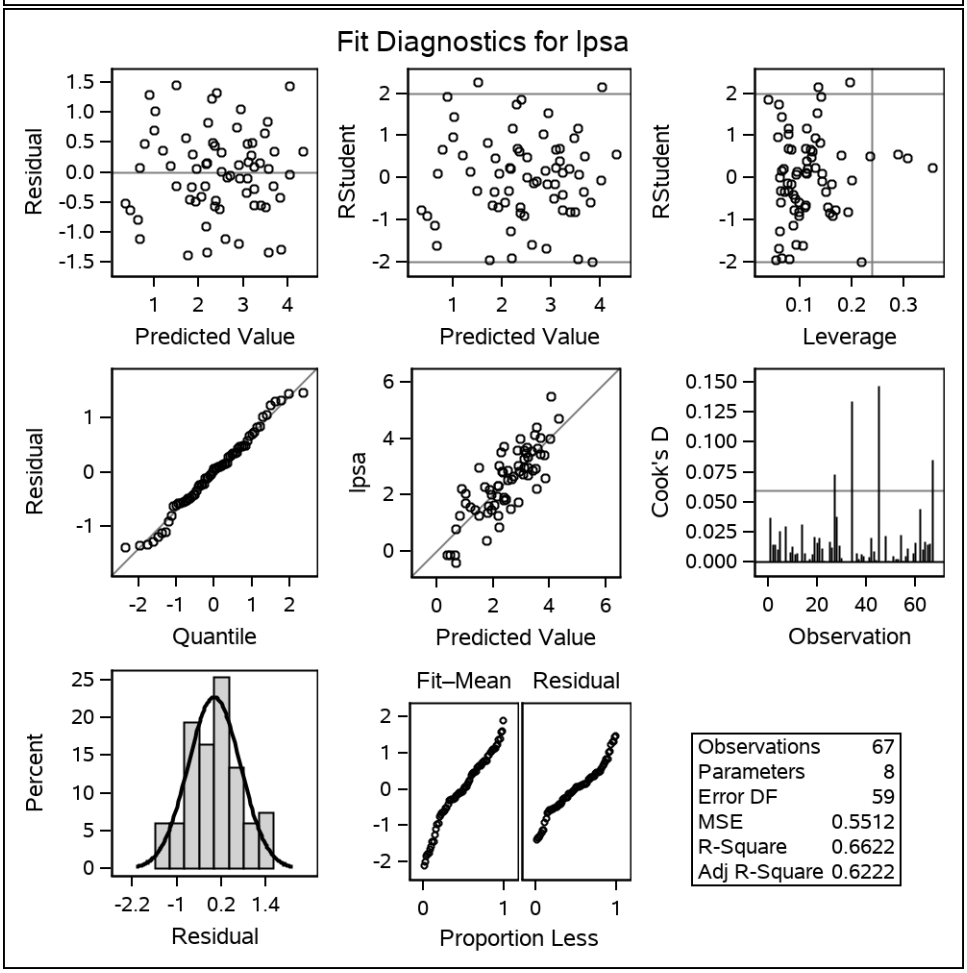
Number of Observations Read	67
Number of Observations Used	67

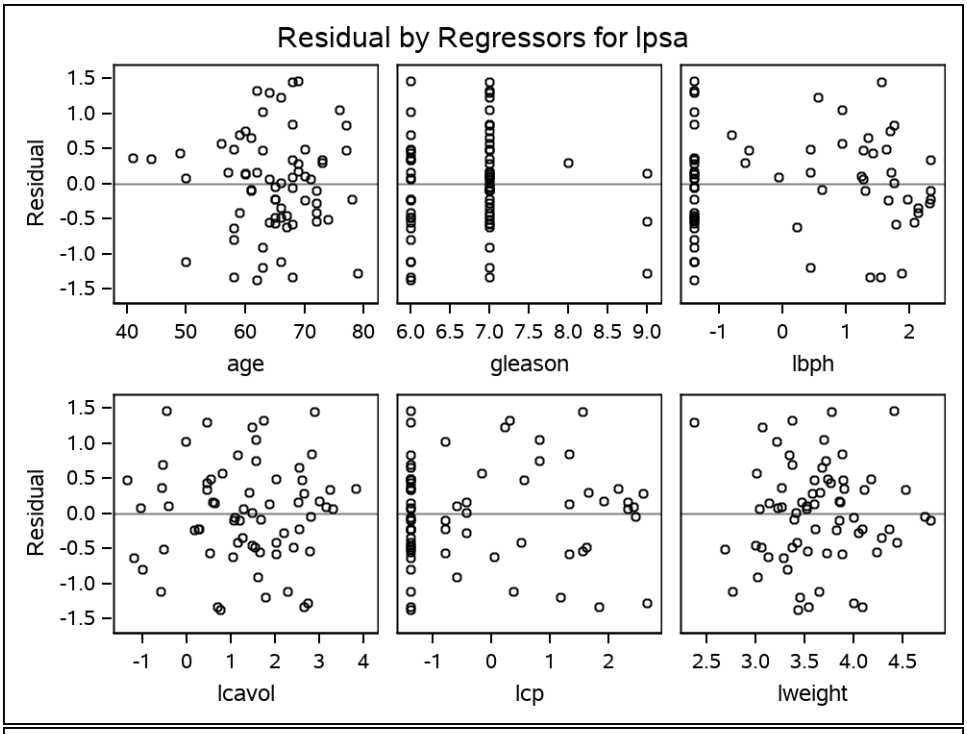
Analysis of Variance					
Source	*DF*	*Sum of Squares*	*Mean Square*	*F Value*	*Pr > F*
Model	7	63.76163	9.10880	16.53	<.0001
Error	59	32.51982	0.55118		
Corrected Total	66	96.28145			

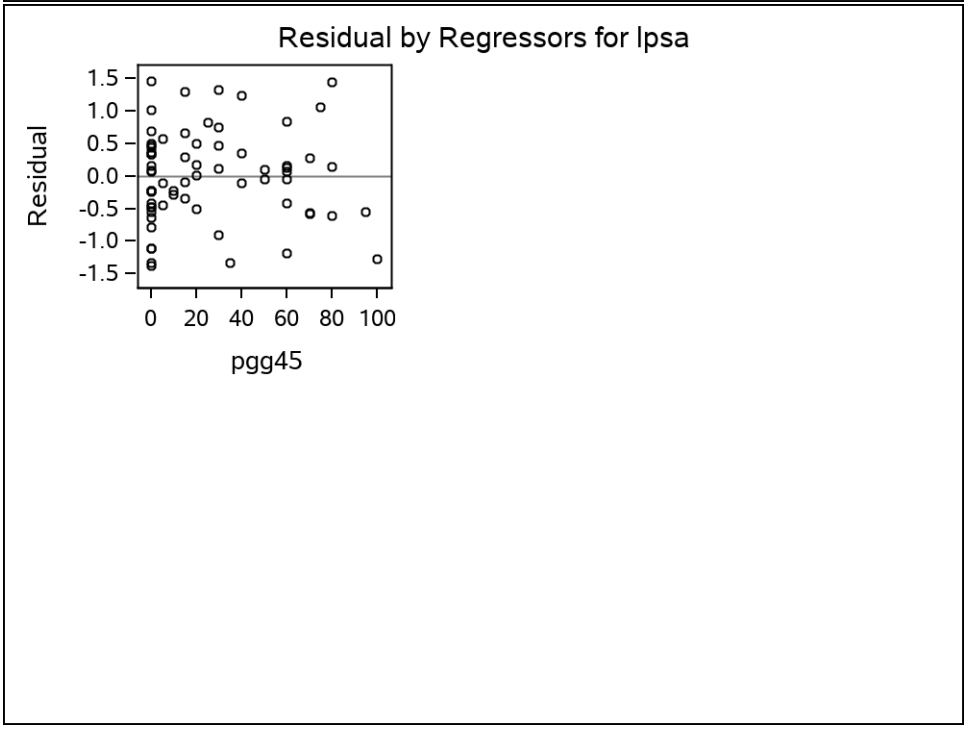
Root MSE	0.74242	*R-Square*	0.6622
Dependent Mean	2.45235	*Adj R-Sq*	0.6222
Coeff Var	30.27377		

Parameter Estimates						
Variable	*Label*	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	Intercept	1	0.62308	1.61724	0.39	0.7014
age	age	1	−0.01924	0.01419	−1.36	0.1802
gleason	gleason	1	−0.08136	0.20850	−0.39	0.6978
lbph	lbph	1	0.11293	0.07219	1.56	0.1231
lcavol	lcavol	1	0.64217	0.10850	5.92	<.0001
lcp	lcp	1	−0.11545	0.10862	−1.06	0.2921
lweight	lweight	1	0.67646	0.23116	2.93	0.0049
pgg45	pgg45	1	0.01117	0.00563	1.98	0.0519


Ridge Regression Analysis for lpsa
VIF
Standardized Coefficient
Ridge Parameter
age
gleason
lbph
lcavol
lcp
lweight
pgg45


Fit Diagnostics for lpsa
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
lpsa
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Observations 67
Parameters 8
Error DF 59
MSE 0.5512
R-Square 0.6622
Adj R-Square 0.6222


Residual by Regressors for lpsa
Residual
age
gleason
lbph
lcavol
lcp
lweight


Residual by Regressors for lpsa
Residual
pgg45