

STAT 525

Chapter 9

Model Selection

Dr. Qifan Song

Example: Patients Survival Time (p.350)

- Surgical unit wants to predict survival in patients undergoing a specific liver operation
- Random sample of 54 patients studied
- Y is post-operation survival time
- Four predictor variables:

X_1 — blood clotting score

X_2 — prognostic index (age)

X_3 — enzyme function score

X_4 — liver function score

Grand Model (all predictors considered)

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_4 X_4$$

and all important covariate are included

Grand model always true

(if $\beta_0, \beta_1 X_1$ is true)

$\therefore \beta_i$ in Grand model can be 0

Survival Time as Response

- Often skewed with a few long-lived times
- In this case, we observe all survival times
- Times can be censored if the study were prior to some subjects' deaths
 - Survival analysis techniques could be used (STAT 526)
- Transformation of survival times will be investigated using Box-Cox transformation under the grand model

```
/* CH09TA01.TXT is a tab-delimited file */  
data a1 (drop=age gender alcmod alcheavy);  
    infile 'D:\CH09TA01.TXT' delimiter='09'x;  
    input blood prog enz liver age gender alcmod alcheavy surv logsurv;  
run;
```

Grand model

'09'x delimiter means a tab-delimited file.

```
proc transreg;
  model boxcox(surv/lambda=-1 to 1 by .1) = identity(blood)
                                         identity(prog) identity(enz) identity(liver);
run; quit;
```

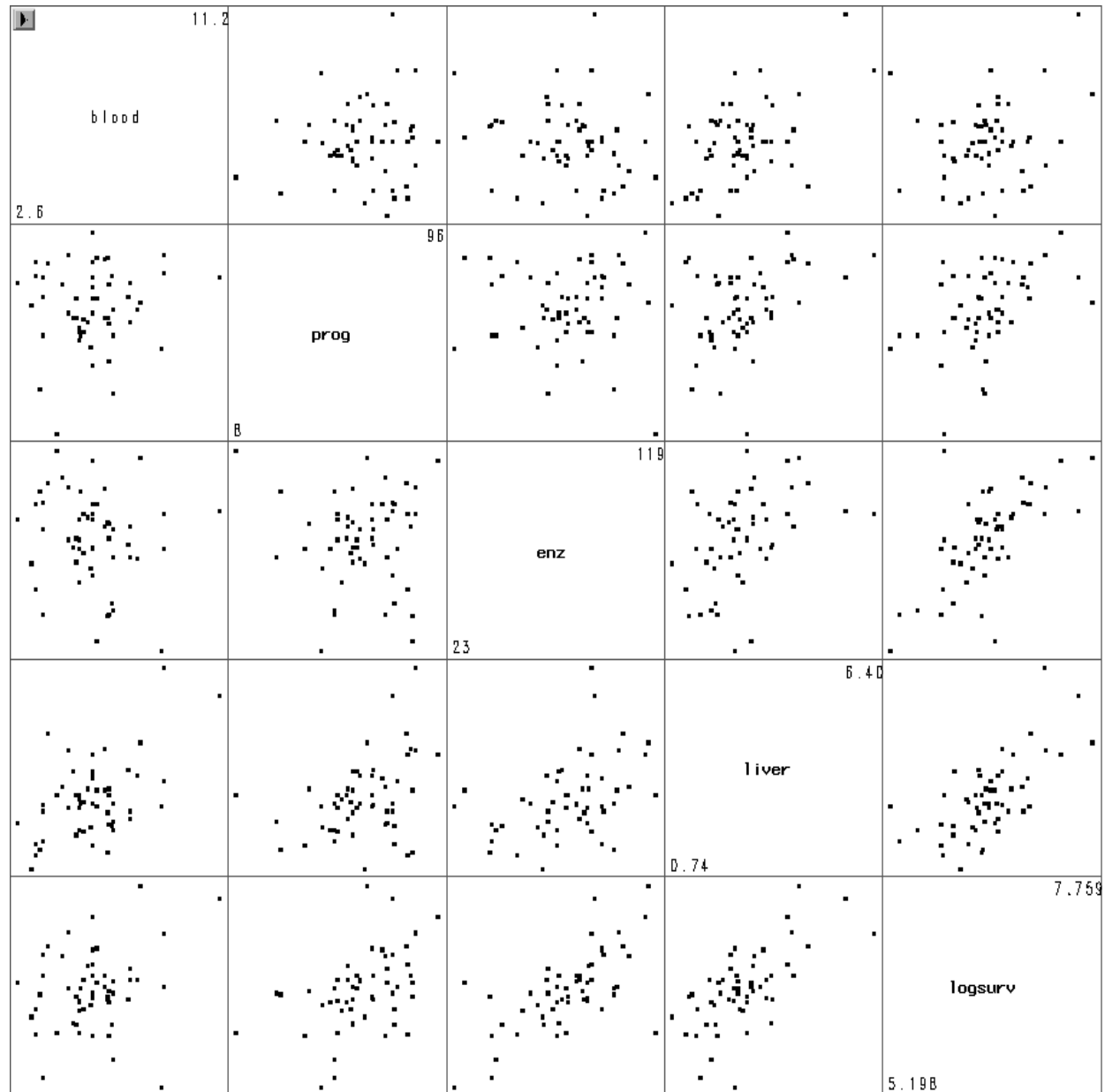
Lambda	R-Square	Log Like
-1.0	0.64	-293.077
-0.9	0.66	-289.803
-0.8	0.68	-286.714
-0.7	0.69	-283.837
-0.6	0.70	-281.203
-0.5	0.72	-278.846
-0.4	0.73	-276.805
-0.3	0.74	-275.119
-0.2	0.75	-273.828 *
-0.1	0.75	-272.971 *
0.0 +	0.76	-272.579 <
0.1	0.76	-272.675 *
0.2	0.76	-273.269 *
0.3	0.76	-274.360 *
0.4	0.75	-275.933
0.5	0.75	-277.961
0.6	0.74	-280.409
0.7	0.73	-283.238
0.8	0.72	-286.406
0.9	0.71	-289.869
1.0	0.69	-293.591

< - Best Lambda

* - Confidence Interval

+ - Convenient Lambda

Scatterplot Matrix



Predictor Summary Statistics

Prelim analysis to justify the grand model

```
/* logsurv = log(surv) */
proc corr;
    var logsurv blood prog enz liver;
run; quit;
```

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
logsurv	54	6.43048	0.49158	347.24600	5.19800	7.75900
blood	54	5.78333	1.60303	312.30000	2.60000	11.20000
prog	54	63.24074	16.90253	3415	8.00000	96.00000
enz	54	77.11111	21.25378	4164	23.00000	119.00000
liver	54	2.74426	1.07036	148.19000	0.74000	6.40000

Pearson Correlation Coefficients, N = 54

Prob > |r| under H0: Rho=0

	logsurv	blood	prog	enz	liver
logsurv	1.00000	0.24619	0.46994	0.65389	0.64926
		0.0727	0.0003	<.0001	<.0001
blood	0.24619	1.00000	0.09012	-0.14963	0.50242
	0.0727		0.5169	0.2802	0.0001
prog	0.46994	0.09012	1.00000	-0.02361	0.36903
	0.0003	0.5169		0.8655	0.0060
enz	0.65389	-0.14963	-0.02361	1.00000	0.41642
	<.0001	0.2802	0.8655		0.0017
liver	0.64926	0.50242	0.36903	0.41642	1.00000
	<.0001	0.0001	0.0060	0.0017	

Variable Selection

Grand model

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

2^{p-1}

All 2^{p-1} subset model comparison

- Two distinct questions

- What is the appropriate subset size?

adjusted R^2 , C_p , MSE, PRESS, AIC, SBC

- What is the best model for a fixed size?

R^2

- * May result in several worthy models
 - * Use subject matter to make final decision

Mallows' C_p Criterion

- Compares total mean squared error with σ^2
- Squared error

$$\begin{aligned}
 (\hat{Y}_i - \mu_i)^2 &= (\hat{Y}_i - E(\hat{Y}_i) + E(\hat{Y}_i) - \mu_i)^2 \\
 &= (E(\hat{Y}_i) - \mu_i)^2 + (\hat{Y}_i - E(\hat{Y}_i))^2 \\
 &= \text{Bias}^2 + (\hat{Y}_i - E(\hat{Y}_i))^2
 \end{aligned}$$

true
response
mean

if true $\beta_0 + \beta_1 X_1$

We have $\beta_0 + \beta_1 X_2$

or
 $\beta_0 + \beta_1 X_1 + \beta_2 X_2$

may have large diff

- Mean value is $(E(\hat{Y}_i) - \mu_i)^2 + \sigma^2(\hat{Y}_i)$
bias² + error
- Total mean value is $\sum (E(\hat{Y}_i) - \mu_i)^2 + \sum \sigma^2(\hat{Y}_i)$ based on true information
- Criterion measure

$$\begin{aligned}
 \text{true parameter } \Gamma_p &= \frac{\sum (E(\hat{Y}_i) - \mu_i)^2 + \sum \sigma^2(\hat{Y}_i)}{\sigma^2} \\
 &= \frac{\sum \text{Bias}^2 + \sum \text{Var}(\text{prediction})}{\text{Var}(\text{error})}
 \end{aligned}$$

- Estimate σ^2 from the full model ($P - 1$ predictors in total)

- $\hat{\sigma}^2 = \text{MSE}(X_1, X_2, \dots, X_{P-1}) = \text{MSE}_P$ *MSE in grand (all p predictors)*

- Consider a model with $p - 1$ predictors

- Can show $E(\text{SSE}_p) = \sum (E(\hat{Y}_i) - \mu_i)^2 + (n - p)\sigma^2$

- * Estimate $\sum (E(\hat{Y}_i) - \mu_i)^2$ by $\text{SSE}_p - (n - p) \text{MSE}_P$

- Note $\sum \sigma^2(\hat{Y}_i) = \text{Trace}\{\sigma^2(\hat{\mathbf{Y}})\} = \sigma^2 \text{Trace}\{\mathbf{H}\} = p\sigma^2$

- * Estimate $\sum \sigma^2(\hat{Y}_i)$ by $p\text{MSE}_P$

$$\text{Cov}(\hat{\mathbf{Y}})$$

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

$$\text{Cov}(\hat{\mathbf{Y}}) = \mathbf{H} \text{Cov}(\mathbf{Y}) \mathbf{H}^T$$

$$= \sigma^2 \mathbf{I} - \mathbf{H}$$

$$= \sigma^2 \mathbf{H}$$

- Putting it together, Γ_p is estimated by

$$\begin{aligned} C_p &= \frac{(\text{SSE}_p - (n - p)\text{MSE}_P) + p\text{MSE}_P}{\text{MSE}_P} \\ &= \frac{\text{SSE}_p}{\text{MSE}(X_1, X_2, \dots, X_{P-1})} - (n - 2p) \end{aligned}$$

- A good model has no bias

all true predictors
 \Rightarrow no bias for μ

$$\Gamma_p = \frac{0 + p\sigma^2}{\sigma^2} = p; \quad E(C_p) \approx p;$$

- A bad model is biased

$$\Gamma_p > \frac{0 + p\sigma^2}{\sigma^2} = p; \quad E(C_p) > p;$$

- When plotting models against p

- Biased models will fall above $C_p = p$
- Unbiased models will fall around the line $C_p = p$

Adjusted R^2 Criterion

- Takes into account the number of parameters in the model
- Switches from SS's to MS's

$$\underline{R_a^2} = 1 - \left(\frac{n-1}{n-p} \right) \frac{\text{SSE}}{\text{SSTO}} = 1 - \frac{\text{MSE}}{\text{MSTO}}$$

*in math
can be negative*

- Choose model which maximizes R_a^2
- Same approach as choosing smallest MSE

PRESS_p Criterion

- Looks at the **P**REdiction **S**um of **S**quares which quantifies how well the fitted values can predict the observed responses
 - For each case i , predict Y_i using model generated from other $n - 1$ cases
 - $\text{PRESS} = \sum (Y_i - \hat{Y}_{i(i)})^2$
- Want to select model with small PRESS
- Can calculate this in one fit (Chapter 10)

prediction for Y_i , but not gain for X_i, Y_i

usually

$$\begin{pmatrix} X_1 & Y_1 \\ \vdots & \vdots \\ X_n & Y_n \end{pmatrix} \rightarrow \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \rightarrow \hat{Y}_{i(1)}$$

remove one sample

$$\begin{pmatrix} Y_2 & Y_2 \\ \vdots & \vdots \\ X_n & Y_n \end{pmatrix} \rightarrow \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \xrightarrow{X_1} \hat{Y}_{i(1)}$$

Other Approaches

$$Y_i = \beta_0 + \beta_1 X_i$$

$$\downarrow$$

$$\max_{\beta_0, \beta_1, \sigma^2} \left\{ -2 \log \frac{1}{(\sqrt{2\pi})^n} \exp \left\{ - \frac{\sum (Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2} \right\} \right\}$$

- Criterion based on minimizing $-2\log(\text{likelihood})$ plus a penalty for more complex model
- AIC - Akaike's information criterion (similar to C_p)

$$n \log \left(\frac{\text{SSE}_p}{n} \right) + 2p$$

when $n \uparrow \rightarrow$ not sufficient
penalty

- SBC - Schwarz Bayesian Criterion

$$n \log \left(\frac{\text{SSE}_p}{n} \right) + p \log(n)$$

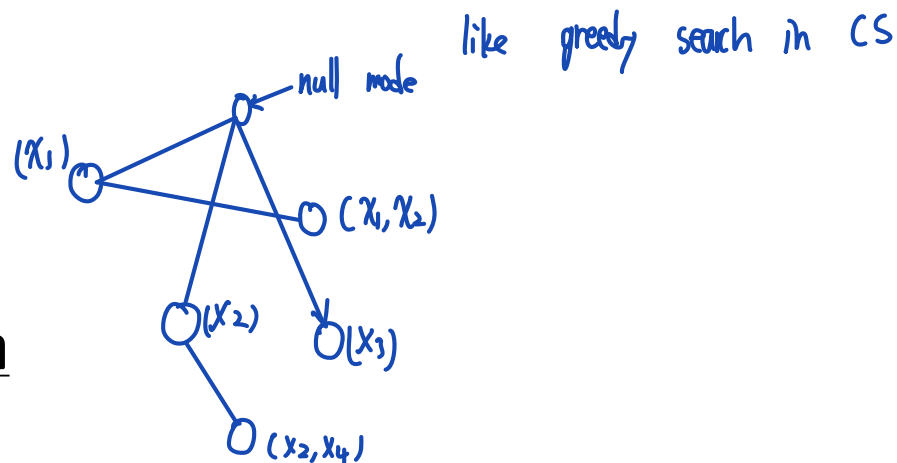
when $n \uparrow$, BIC more consistent
stronger penalty

– It is also called BIC, i.e, Bayesian Information Criterion

- If n is very large, AIC tends to overselect the model, while BIC is consistent.

SELECTION in SAS

- Helpful options in MODEL statement
 - SELECTION = to choose model selection procedure and criterion
 - * FORWARD (step up)
 - * BACKWARD (step down)
 - * STEPWISE (forward with backward glance)
 - * RSQUARE, ADJRSQ, CP (all subset selection using the specified criterion)
 - INCLUDE = *n* forces first *n* variables into all models
 - BEST = *n* limits output to the best *n* models
 - START = *n* limits output to models with $\geq n$ *X*'s
 - B will include parameter estimates



Forward Selection

start with nothing

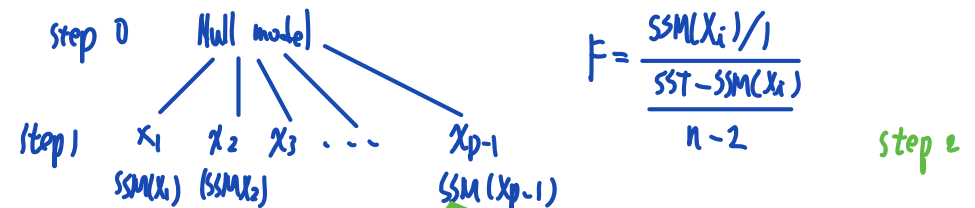
add predictors one by one

/* SLE: significance level for entry into the model */

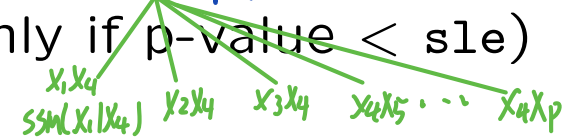
/* Default: sle=0.5

model logsurv = blood prog enz liver/selection=f sle=0.5;

- Start with no variables



- Add one variable with best F-value (only if p-value < sle)
on Forward Selection, no reduced or Full models are true
hence, p-value is relative
- Add the next variable with best F-value given the previous variables in the model (only if p-value < sle)
- Stop if no variables can be added with p-value < sle



Backward Elimination

```
/* SLS: significance level for staying in the model */  
/* Default: sls = 0.10 */  
model logsurv = blood prog enz liver/selection=b sls=0.10;
```

- Start with all the variables
- Delete the variable that has the smallest extra SS (only if p-value $>$ sls)
- Delete the next variable that has the smallest extra SS (only if p-value $>$ sls)
- Stop when all variables have p-value $<$ sls

Stepwise Search

```
/* Default: sle = 0.15, sls=0.15 */  
model logsurv = blood prog enz liver/selection=stepwise;
```

- Start with no variables
- Add variables sequentially as in forward selection, using `sle`
- 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 `sls`
- Stop when nothing can be added, and nothing non-significant can be removed
- Fix $sle \leq sls$ to void cycling.

All Subset Selection

- Select from all the possible models
 - `selection=rsquare`
 - `selection=adjrsq`
 - `selection=cp`

Models of Same Subset Size

- Can also use R^2 or SSE
- May result in several worthy models
- Use knowledge on subject matter to make final decision
- Decision not that important if goal is prediction

Example: All Subset Selection in SAS

```
/* ---- Variable Selection: Quality of Fit ---- */
proc reg data=a1;
    model logsurv = blood prog enz liver / ad R2 cp intercept
    selection = rsquare R2 adjrsq cp b;
run;
```

which selection method you want to use

Number in Model	R-Square	Adjusted R-Square	C(p)
<i>size</i> 1	0.4276	0.4166	66.4889
1	0.4215	0.4104	67.7148
1	0.2208	0.2059	108.5558
1	0.0606	0.0425	141.1639

2	0.6633	0.6501	20.5197
2	0.5995	0.5838	33.5041
2	0.5486	0.5309	43.8517
2	0.4830	0.4627	57.2149
2	0.4301	0.4078	67.9721
2	0.2627	0.2338	102.0313

3	0.7573	0.7427	3.3905
3	0.7178	0.7009	11.4237
3	0.6121	0.5889	32.9320
3	0.4870	0.4562	58.3917

4	0.7592	0.7396	5.0000

Number in Model	R-Square	-----Parameter Estimates-----				
		Intercept	blood	prog	enz	liver
1	0.4276	5.26426	.	.	0.01512	.
1	0.4215	5.61218	.	.	.	0.29819
1	0.2208	5.56613	.	0.01367	.	.
1	0.0606	5.99386	0.07550	.	.	.

2	0.6633	4.35058	.	0.01412	0.01539	.
2	0.5995	5.02818	.	.	0.01073	0.20945
2	0.5486	4.54623	0.10792	.	0.01634	.
2	0.4830	5.24574	.	0.00776	.	0.25299
2	0.4301	5.73422	-0.03282	.	.	0.32288
2	0.2627	5.23573	0.06302	0.01313	.	.

3	0.7573	3.76618	0.09546	0.01334	0.01645	.
3	0.7178	4.40582	.	0.01101	0.01261	0.12977
3	0.6121	4.78168	0.04482	.	0.01220	0.16360
3	0.4870	5.34144	-0.02272	0.00752	.	0.27147

4	0.7592	3.85195	0.08368	0.01266	0.01563	0.03216

```

/* ---- PRESS: Quality of Prediction ---- */
/* OUTEST: outputs estimates & model fit summary statistics */
/* PRESS: outputs the PRESS statistic of the (final selected) model to the
OUTEST= data set, use multiple model statement to obtain multiple PRESS
statistics */

```

```

/* Model with Three/Four Predictors */
proc reg data=a1 outest=sumstats1 press;
    model logsurv = blood prog enz;
    model logsurv = blood prog enz liver;
run;

```

```

proc print data=sumstats1; run; quit;

```

Obs	_MODEL_	...	_RMSE_	_PRESS_	Intercept	...	enz	liver	logsurv
1	MODEL1	...	0.24934	3.91424	3.76618	...	0.016452	.	-1
2	MODEL2	...	0.25087	4.06857	3.85195	...	0.015632	0.032161	-1

Example: Stepwise Selection in SAS

```
data a1;
  infile 'D:\nobackup\tmp\CH09TA01.TXT' delimiter='09'x;
  input blood prog enz liver age gender alcmod alcheavy surv logsurv;
run;

proc reg data=a1;
  model logsurv=blood prog enz liver age gender alcmod alcheavy/
    selection=stepwise;
run; quit;
```

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	enz		1	0.4276	0.4276	117.409	38.84	<.0001
2	prog		2	0.2357	0.6633	50.4716	35.70	<.0001
3	alcheavy		3	0.1147	0.7780	18.9145	25.85	<.0001
4	blood		4	0.0519	0.8299	5.7508	14.93	0.0003
5	gender		5	0.0076	0.8374	5.5406	2.23	0.1418

LASSO selection

- LASSO (least absolute shrinkage and selection operator):

$$\mathbf{b} = \arg \min \|\mathbf{Y} - \mathbf{X}\mathbf{b}\|, \text{ subject to } \sum |b_i| \leq t$$

- Equivalent estimation:

$$\mathbf{b} = \arg \min \|\mathbf{Y} - \mathbf{X}\mathbf{b}\| + \lambda \sum |b_i|$$

- a sequence of models, from null to full, as t increases
- One can pick one among them, depending on some criterion.

LASSO in SAS

```
proc glmselect data=a1 plot=CriterionPanel ;  
    model logsurv=blood prog enz liver age gender alcmod alcheavy  
        / selection=LASSO(choose=SBC steps=10);  
run;
```

LASSO Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	SBC
0	Intercept		1	-73.7135
1	enz		2	-70.2835
2	liver		3	-89.6791
3	prog		4	-98.5064
4	alcheavy		5	-135.1305
5	blood		6	-142.6026
6	gender		7	-142.3672
7	age		8	-144.4504*
8	alcmod		9	-142.8701

* Optimal Value of Criterion

Plus additional table, e.g. parameter estimations, and figures.

Cross Validation

```
* ---- select random 30 subjects ----;
proc surveyselect data=a1 method=srs n=30 out=a3 outall;
run;

* ---- take the 30 subjects in the training set ----;
data train; set a3; if Selected=1; run;

* ---- the remainders in the validation set ----;
data val; set a3; if Selected ne 1; run;

* ---- regression model on the training set ----;
proc reg data=train outest=trainStats;
    model lsurv = blood prog enz;
run;

* ---- predict survival on the validation set ----;
proc score data=val score=trainStats
    out=valStats type=parms;
var blood prog enz ;
run;

* ---- find MSPR ----;
data valSummary; set valStats;
    spr = (lsurv - MODEL1)*(lsurv - MODEL1);
run;
proc univariate data=valSummary; var spr; run;
```

Chapter Review

- Variable Selection Criteria
 - R^2 , Adjusted R^2
 - C_p
 - PRESS
 - AIC, SBC (aka BIC)
- Automatic Search Procedures
 - Forward Selection
 - Backward Elimination
 - Stepwise Search
 - All Subset Selection
 - LASSO