Machine Learning: In the trenches

Michigan SAS Users Conference

Brenda W. Gillespie

May 30, 2019

We will focus on the Machine Learning options of SAS GLMSelect

- GLMSelect models continuous outcome variables as a function of multiple predictor variables, including categorical variables, interactions, splines, etc.
  - The goal is usually prediction, with a **focus on avoiding over-fitting**
  - Can handle huge numbers of variables (more than # observations)
  - Also useful with modest numbers of variables where avoiding over-fitting is important

- GLMSelect has many options. How to choose?

Machine Learning: 4 methods

LAR (least angle regression)

Like forward selection, starts with no effects in the model and adds effects.

The parameter estimates at any step are "shrunk" when compared to the corresponding least squares estimates.

Classification variables are split.

LASSO (Least Absolute Shrinkage and Selection Operator)

Adds and deletes parameters based on ordinary least squares with a **constraint is based on the sum of the absolute regression coefficients**.

Classification variables are split.

ELASTICNET

Estimates parameters based on ordinary least squares with a **constraint is based on a linear combination of the sum of the absolute regression coefficients (like LASSO) and the sum of the squared regression coefficients (like Ridge)**

Classification variables are split.

GROUPLASSO

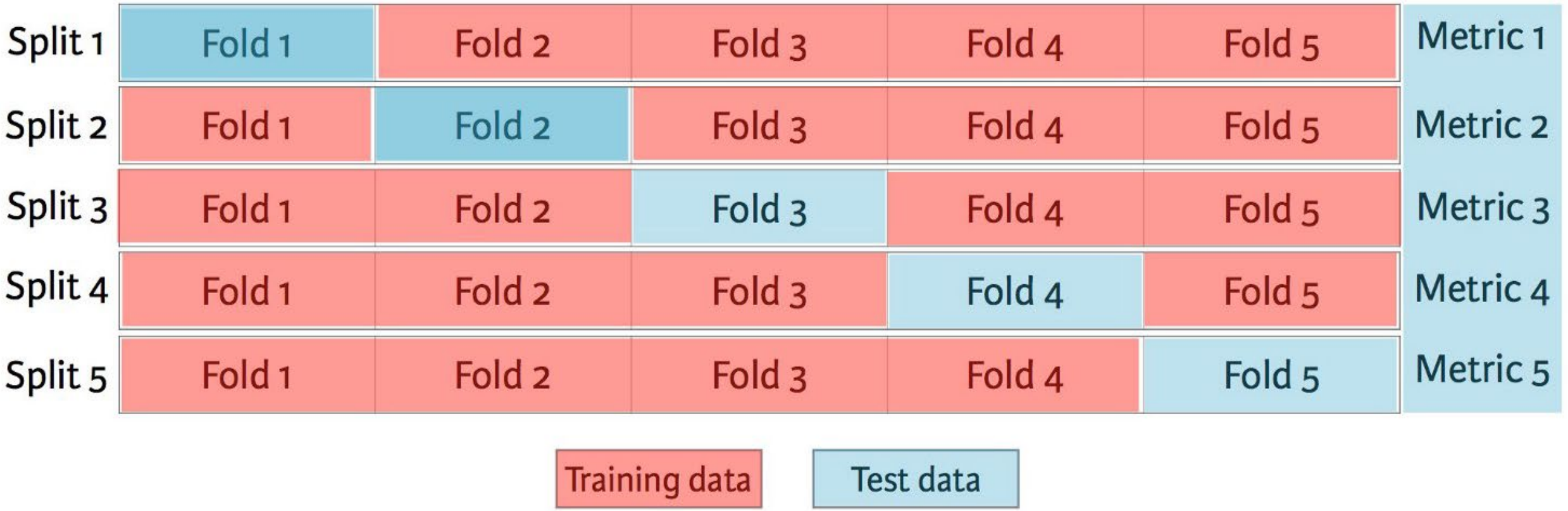
A variant of LASSO that estimates parameters based on a version of ordinary least squares in which the sum of the Euclidean norms of a group of regression coefficients is constrained (e.g., all parameters associated with a spline).

Proc GLMSelect;
MODEL Y = X1 – X500 / <options>; run;

Table 49.5: MODEL Statement Options

Option	Description
CVDETAILS=	Requests details when cross validation is used
CVMETHOD=	Specifies how subsets for cross validation are formed
DETAILS=	Specifies details to be displayed
FUZZ=	Specifies the tolerance range for criterion comparisons
HIERARCHY=	Specifies the hierarchy of effects to impose
NOINT	Specifies models without an explicit intercept
ORDERSELECT	Requests that parameter estimates be displayed in the order in which the parameters first entered the model
SELECTION=	Specifies the model selection **method**
SHOWPVALUES	Requests p-values in "ANOVA" and "Parameter Estimates" tables
STATS=	Specifies additional statistics to be displayed

5-fold Cross Validation



For each of the *5* Splits, 4 parts are used as a training set (red) and the 'left out' set is used as the test set (blue). The average error across all *5* *test-set* trials is computed.

Eijaz Allibhai, www.datacamp.com

CVMETHOD=BLOCK(n), SPLIT(n), RANDOM(n), INDEX (n) (n)=number of CV groups. For Training data.

- BLOCK requests that parts be formed of n blocks of consecutive training observations.
- SPLIT requests that the ith part consist of training observations (i, i+n, i+2n, …)
- RANDOM assigns each training observation randomly to one of the n parts.
- INDEX(variable) – You specify the sets based on the value of a named variable.
- Defaults: n=5 with CVMETHOD=BLOCK, SPLIT, or RANDOM.

  CVMETHOD=RANDOM(5).

Table 49.6: Applicable **SELECTION=** Method (columns) & Options (rows)

Option	FORWARD	BACKWARD	STEPWISE	LAR	LASSO	ELASTICNET	GROUPLASSO
STOP=	X	X	X	X	X	X	X
CHOOSE=	X	X	X	X	X	X	X
STEPS=	X	X	X	X	X	X	X
MAXSTEP=	X	X	X	X	X	X	X
SELECT=	X	X	X				
INCLUDE=	X	X	X				
SLENTRY=	X		X	X	X	X	
SLSTAY= (SL=signif level)		X	X		X	X	
DROP=			X				
ADAPTIVE					X		
LSCOEFFS				X	X		
L1= regularization or constraint parameter					X	X	X
L1CHOICE=					X	X	
L2= ridge parameter						X	
L2STEPS=						X	
L2LOW=						X	
L2HIGH=						X	
L2SEARCH=						X	
ENSCALE						X	
SCREEN=					X	X	

Table 49.6: Applicable **SELECTION=** Method (columns) & Options (rows)

Option	Method	
	LASSO	ELASTICNET
STOP= (when to stop adding variables)	X	X
CHOOSE= (how to choose the best fit within the stop range)	X	X
STEPS=	X	X
MAXSTEP=	X	X
SLENTRY= (SL=signif level)	X	X
SLSTAY=	X	X
ADAPTIVE	X	
LSCOEFFS	X	
L1= lasso constraint parameter	X	X
L1CHOICE=	X	X
L2= ridge parameter		X
L2STEPS=		X
L2LOW=		X
L2HIGH=		X
SCREEN=	X	X

STOP and then CHOOSE

- **STOP** specifies when PROC GLMSELECT is to stop the selection process
- **CHOOSE** specifies the criterion for choosing the 'best' model.
  - The criterion, which is the (average of the n) residual SS by n-fold cross-validation, is evaluated at each step of the selection process; the model with the best value of the criterion is chosen. (The default chooses the model at the last possible step.)

SAS code: Selection=LASSO
(**stop=30** choose=CV)

The graph shows each step.
The best model choice by
Cross-validation is at step 12

[Same for Elastic Net]

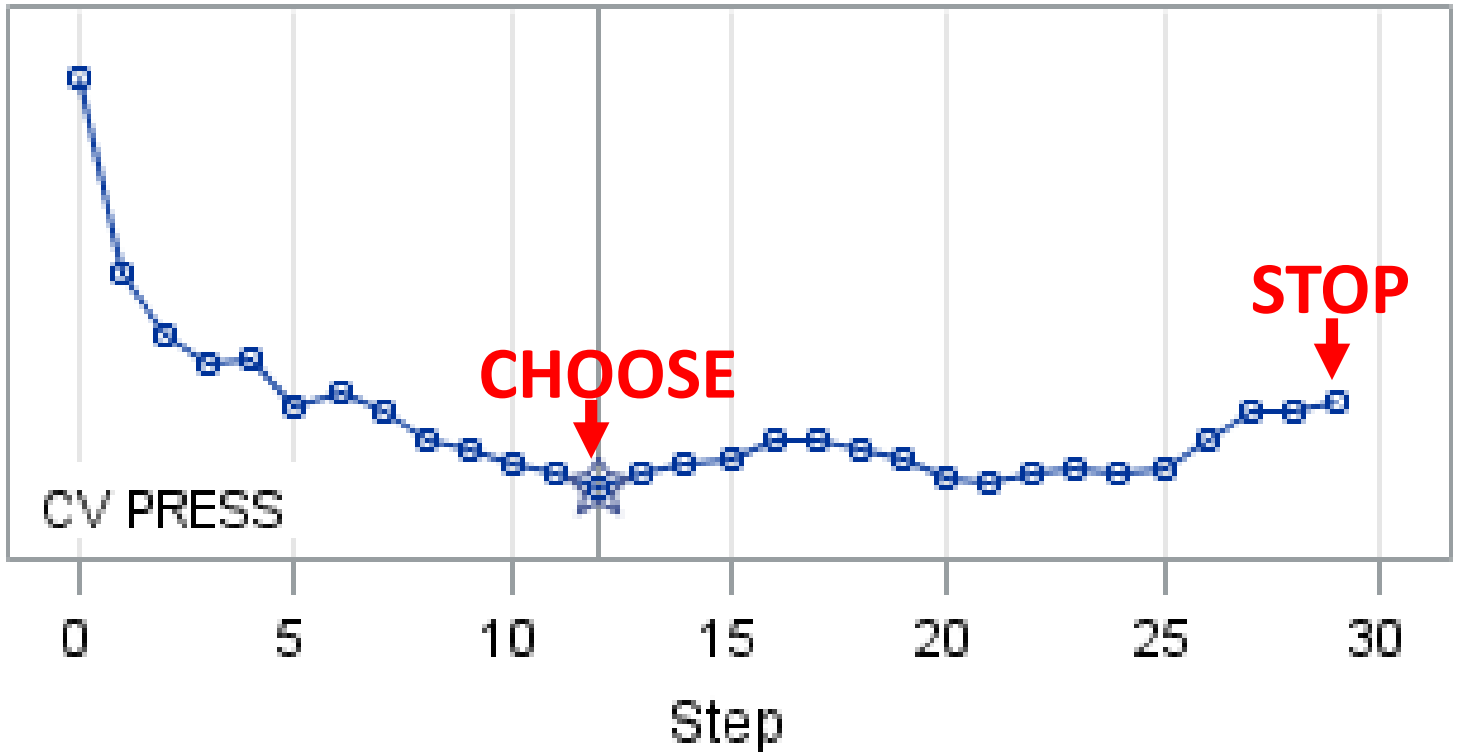


Table 49.8: Criteria for the STOP= Option

Option	Criteria [Highlighted options cannot be used in CHOOSE]
Number	Stop at specified step number (e.g., 30)
ADJRSQ	Adjusted R-square statistic [NOTE: If STOP not specified, STOP=SBC]
AIC	Akaike's information criterion
AICC	Corrected Akaike's information criterion
BIC	Sawa Bayesian information criterion
CP	Mallows' C_p statistic
CV	Predicted residual sum of square with k-fold cross validation
L1	The LASSO regularization or constraint parameter
PRESS	Predicted residual sum of squares
SBC	Schwarz Bayesian information criterion
SL	Significance level
VALIDATE	Average square error for the validation data

Table 49.7: Criteria for the CHOOSE= Option

Criterion	[Highlighted options cannot be used in STOP]
ADJRSQ	Adjusted R-square statistic
AIC	Akaike's information criterion
AICC	Corrected Akaike's information criterion
BIC	Sawa Bayesian information criterion
CP	Mallows' C_p statistic
CV	Predicted residual sum of square with k-fold cross validation
CVEX	Predicted residual sum of square with k-fold external cross validation
PRESS	Predicted residual sum of squares
SBC	Schwarz Bayesian information criterion
VALIDATE	Average square error for the validation data

Proc GLMSelect;
MODEL Y = X1 – X500 / SELECTION= method (<method options>)

Examples:

MODEL Y = X1 – X500 / SELECTION=LASSO (ADAPTIVE CHOOSE=AIC)

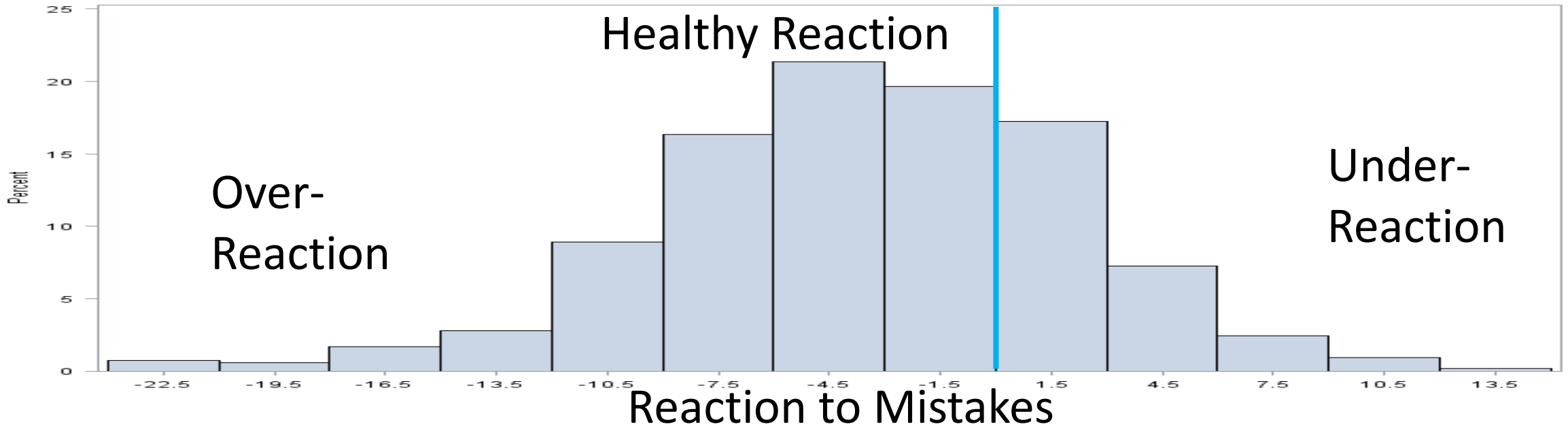
- In the LASSO method, **ADAPTIVE** requests that adaptive weights be applied to each of the coefficients (either specified or obtained from OLS)

- MODEL Y = X1 – X500
  / SELECTION=LASSO (SCREEN=SASVI stop=50 choose=BIC)

- **SCREEN** gives options for faster processing when you have a very large number of possible predictors but expect a relatively small subset of true predictors
  - SASVI = safe screening technique

Example: Reactions to making mistakes

- We are interested in predicting reactions to making mistakes.
- Some reaction to making mistakes is healthy
- Over-reaction to mistakes can be debilitating
- Under-reaction misses the chance for learning and correction



Goal: To Predict Reaction to Mistakes

- N = 539 subjects
- Y = Mistake_reaction (continuous variable)

- Potential predictors of reaction to be tested:
  - Age
  - Gender
  - Test Accuracy
  - 117 other variables measuring characteristics such as introversion/extroversion or behavioral disorders, measured by questionnaire

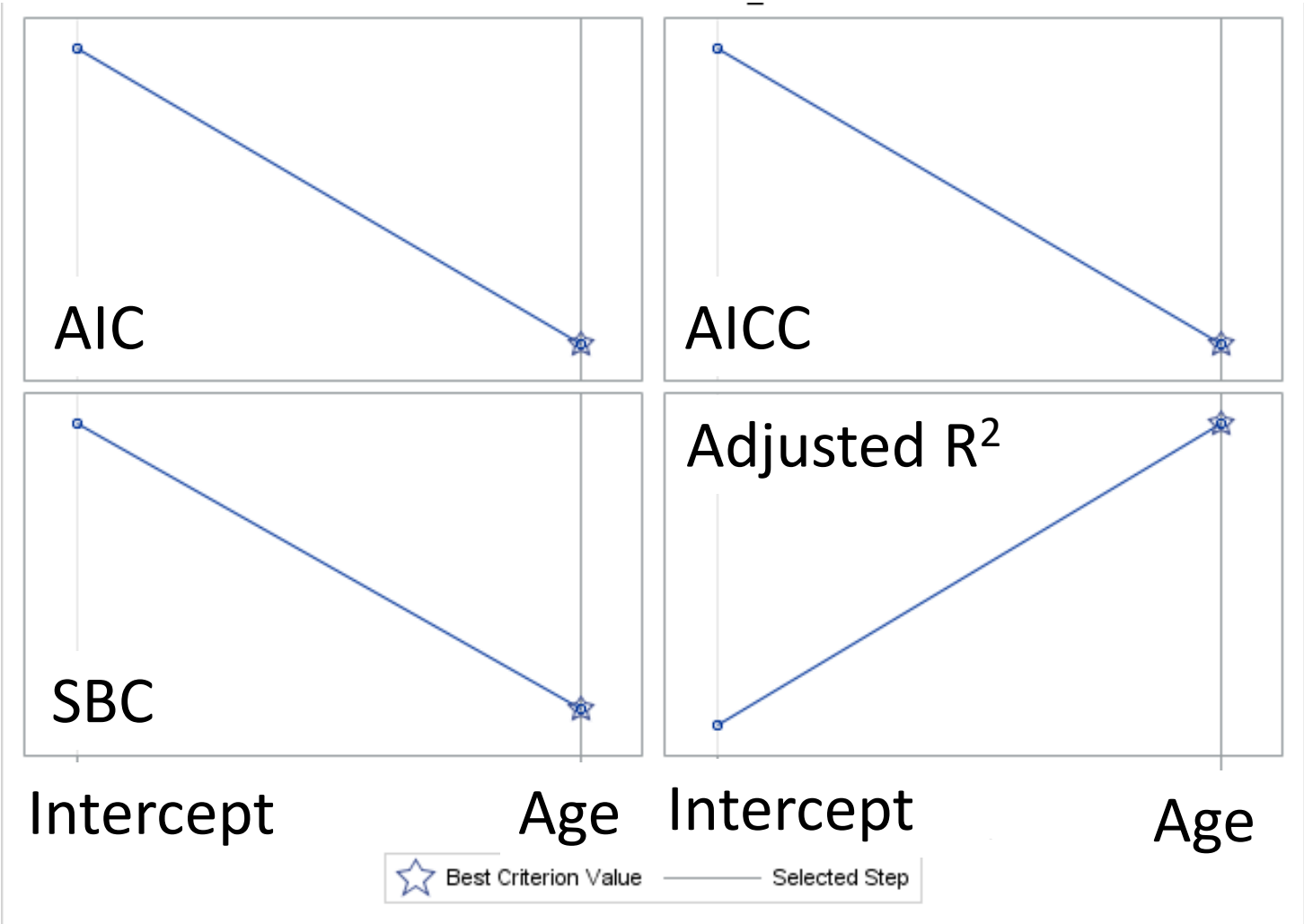
```
proc glmselect plots=all;  /*standardized dataset done automatically*/
        model Mistake_reaction = Age Gender Accuracy  X1 – X117
                                / selection=lasso;  run;
```

Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	-0.793339
Age_yrs	**1**	**-0.196773**

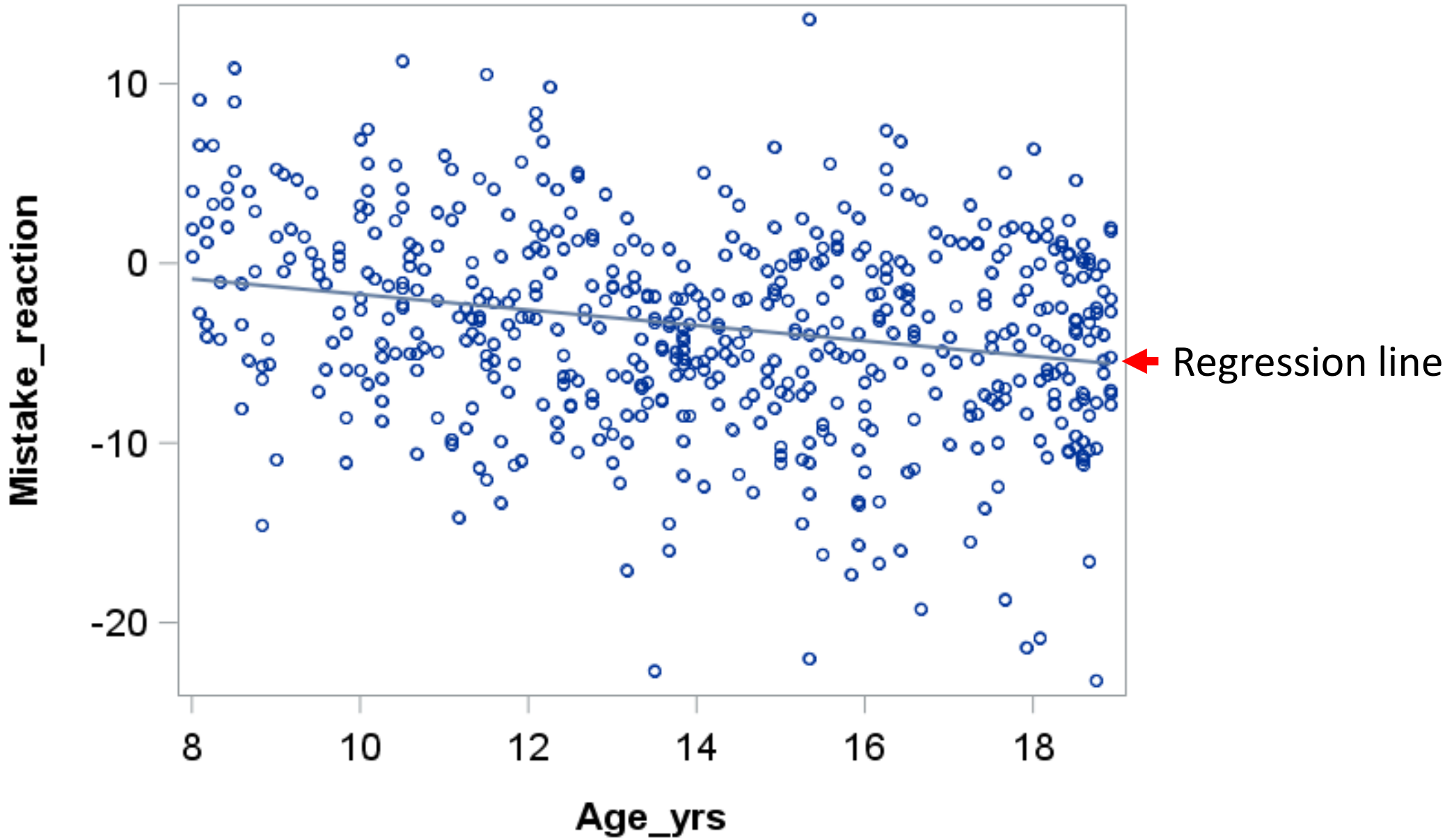
Out of 120 variables, only one survived!

Same result with Elastic Net.

If not specified, STOP=SBC

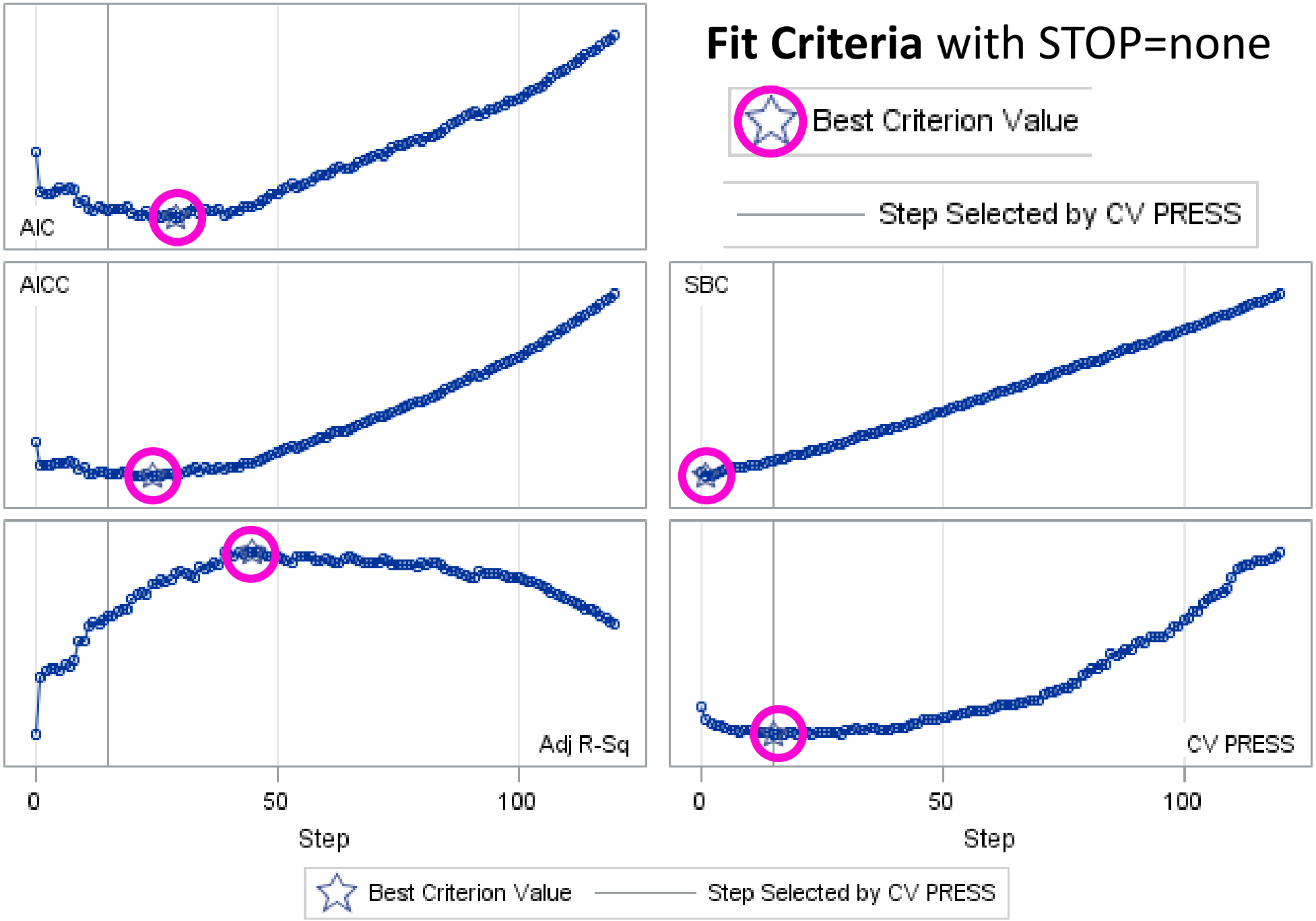


Scatterplot of Reaction to Mistakes by Age (years)

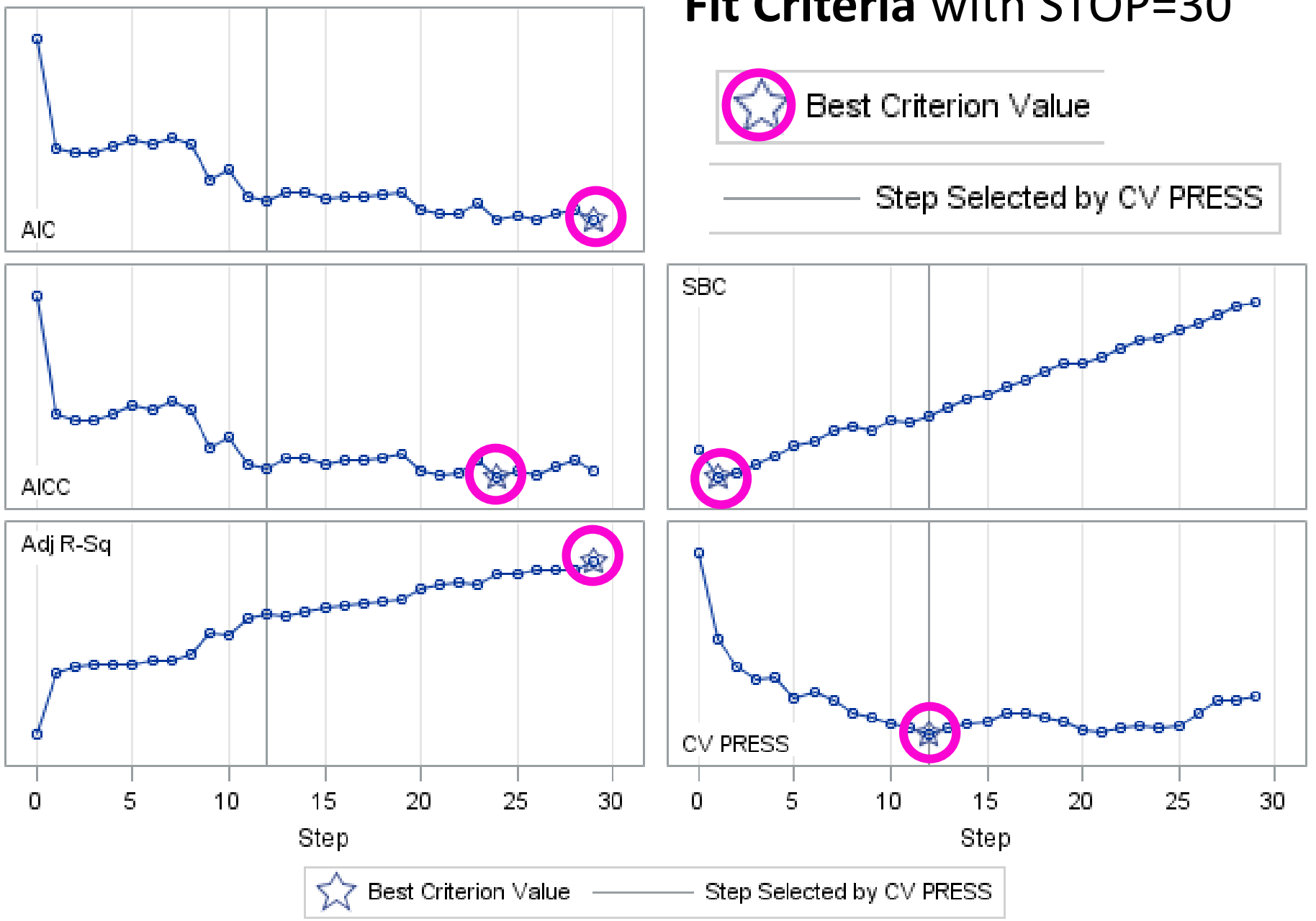


Try LASSO with Options

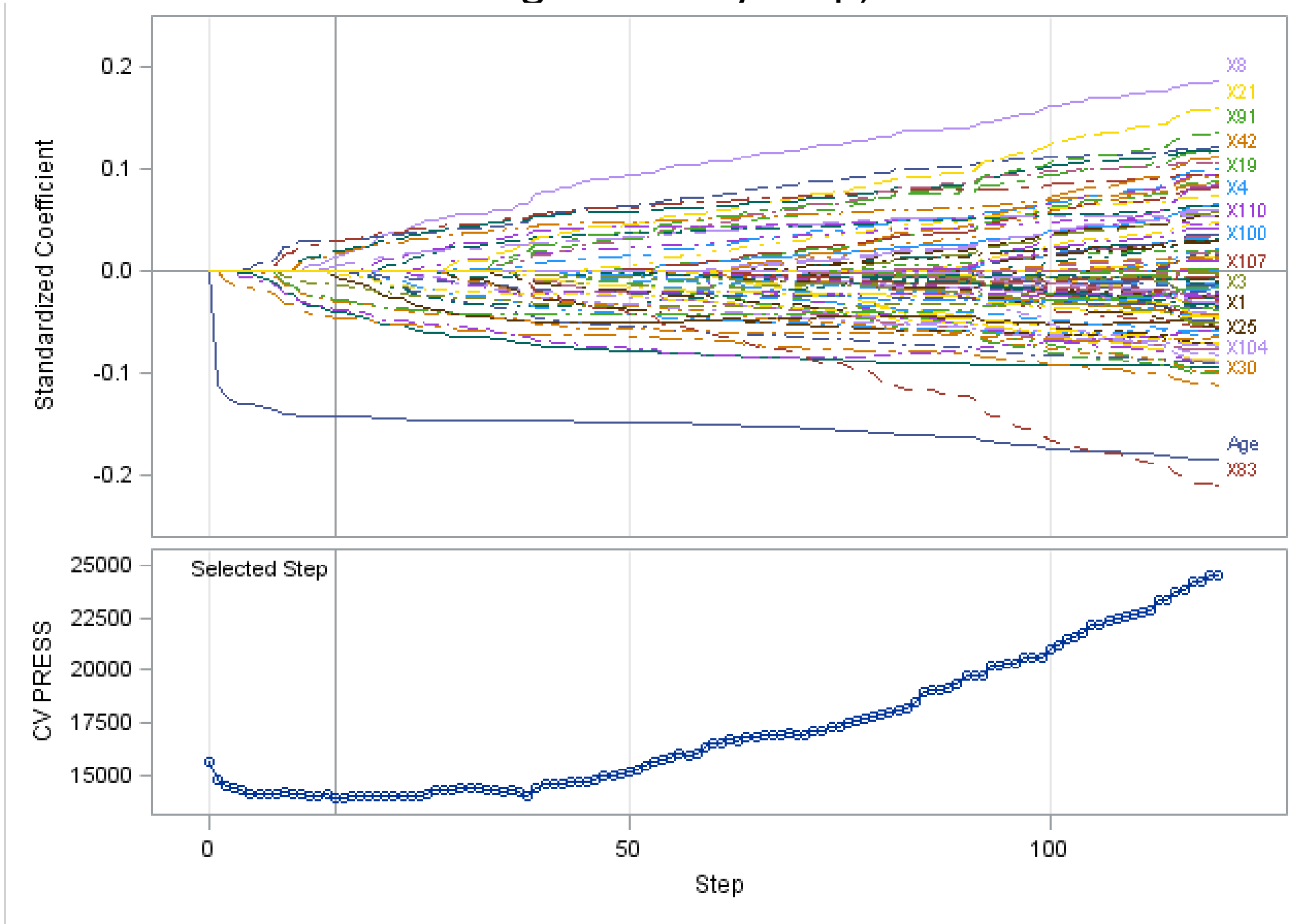
```
ods graphics on;
proc glmselect plots=all;
model Mistake_reaction = Age  Gender Accuracy  X1 – X117
  / CVMETHOD=RANDOM(5)  selection=LASSO (stop=none choose=CV) ;
run;
ods graphics off;

/* Later we will use (stop=30 choose=CV)  */
/*It's handy to turn ODS graphics on and off before and after each
proc glmselect.  These will not be shown in the next slides. */
```


Fit Criteria with STOP=none
Best Criterion Value
Step Selected by CV PRESS
AIC
AICC
SBC
Adj R-Sq
CV PRESS
0
50
100
Step
0
50
100
Step
Best Criterion Value
Step Selected by CV PRESS

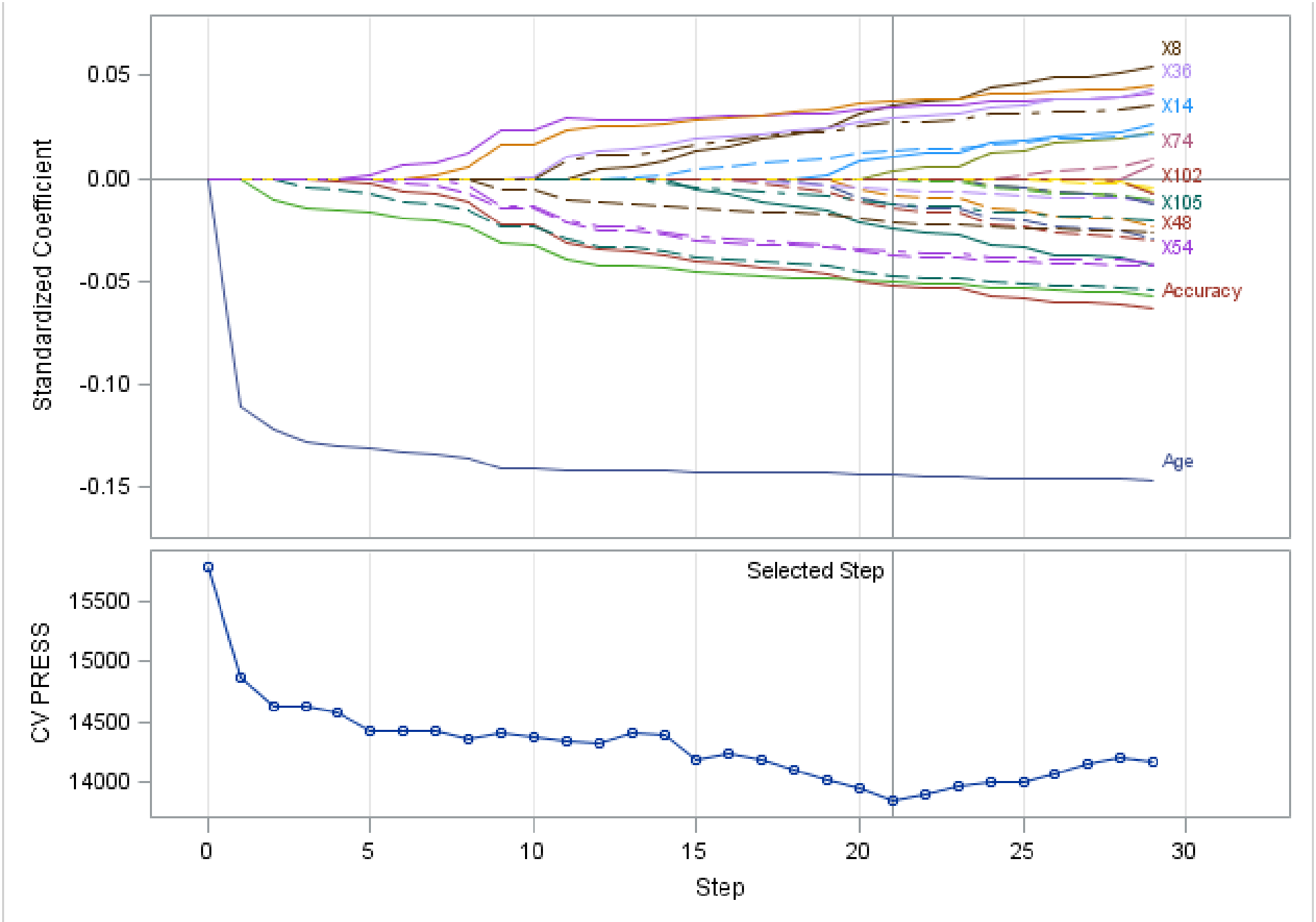
Fit Criteria with STOP=30



Coefficient Progression by Step; STOP=none



Coefficient Progression by Step; STOP=30



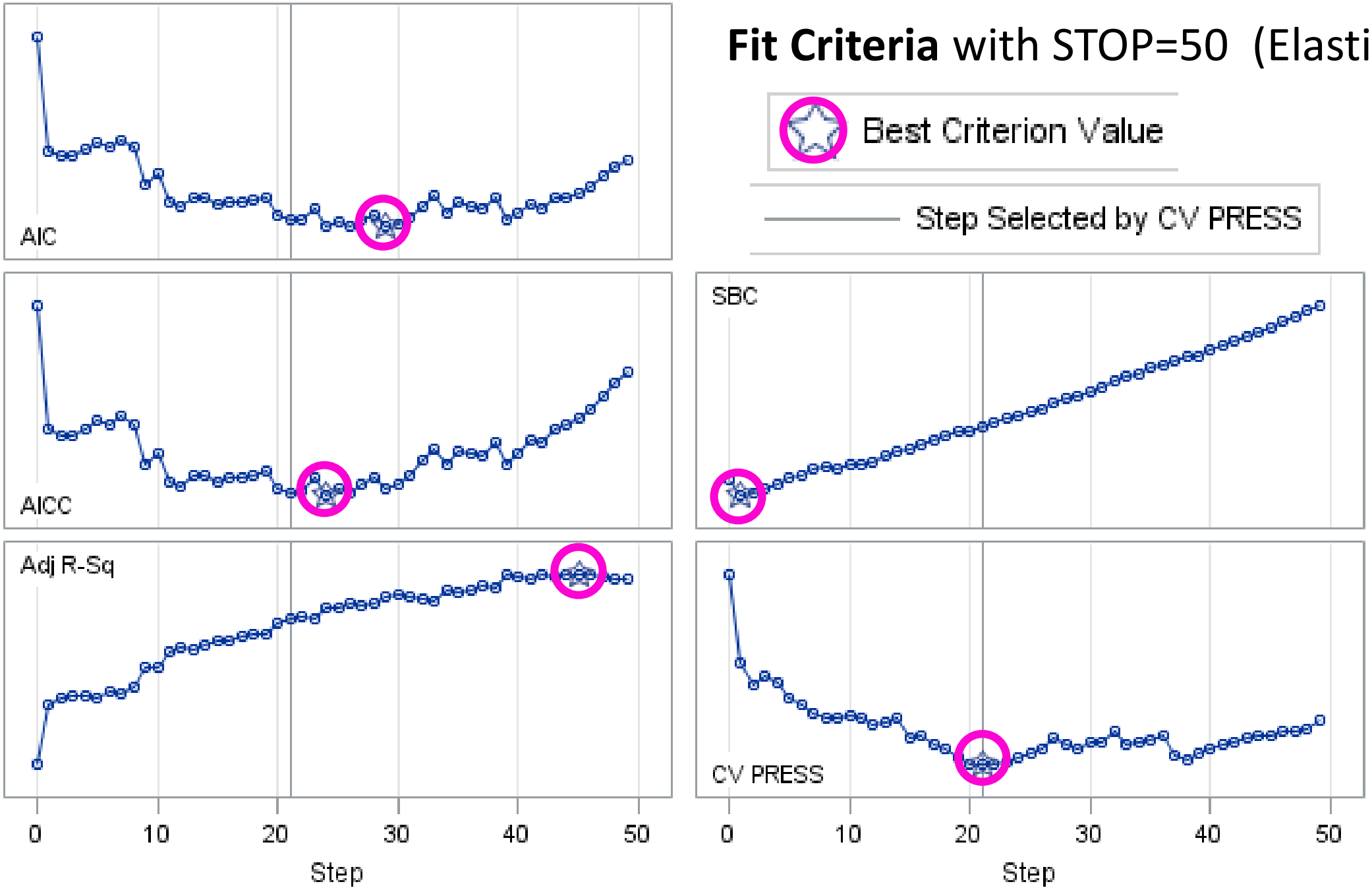
LASSO vs. Elastic Net

- LASSO
  - More parsimonious
  - Chooses only one of two or more correlated variables

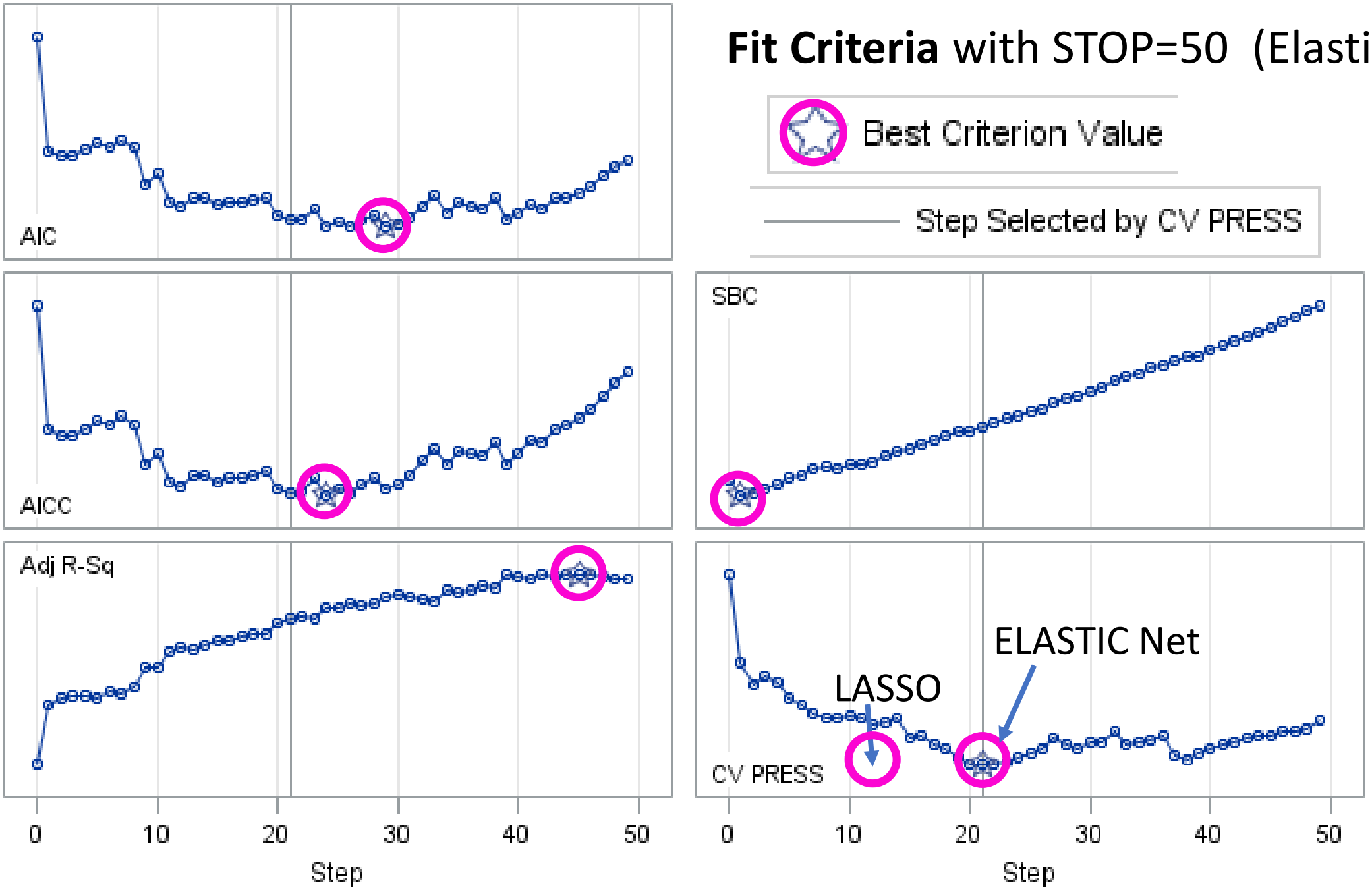
- Elastic Net
  - Allows more variables to enter
  - Allows correlated variables to both/all enter

```
proc glmselect plots=all;  /* ElasticNet with options  */
model Mistake_reaction = Age Gender Accuracy  X1 -- X117
    / CVMETHOD=RANDOM selection=elasticnet (stop=50 choose=CV);
 run;
```

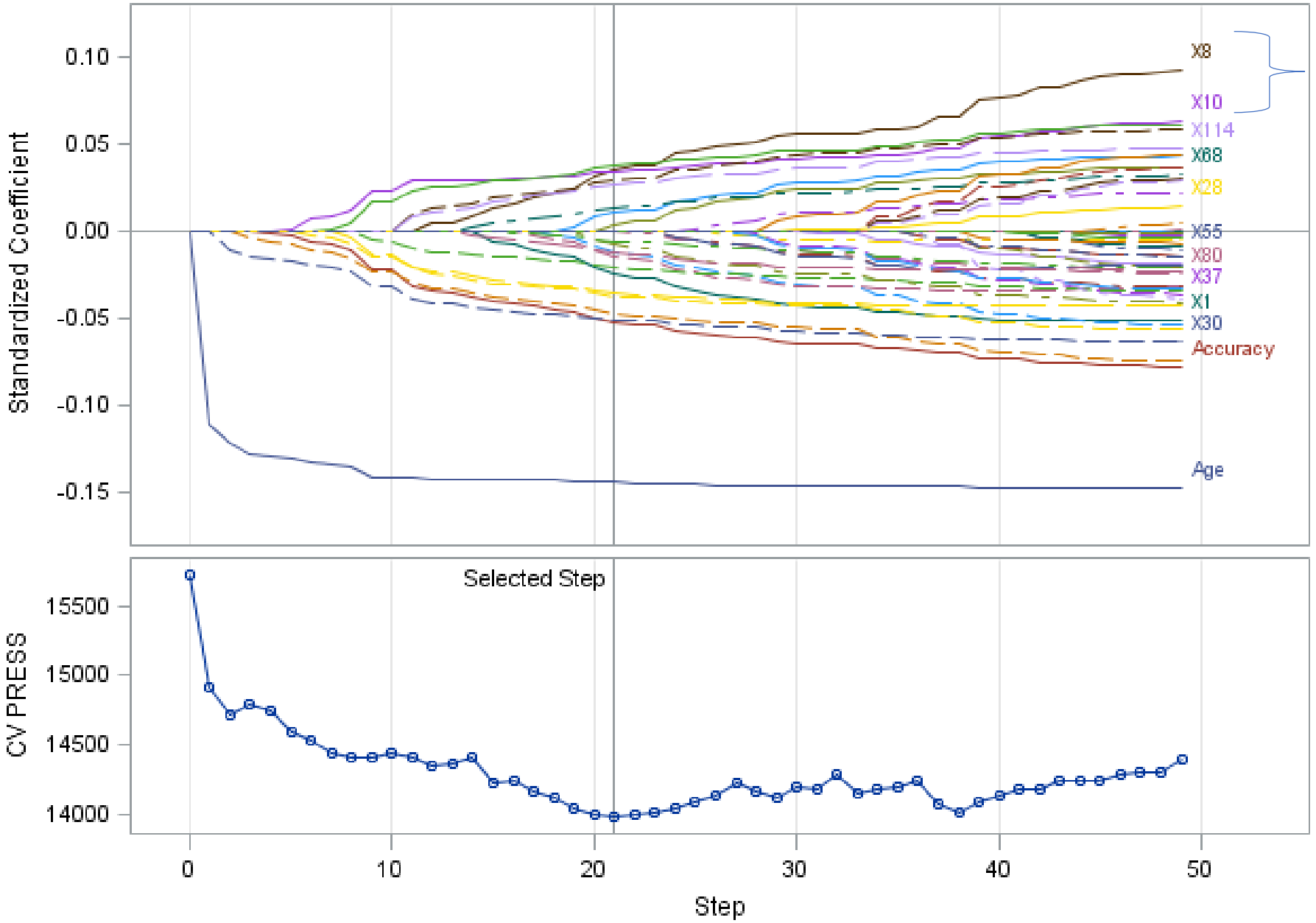
Fit Criteria with STOP=50 (ElasticNet)



Fit Criteria with STOP=50 (ElasticNet)

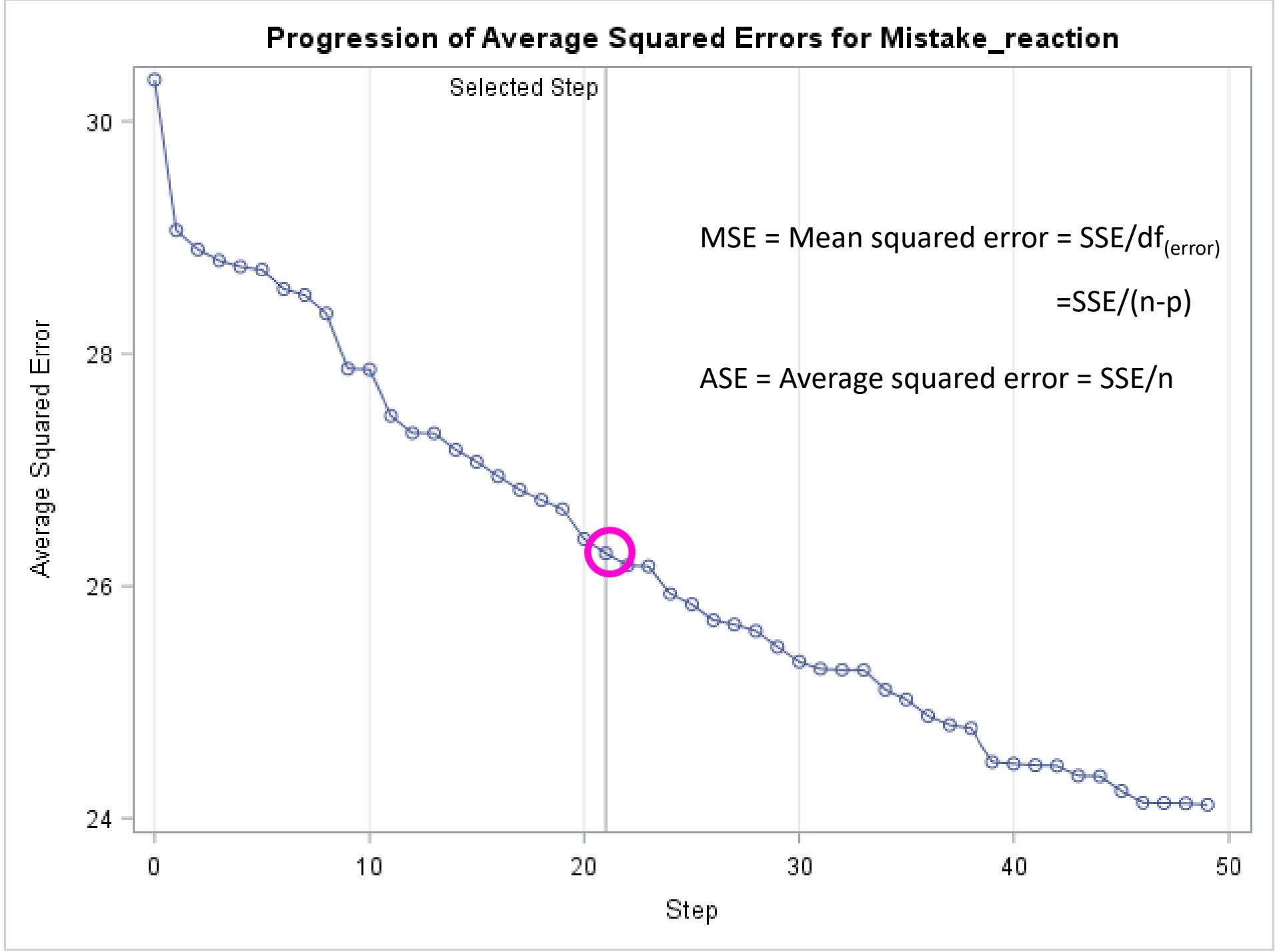


Coefficient Progression by Step; STOP=50 (Elastic Net)



Corr(X8,X10)=0.60
Both are in the Elastic Net model.

In the LASSO model, only X8 was included.


Progression of Average Squared Errors for Mistake_reaction
Selected Step
30
28
26
24
Average Squared Error
0
10
20
30
40
50
Step
MSE = Mean squared error = SSE/df(error)
=SSE/(n-p)
ASE = Average squared error = SSE/n

Things still to try

- Add interactions (selectively or, say, all pairwise)

- Compare prediction error (based on cross-validation) from different models

LASSO vs. Elastic Net

As stated by Zou and Hastie (2005), **the elastic net method can overcome the limitations of LASSO in the following three scenarios**:

- When **p > n, the LASSO method selects at most n variables**
  - The **elastic net method can select more than n variables** because of the ridge regression regularization.

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

- In the **n > p** case, if there are **high correlations between predictors**, 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**.
  - i.e., Carefully select the coefficient for $\sum_{j=1}^{p} \beta_j^2$

Conclusion

- GLMSelect has multiple options and graphics for model selection to optimize prediction while preventing over-fitting
  - If you don’t have a Test or Validation set (or a large sample size to partition off such a set), using k-fold cross-validation is recommended to better estimate prediction error
- Sometimes with so many options, it can be difficult to choose.
- Is it better to be conservative (restrict to fewer variables) or anti-conservative (allow more variables in the model)? Or in between?
- In any case, it is useful to try several options to help make a decision.