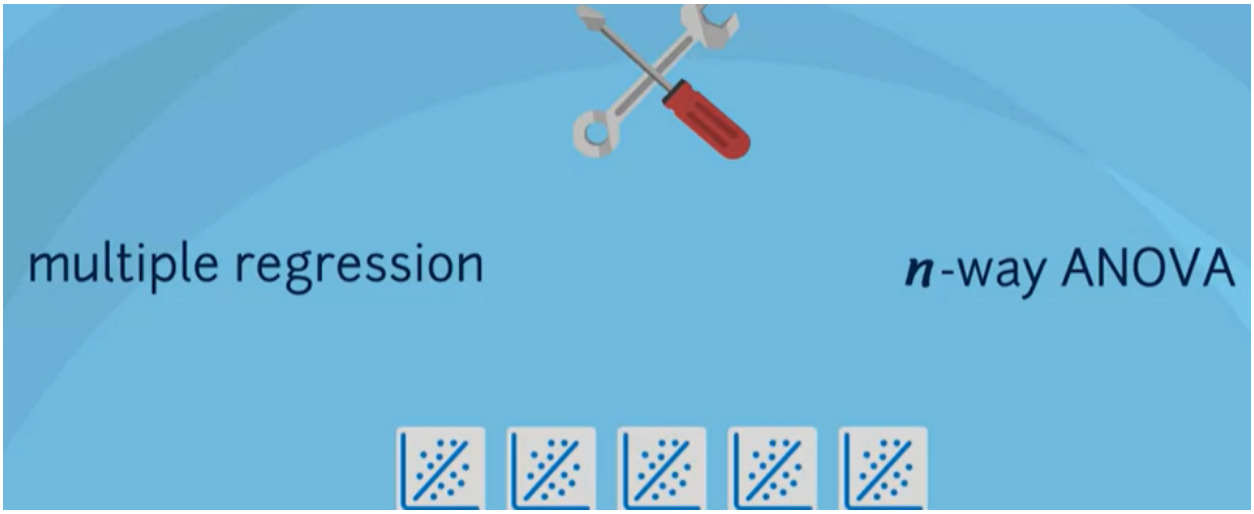
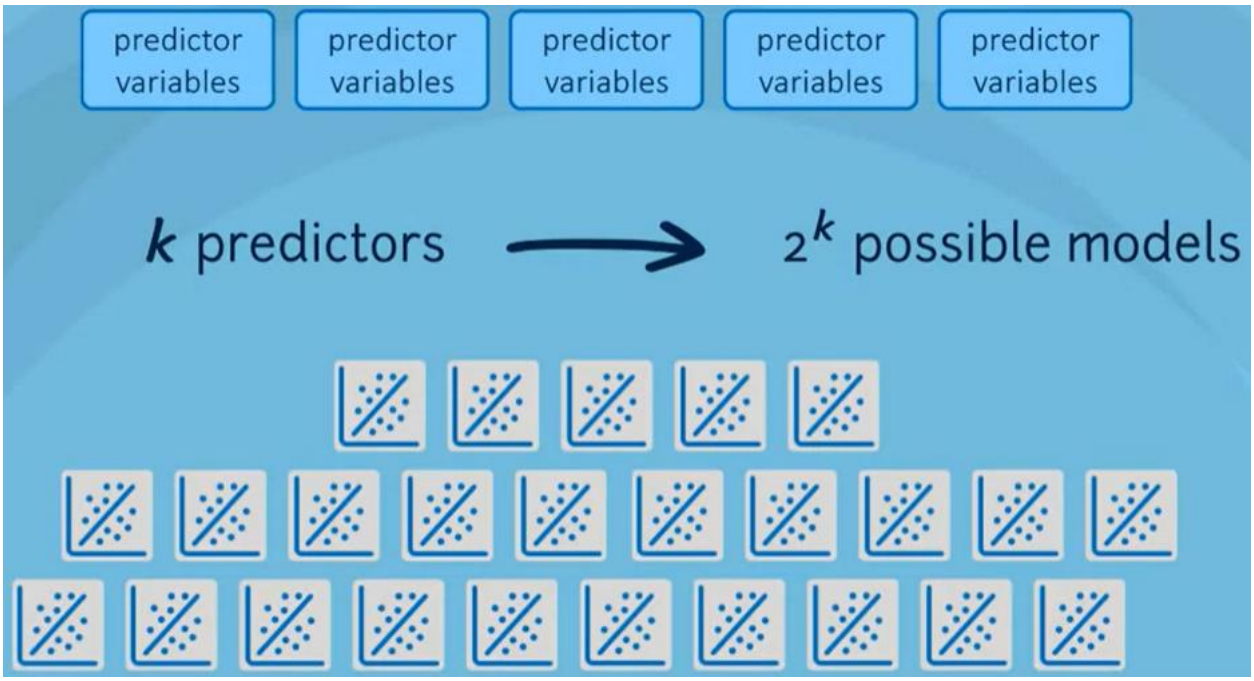
SBA Statistical Business Analyst with SAS

SBA2: Regression Modeling Fundamentals W1

W1 Model Building and Effect Selection

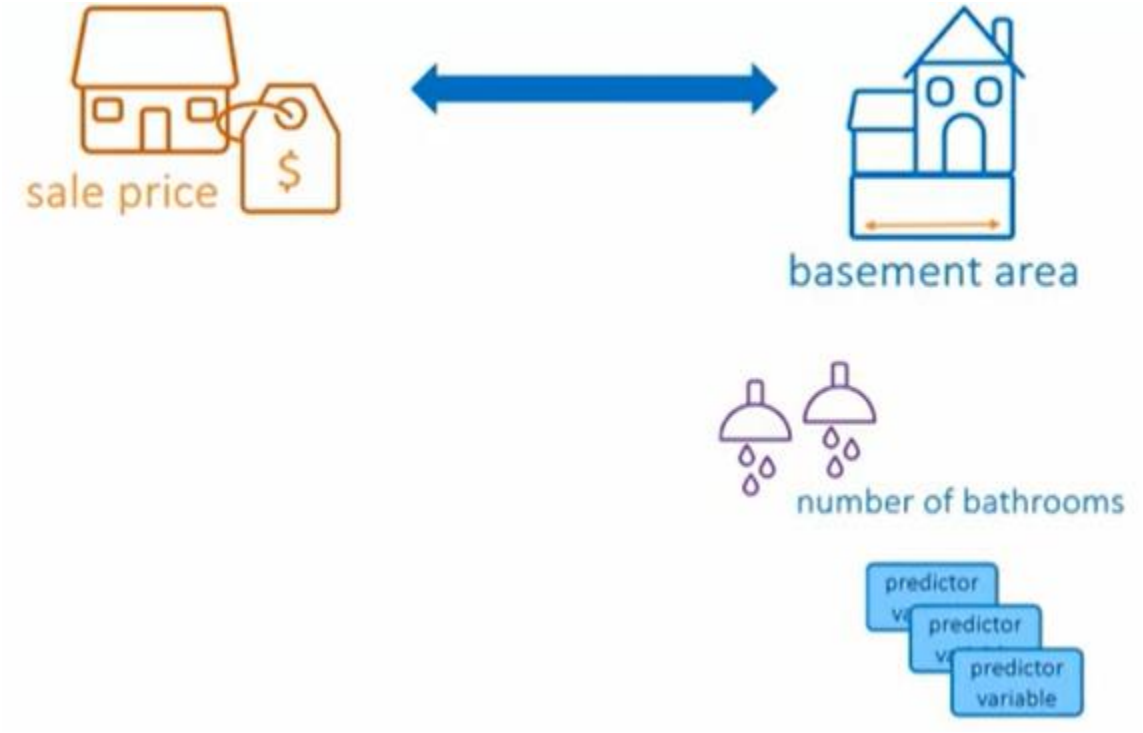
Overview

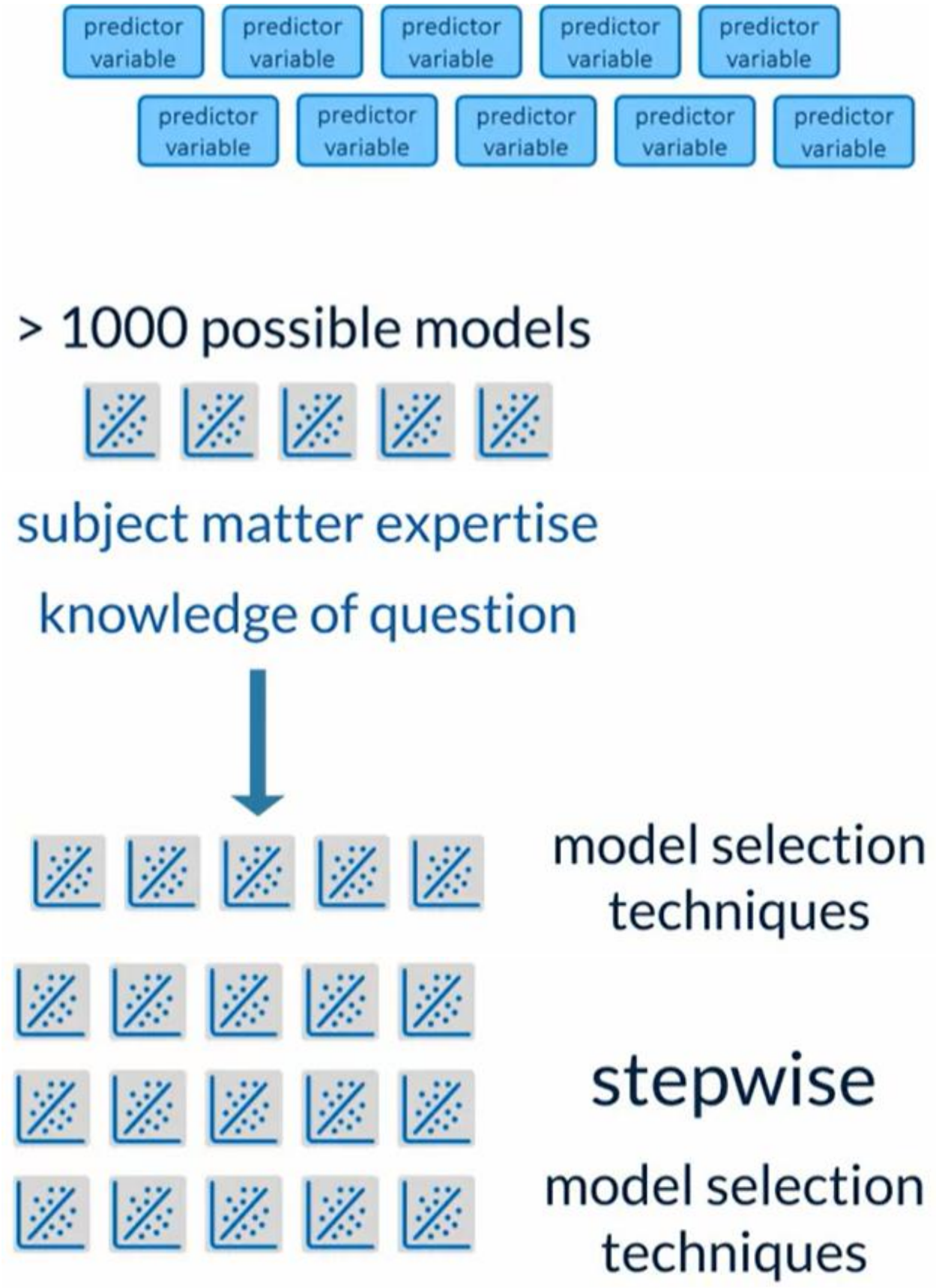




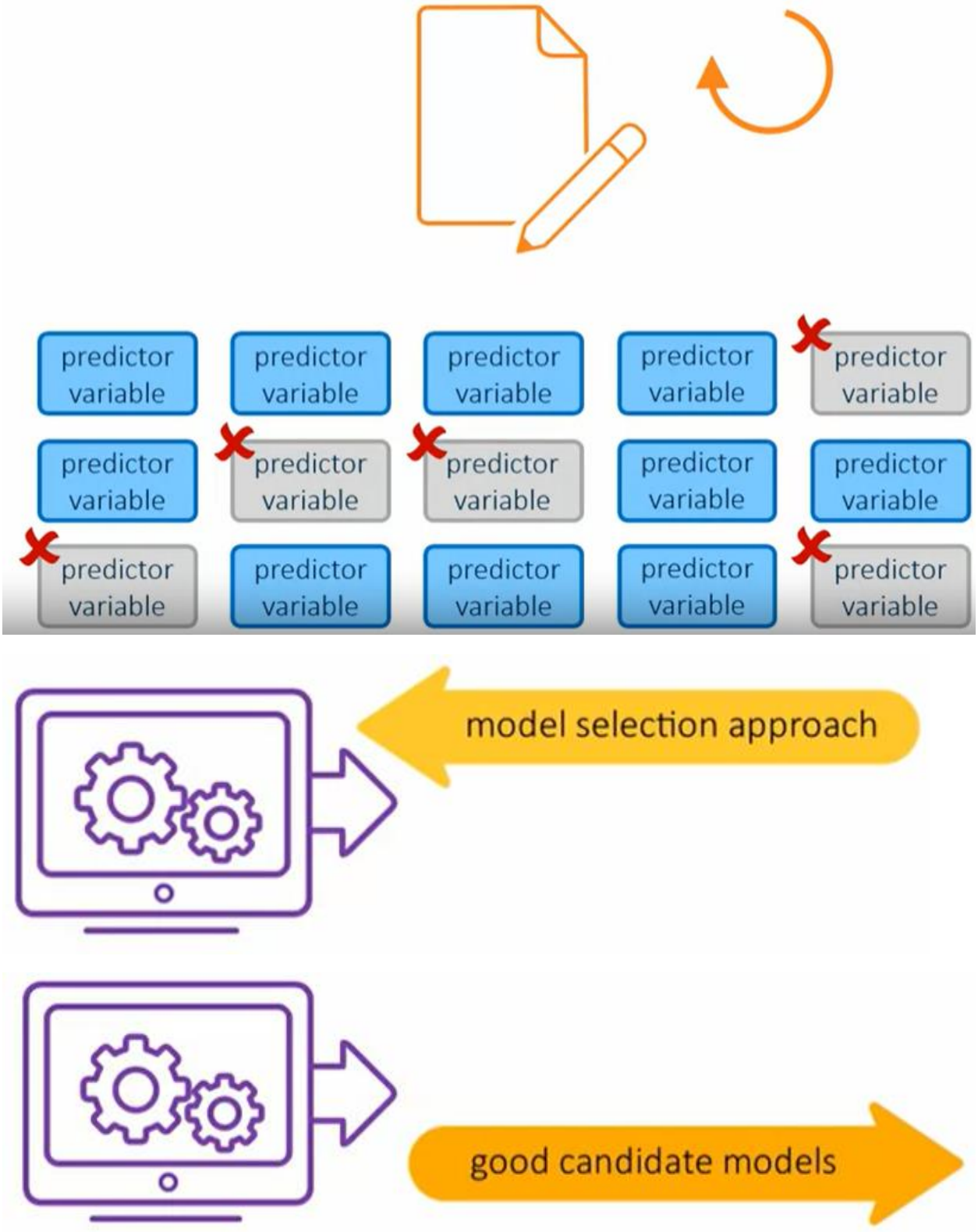
Step Wise Selection Using Significance Level

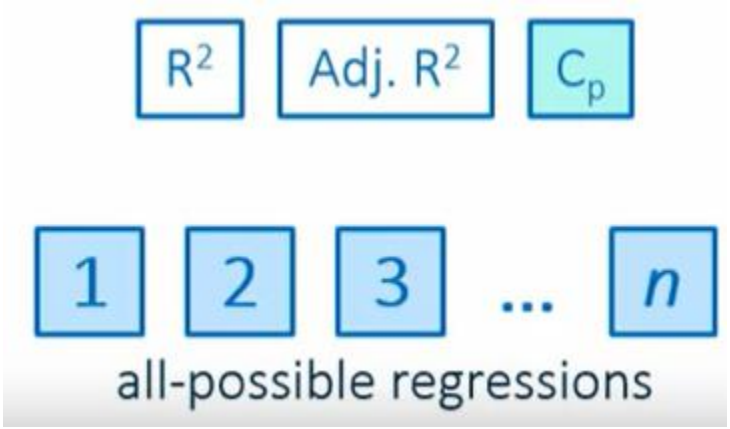
Scenario

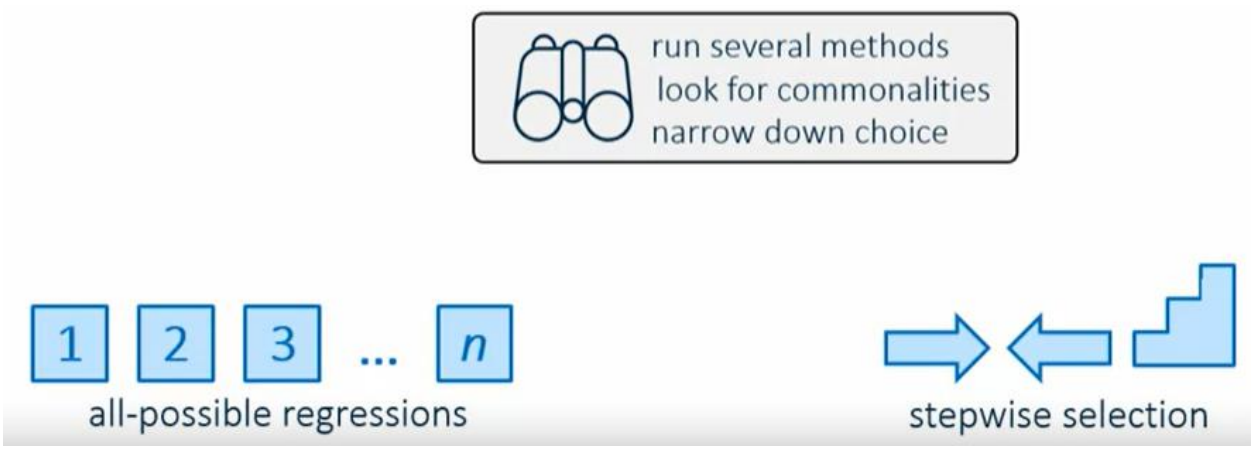



predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
> 1000 possible models
subject matter expertise
knowledge of question
model selection techniques
stepwise
model selection techniques

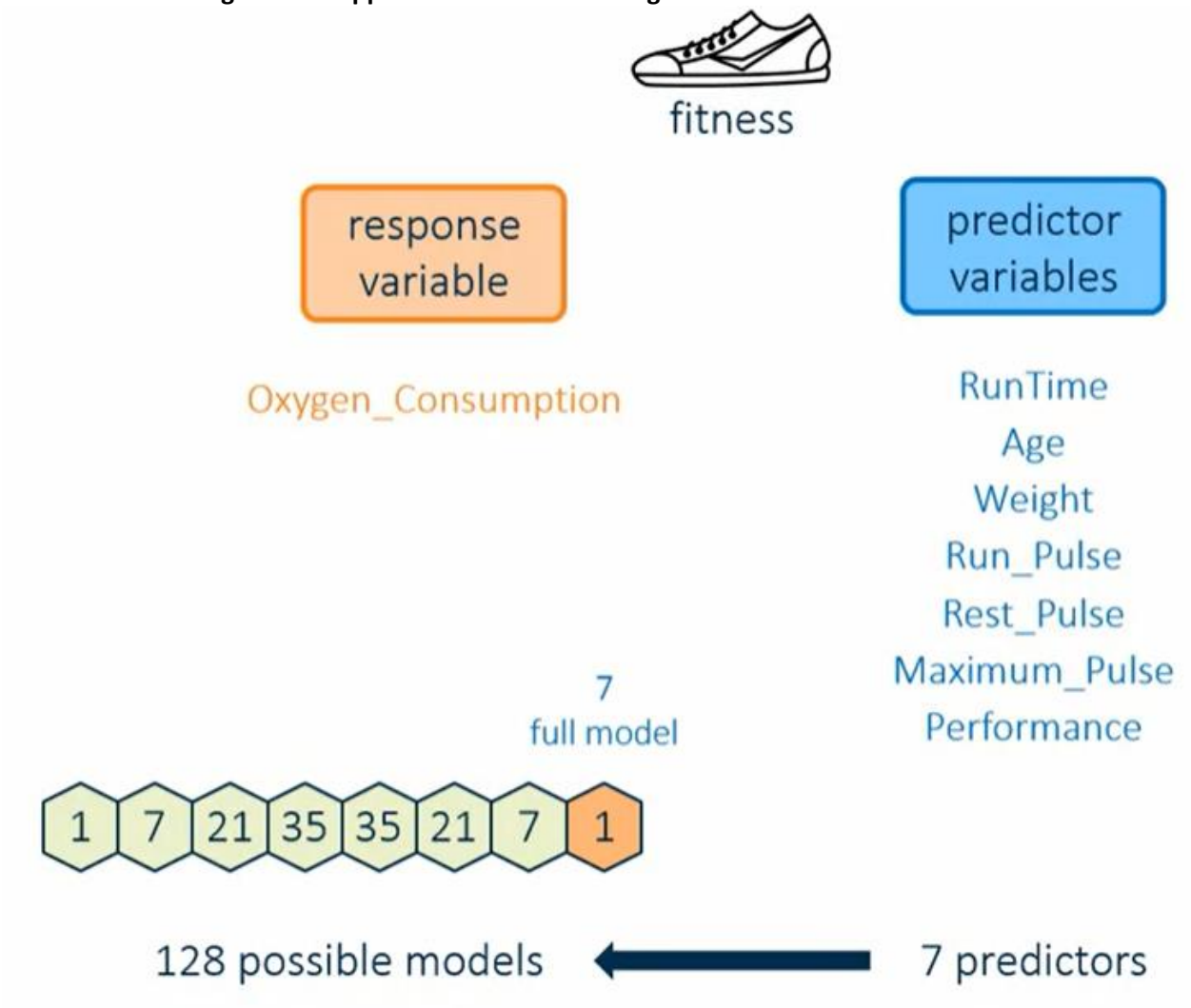
Approaches to Selecting Models

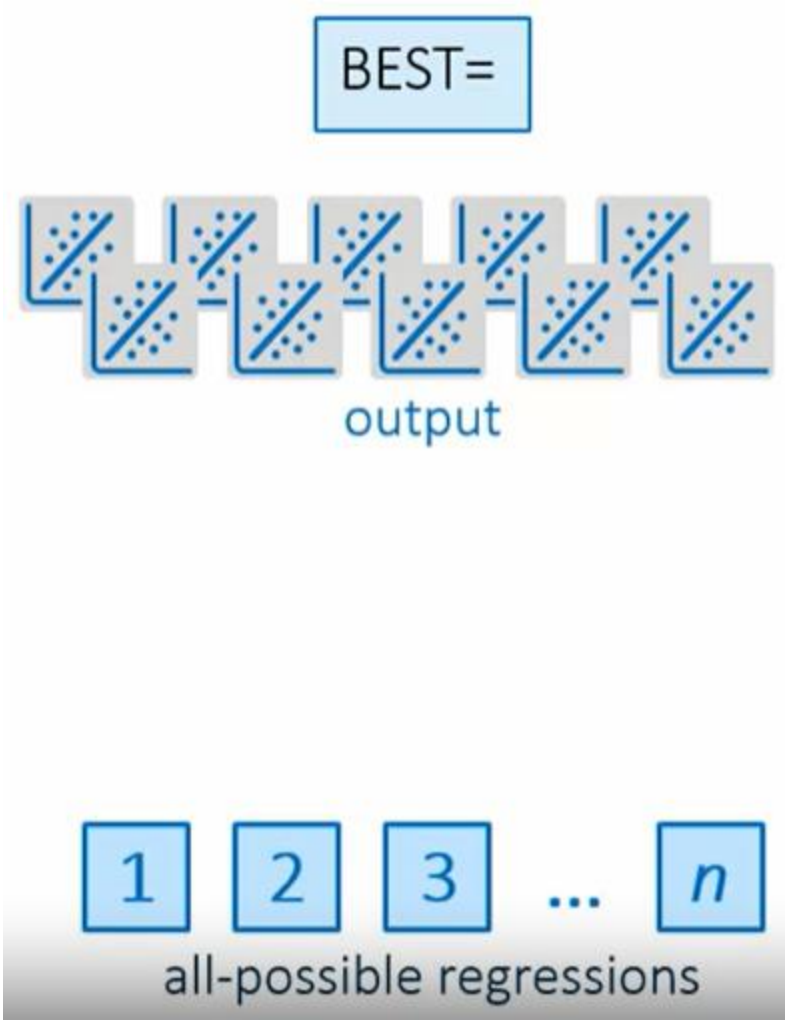



R^2
Adj. R^2
C_p
1
2
3
...
n
all-possible regressions

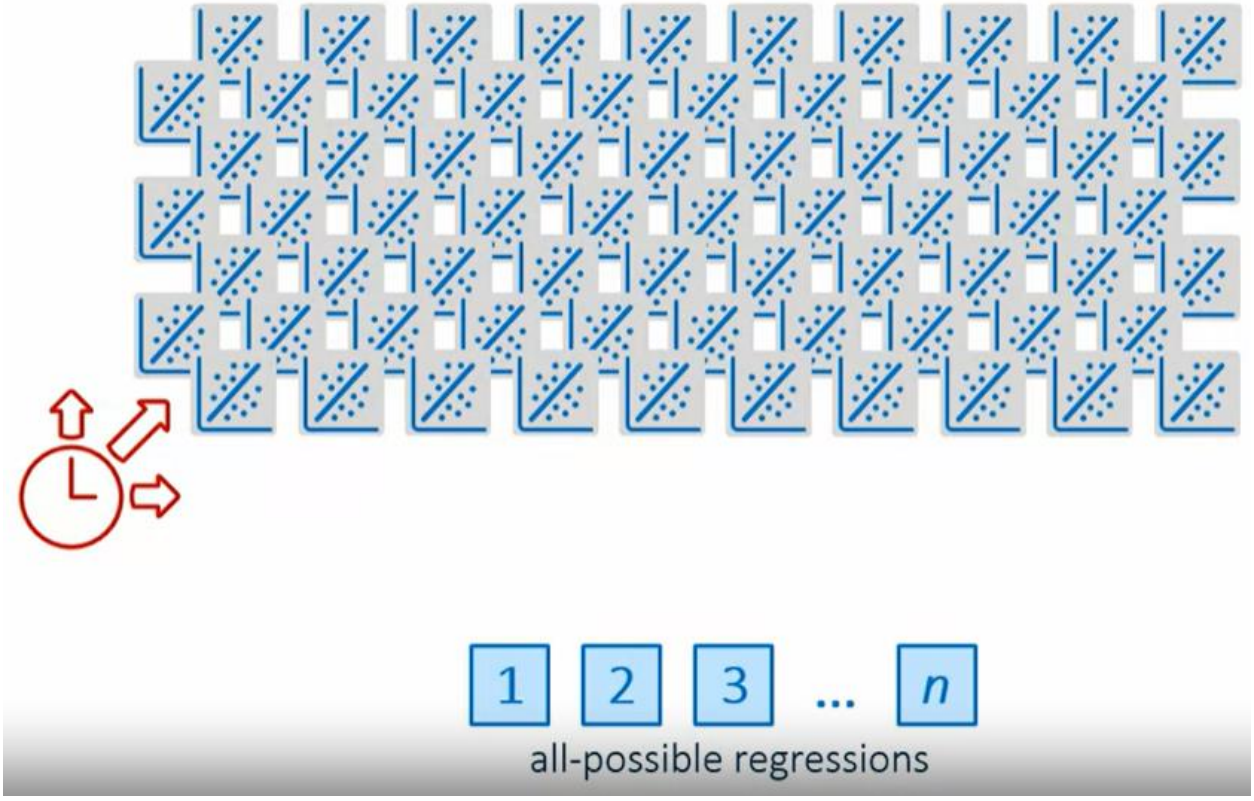

run several methods
look for commonalities
narrow down choice
1
2
3
...
n
all-possible regressions
stepwise selection

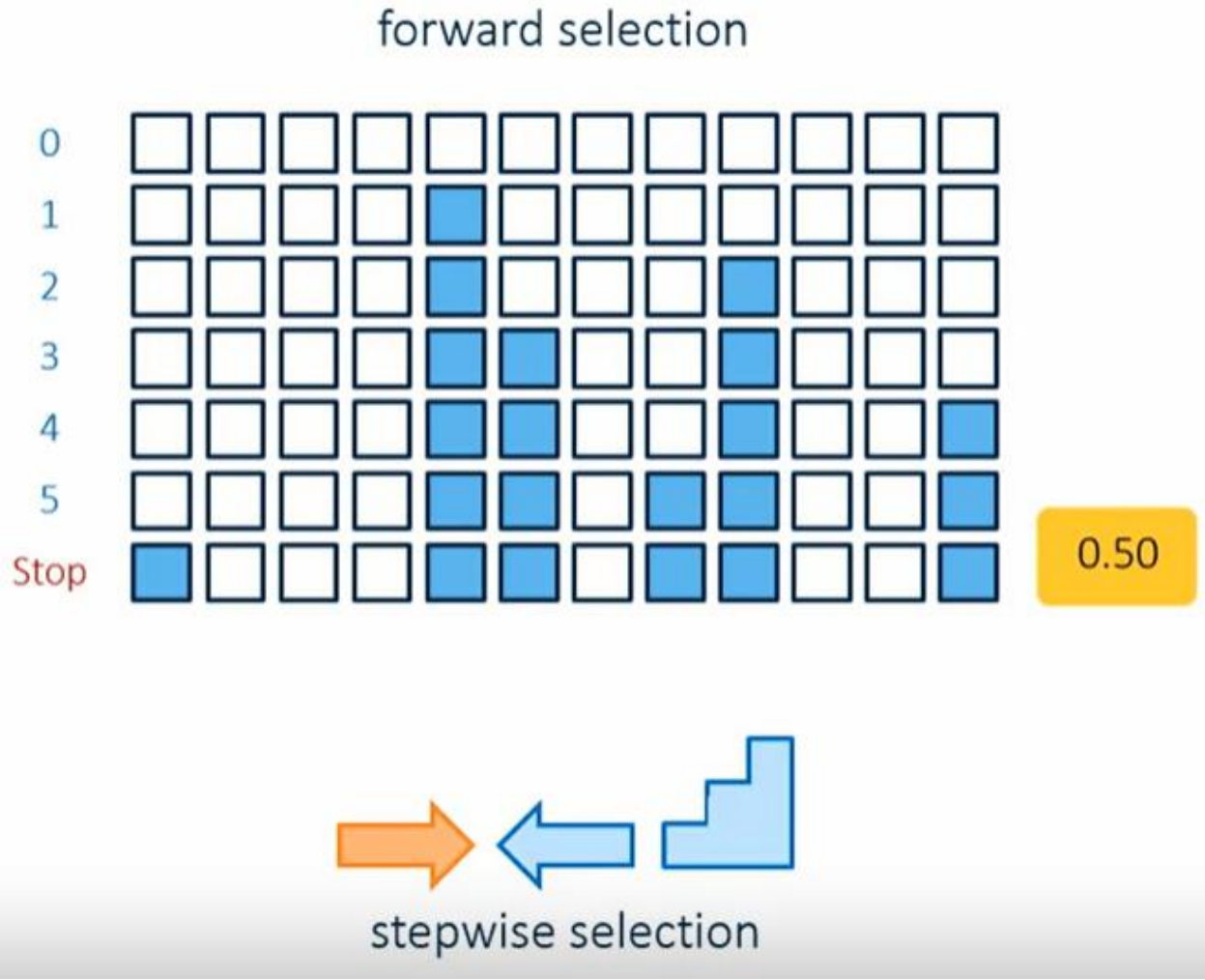
The All-Possible Regressions Approach to Model Building

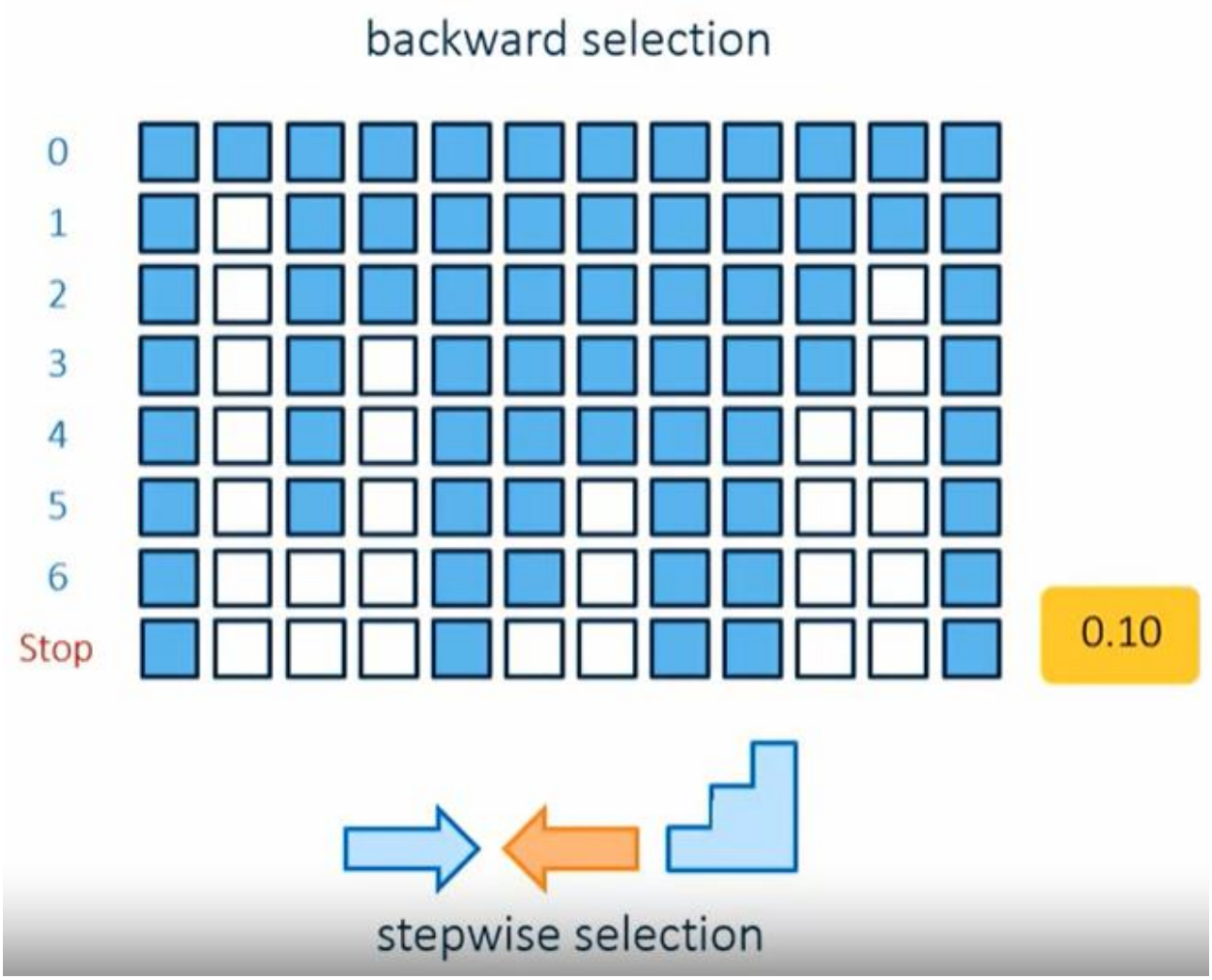



BEST=
output
1
2
3
...
n
all-possible regressions

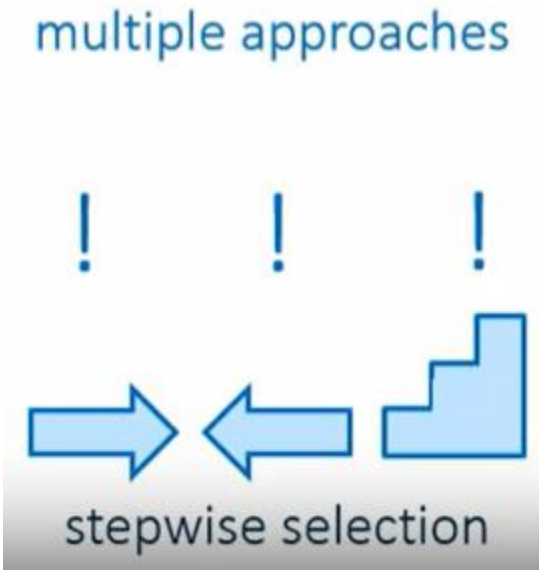
The Step Wise Selection Approach to Model Building



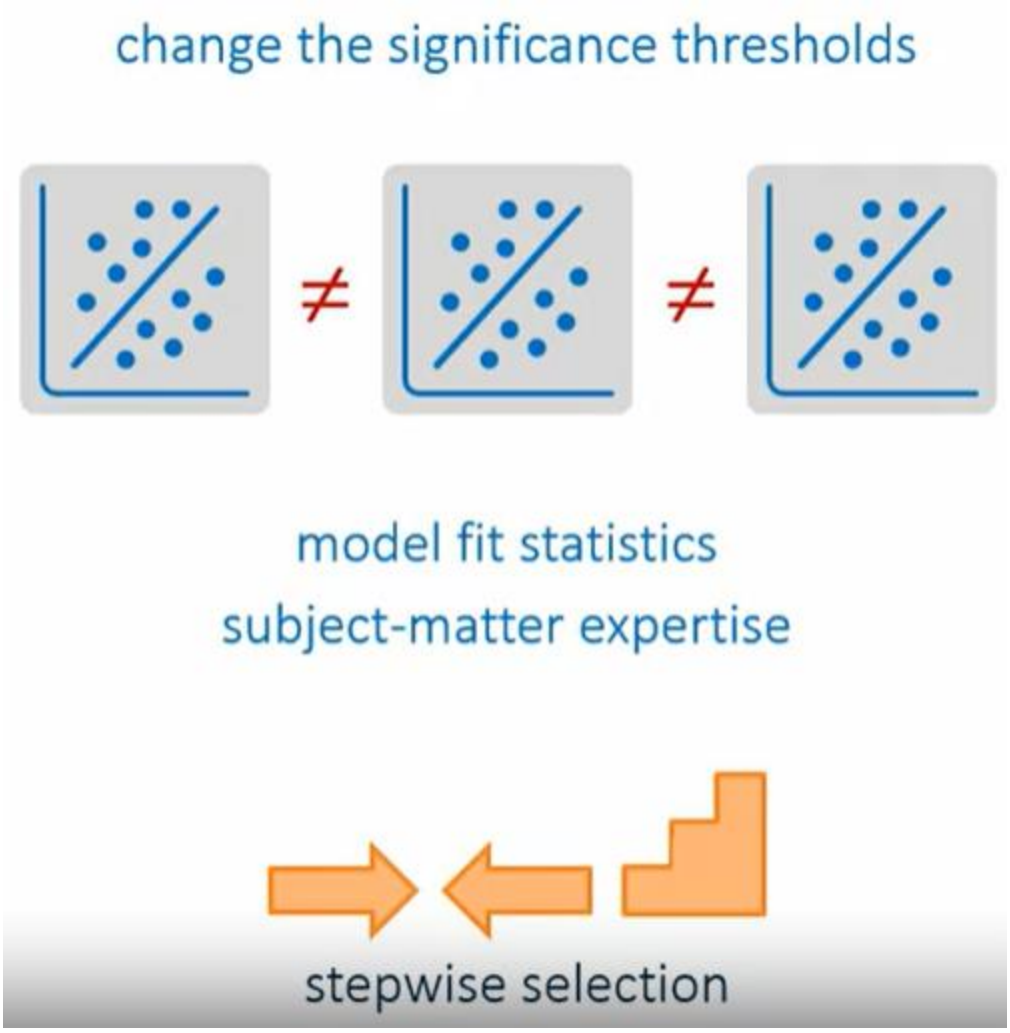

forward selection
0
1
2
3
4
5
Stop
0.50
stepwise selection


backward selection
0
1
2
3
4
5
6
Stop
0.10
stepwise selection

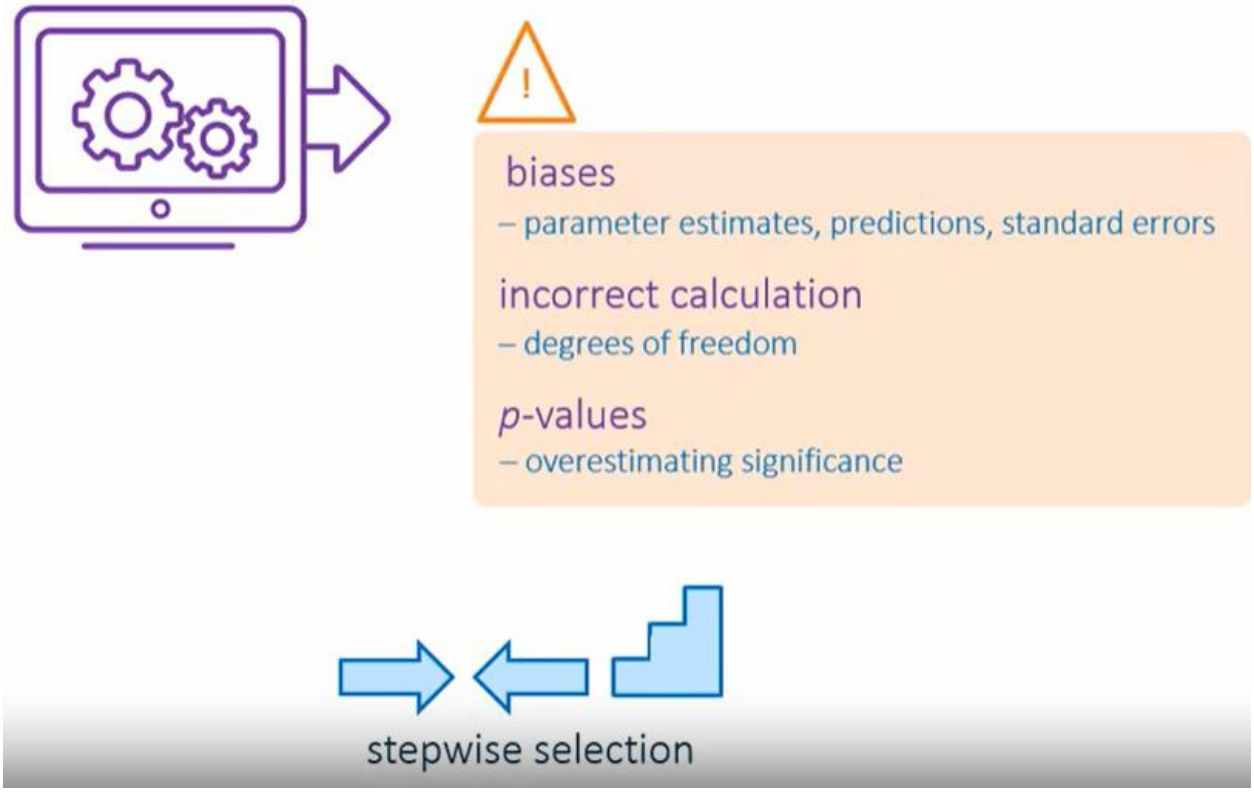

stepwise selection
0
1
2
3
4
5
6
Stop
0.15
stepwise selection

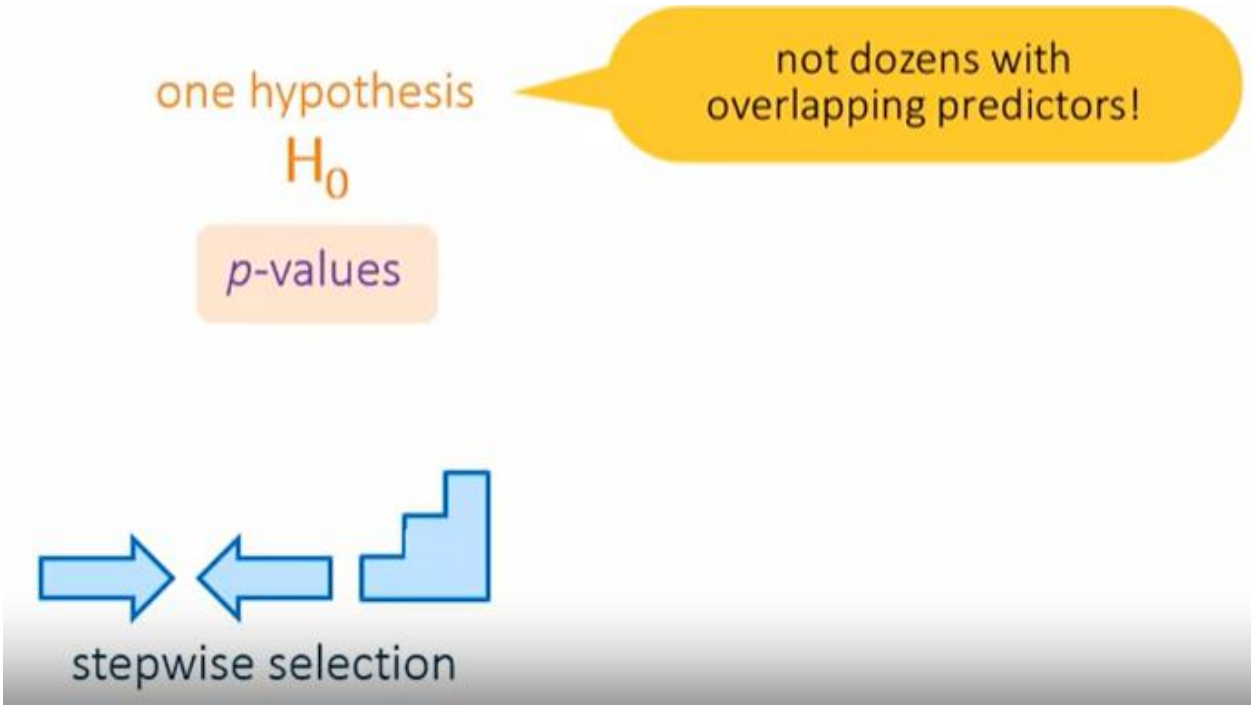

multiple approaches
stepwise selection


stepwise selection


change the significance thresholds
≠
≠
model fit statistics
subject-matter expertise
stepwise selection

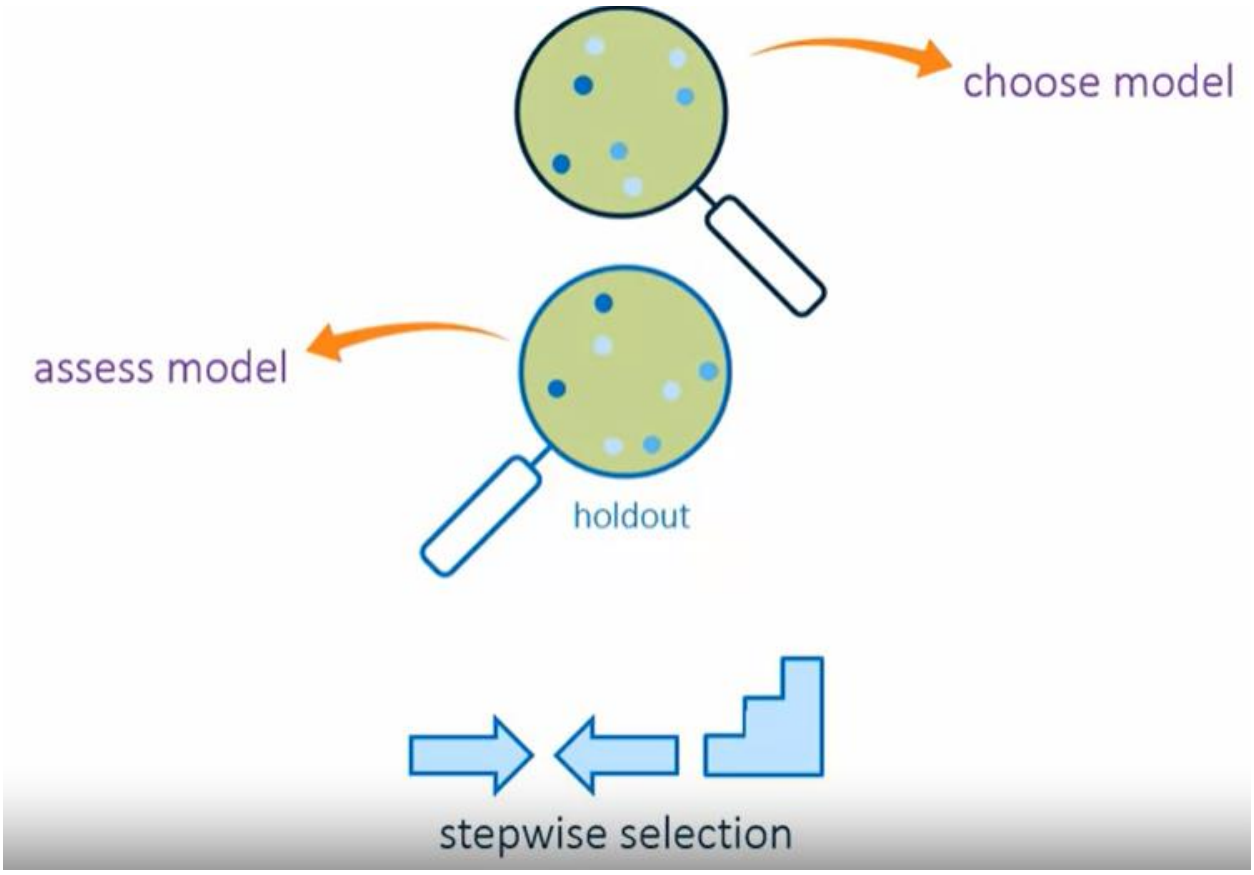
Interpreting p-values and parameter estimates

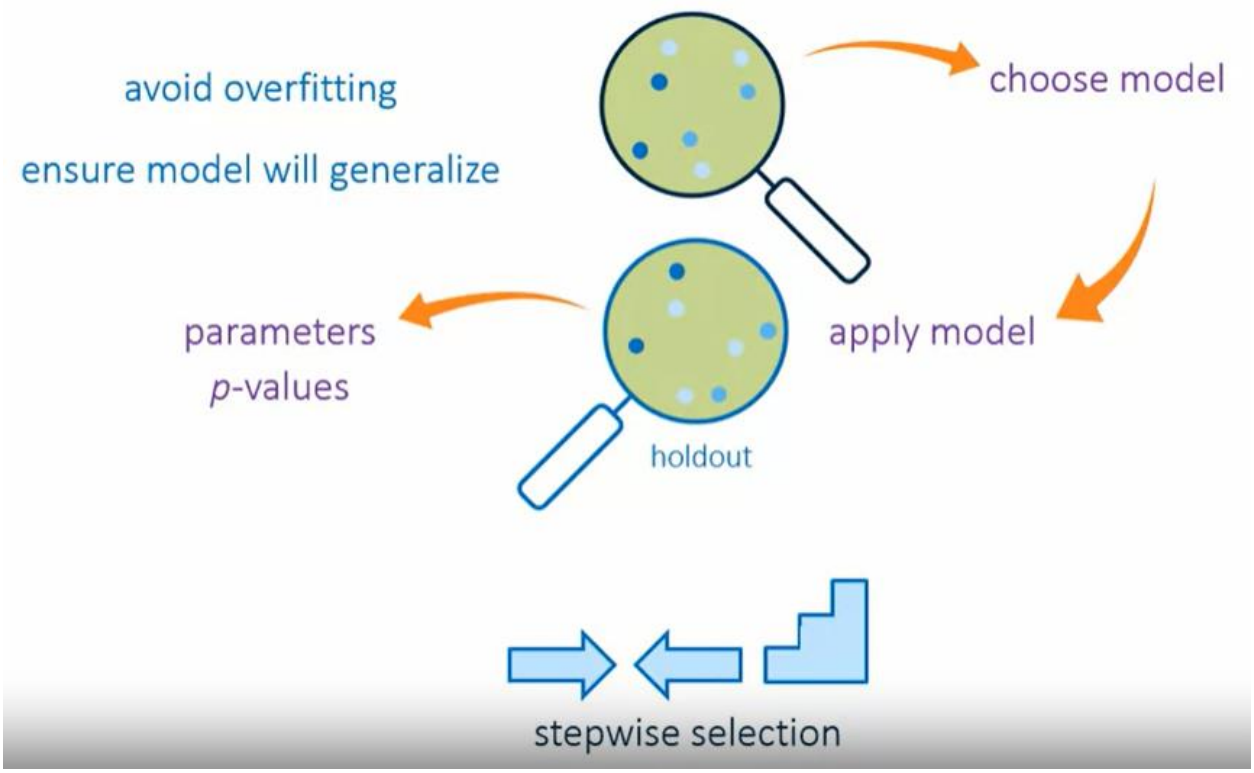




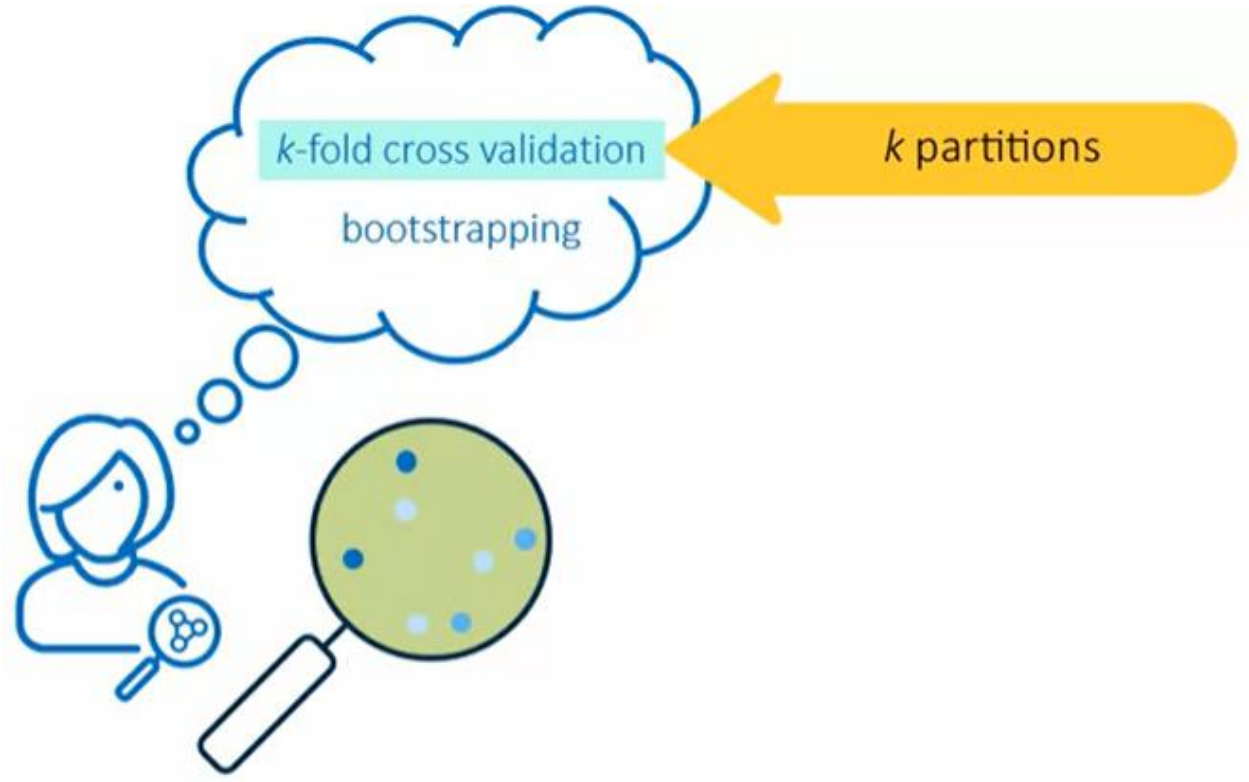
adj. R^2

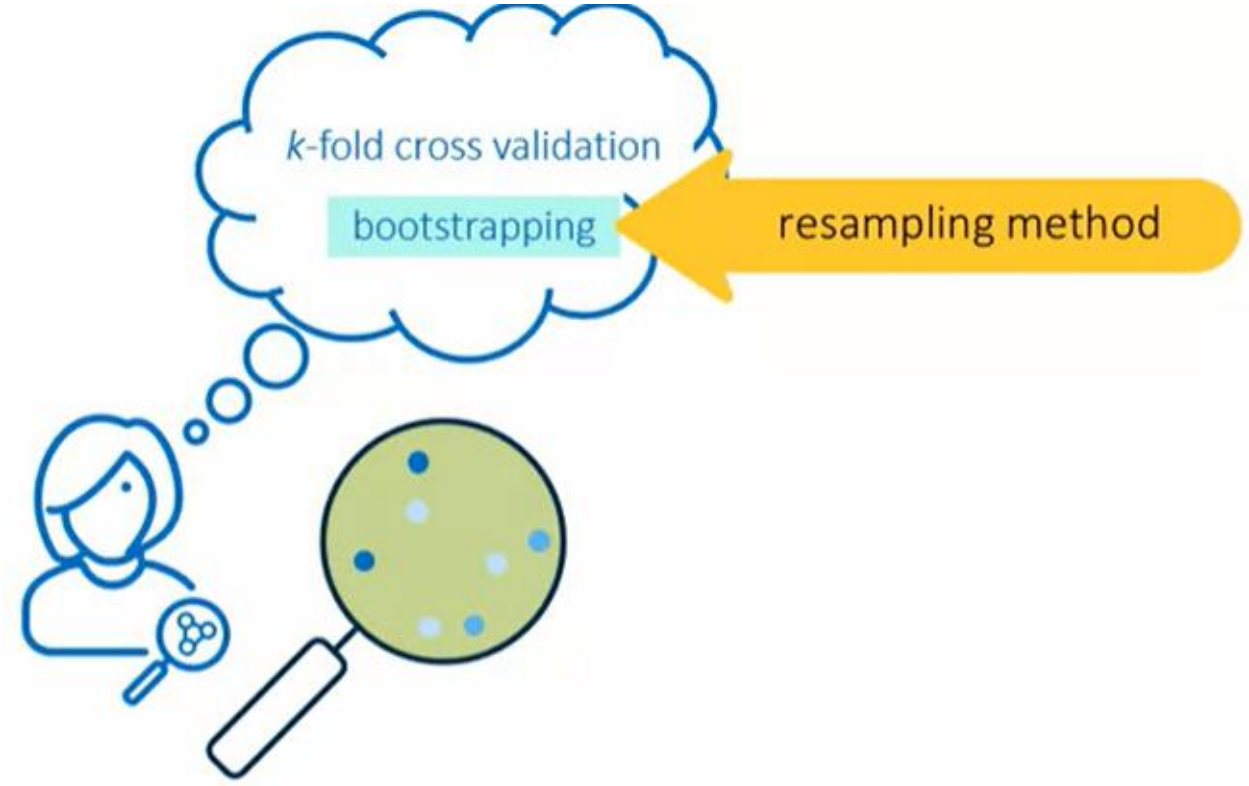
information criteria




avoid overfitting
ensure model will generalize
choose model
parameters
p-values
apply model
holdout
stepwise selection


not
enough
data


k-fold cross validation
bootstrapping
k partitions


k-fold cross validation
bootstrapping
resampling method

Demo Performing Step Wise Regression Using PROC GLMSELECT

```
PROC GLMSELECT DATA=SAS-data-set <options>;
    CLASS variable(s);
    < label: > MODEL dependent = <effects> < / options>;
RUN;
```

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
        Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st104d01.sas*/
ods graphics on;
proc glmselect data=STAT1.ameshousing3 plots=all;
    STEPWISE: model SalePrice = &interval / selection=stepwise details=steps select=SL slstay=0.05 slentry=0.05;
    title "Stepwise Model Selection for SalePrice - SL 0.05";
run;
```

The GLMSELECT Procedure
Stepwise Selection: Step 1

Effect Entered: Basement_Area

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	2.012418E11	2.012418E11	270.16
Error	298	2.219817E11	744904950	
Corrected Total	299	4.232235E11		

Root MSE	
Dependent Mean	
R-Square	
Adj R-Sq	
AIC	6432
AICC	6432
SBC	6138.03102

```
PROC GLM DATA=SAS-data-set <options>;
    CLASS variable(s);
    MODEL dependent-variables=independent-effects < / options>;
    MEANS effects < / options>;
RUN;
```

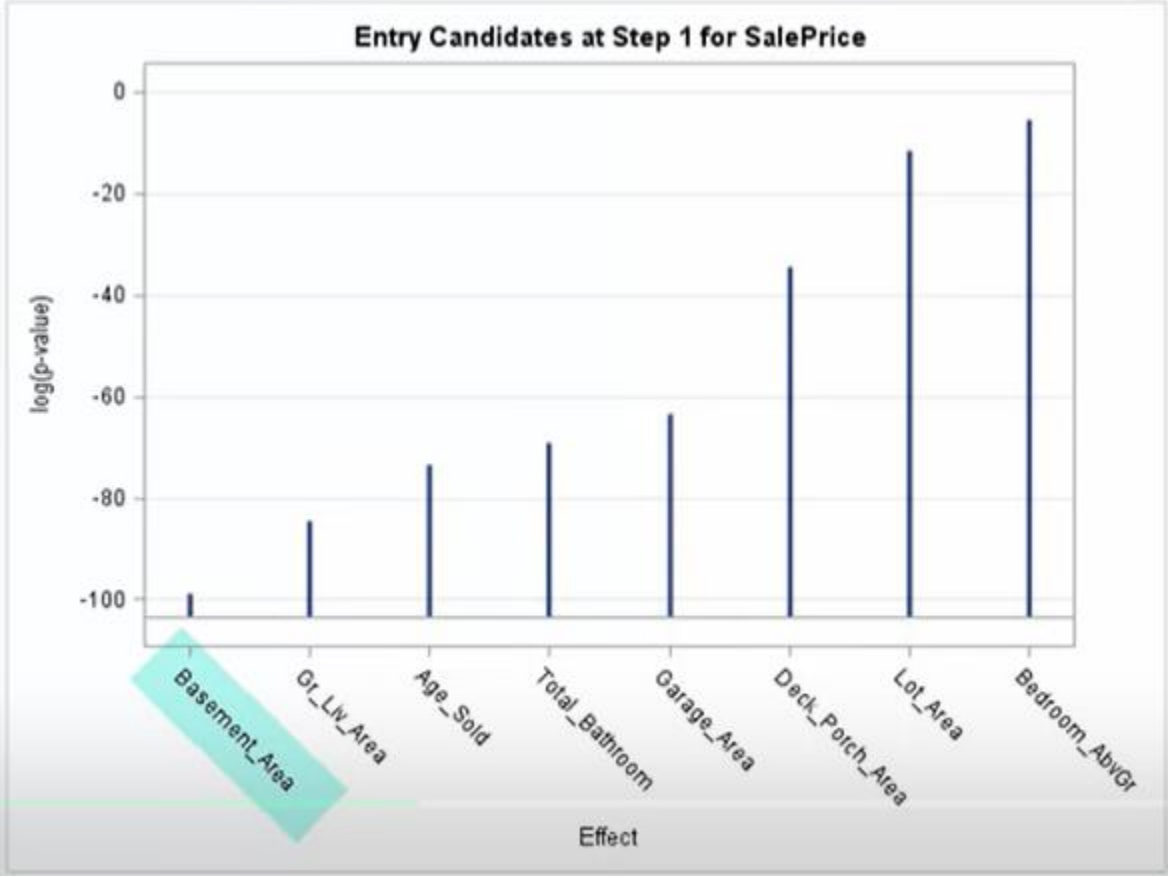
SHOWPVALUES

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	73904	4179.193780	17.68
Basement_Area	1	72.107717	4.387055	16.44

Entry Candidates

Rank	Effect	Log pValue	Pr > F
1	Basement_Area	-98.8577	<.0001
2	Gr_Liv_Area	-84.6132	<.0001
3	Age_Sold	-73.5219	<.0001
4	Total_Bathroom	-69.1880	<.0001
5	Garage_Area	-63.3558	<.0001
6	Deck_Porch_Area	-34.3105	<.0001
7	Lot_Area	-11.6303	<.0001
8	Bedroom_AbvGr	-5.5339	0.0040


Entry Candidates at Step 1 for SalePrice
0
-20
-40
-60
-80
-100
log(p-value)
Basement_Area
Gr_Liv_Area
Age_Sold
Total_Bathroom
Garage_Area
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Effect

The GLMSELECT Procedure
Stepwise Selection: Step 2

Effect Entered: Gr_Liv_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	2.64483E11	1.322415E11	247.42
Error	297	1.587405E11	534479711	
Corrected Total	299	4.232235E11		

Root MSE	23119
Dependent Mean	137525
R-Square	0.6249
Adj R-Sq	0.6224
AIC	6334.02620
AICC	6334.16179
SBC	6043.13755

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	12664	6650.339855	1.90
Gr_Liv_Area	1	69.606974	6.399091	10.88
Basement_Area	1	52.309702	4.137885	12.64

Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Stepwise Selection: Step 3

Effect Entered: Age_Sold

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	3.207148E11	1.069049E11	308.69
Error	296	1.025087E11	346313132	
Corrected Total	299	4.232235E11		

Root MSE	18609
Dependent Mean	137525
R-Square	0.7578
Adj R-Sq	0.7553
AIC	6204.82927
AICC	6205.03335
SBC	5917.64440

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	53400	6235.076995	8.56
Gr_Liv_Area	1	68.108646	5.152294	13.22
Basement_Area	1	36.329120	3.559067	10.21
Age_Sold	1	-543.493346	42.651840	-12.74

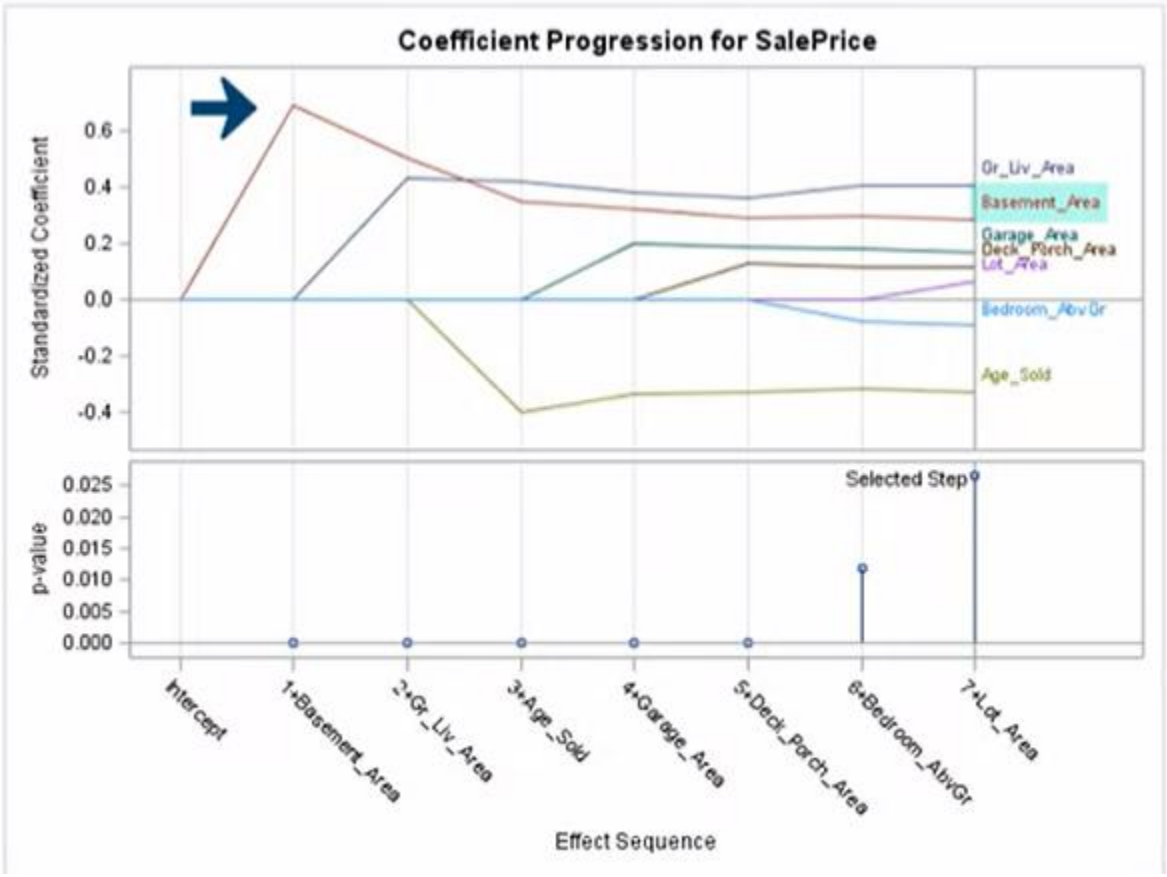
Stepwise Model Selection for SalePrice - SL 0.05

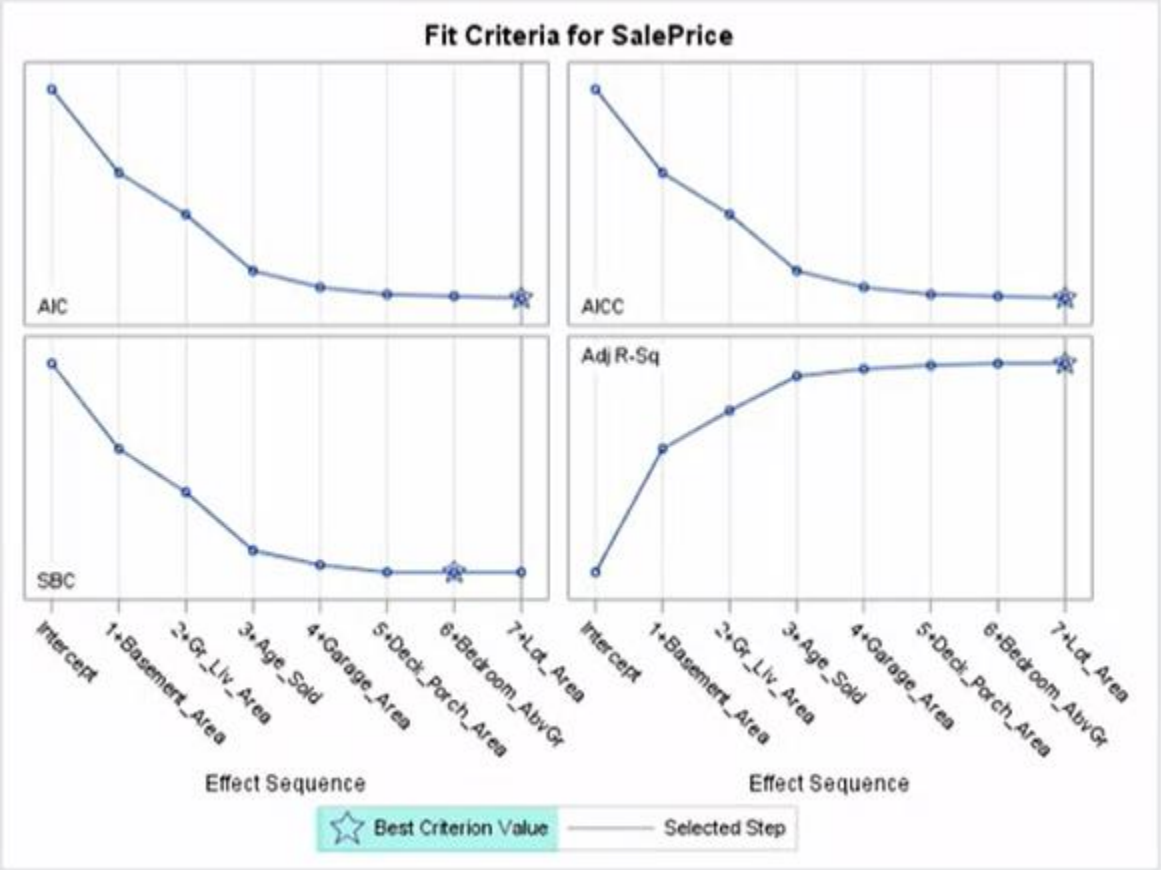
The GLMSELECT Procedure

Stepwise Selection Summary					
Step	Effect Entered	Effect Removed	Number Effects In	F Value	Pr > F
0	Intercept		1	0.00	1.0000
1	Basement_Area		2	270.16	<.0001
2	Gr_Liv_Area		3	118.32	<.0001
3	Age_Sold		4	162.37	<.0001
4	Garage_Area		5	42.30	<.0001
5	Deck_Porch_Area		6	19.99	<.0001
6	Bedroom_AbvGr		7	6.41	0.0119
7	Lot_Area		8	4.97	0.0265

Selection stopped because the candidate for entry has SLE > 0.05 and the candidate for removal has SLS < 0.05.

Stop Details					
Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	Total_Bathroom	0.1167	>	0.0500	(SLE)
Removal	Lot_Area	0.0265	<	0.0500	(SLS)


Coefficient Progression for SalePrice
Standardized Coefficient
0.6
0.4
0.2
0.0
-0.2
-0.4
Gr_Liv_Area
Basement_Area
Garage_Area
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Age_Sold
p-value
0.025
0.020
0.015
0.010
0.005
0.000
Selected Step
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
Effect Sequence


Fit Criteria for SalePrice
AIC
AICC
SBC
Adj R-Sq
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
Effect Sequence
Effect Sequence
Best Criterion Value
Selected Step

Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 7).

Effects:	Intercept Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr

Effects: Intercept Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	7	3.424508E11	48921543221	176.86
Error	292	80772716963	276618894	
Corrected Total	299	4.232235E11		

Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	47463	5880.674041	8.07
Gr_Liv_Area	1	65.303724	5.436672	12.01
Basement_Area	1	29.849078	3.345400	8.92
Garage_Area	1	36.309606	6.452405	5.63
Deck_Porch_Area	1	32.052554	7.967677	4.02
Lot_Area	1	0.708127	0.317512	2.23
Age_Sold	1	-447.199882	41.019314	-10.90
Bedroom_AbvGr	1	-5042.766498	1687.928168	-2.99

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
      Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st104d01.sas*/
ods graphics on;
proc glmselect data=STAT1.ameshousing3 plots=all;
       STEPWISE: model SalePrice = &interval / selection=stepwise details=steps select=SL slstay=0.05
       slentry=0.05;
       title "Stepwise Model Selection for SalePrice - SL 0.05";
run;


/*Optional Code that will execute forward and backward selection
  Each with slentry and slstay = 0.05.

proc glmselect data=STAT1.ameshousing3 plots=all;
       FORWARD: model SalePrice = &interval / selection=forward details=steps select=SL slentry=0.05;
       title "Forward Model Selection for SalePrice - SL 0.05";
run;

proc glmselect data=STAT1.ameshousing3 plots=all;
```

```
        BACKWARD: model SalePrice = &interval / selection=backward details=steps select=SL slstay=0.05;
        title "Backward Model Selection for SalePrice - SL 0.05";
run;
*/
```

Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure

Data Set	STAT1.AMESHOUSING3
Dependent Variable	SalePrice
Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.05
Stay Significance Level (SLS)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	300
Number of Observations Used	300

Dimensions	
Number of Effects	9
Number of Parameters	9

Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Stepwise Selection: Step 0

Effect Entered: Intercept

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	0	0	.	.
Error	299	4.232235E11	1415463276	
Corrected Total	299	4.232235E11		

Root MSE	37623
Dependent Mean	137525
R-Square	0.0000
Adj R-Sq	0.0000
AIC	6624.21515
AICC	6624.25555
SBC	6325.91893

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	137525	2172.144314	63.31

Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Stepwise Selection: Step 1

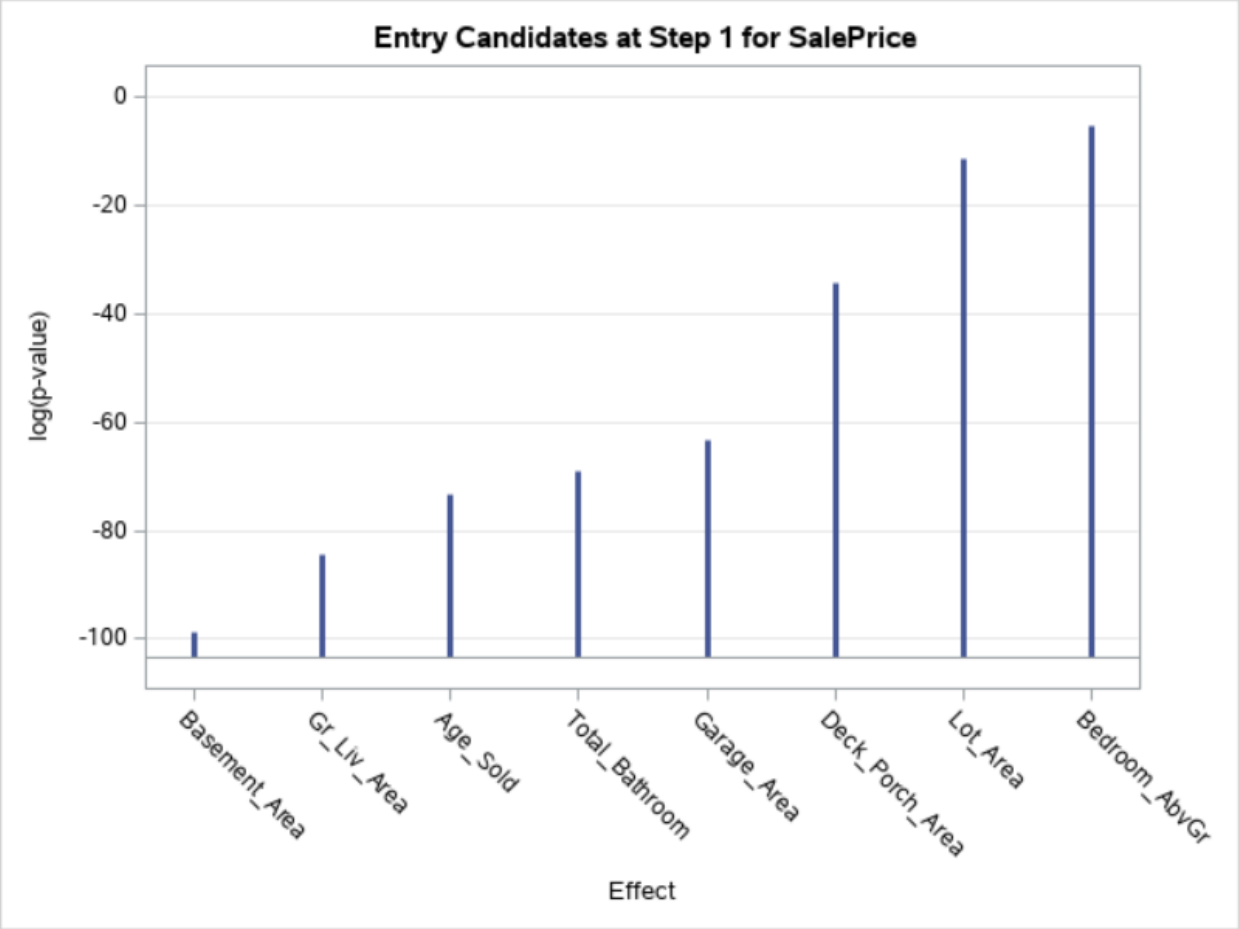
Effect Entered: Basement_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	2.012418E11	2.012418E11	270.16
Error	298	2.219817E11	744904950	
Corrected Total	299	4.232235E11		

Root MSE	27293
Dependent Mean	137525
R-Square	0.4755
Adj R-Sq	0.4737
AIC	6432.62346
AICC	6432.70454
SBC	6138.03102

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	73904	4179.193780	17.68
Basement_Area	1	72.107717	4.387055	16.44

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Basement_Area	-98.8577	<.0001
2	Gr_Liv_Area	-84.6132	<.0001
3	Age_Sold	-73.5219	<.0001
4	Total_Bathroom	-69.1880	<.0001
5	Garage_Area	-63.3558	<.0001
6	Deck_Porch_Area	-34.3105	<.0001
7	Lot_Area	-11.6303	<.0001
8	Bedroom_AbvGr	-5.5339	0.0040


Entry Candidates at Step 1 for SalePrice
log(p-value)
0
-20
-40
-60
-80
-100
Basement_Area
Gr_Liv_Area
Age_Sold
Total_Bathroom
Garage_Area
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Effect

Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Stepwise Selection: Step 2

Effect Entered: Gr_Liv_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	2.64483E11	1.322415E11	247.42
Error	297	1.587405E11	534479711	
Corrected Total	299	4.232235E11		

Root MSE	23119
Dependent Mean	137525
R-Square	0.6249
Adj R-Sq	0.6224
AIC	6334.02620
AICC	6334.16179
SBC	6043.13755

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	12664	6650.339855	1.90
Gr_Liv_Area	1	69.606974	6.399091	10.88
Basement_Area	1	52.309702	4.137885	12.64

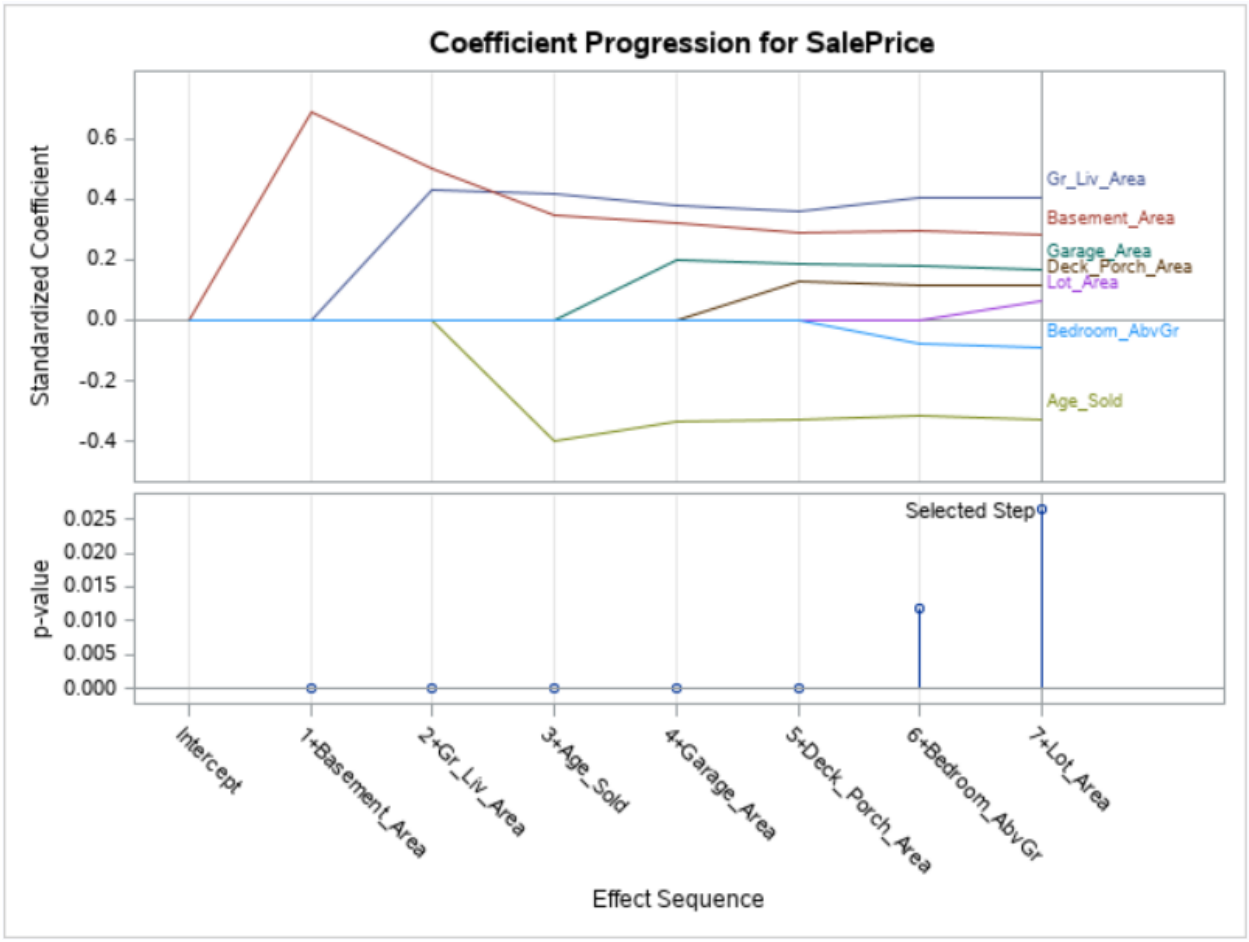
Stepwise Model Selection for SalePrice - SL 0.05

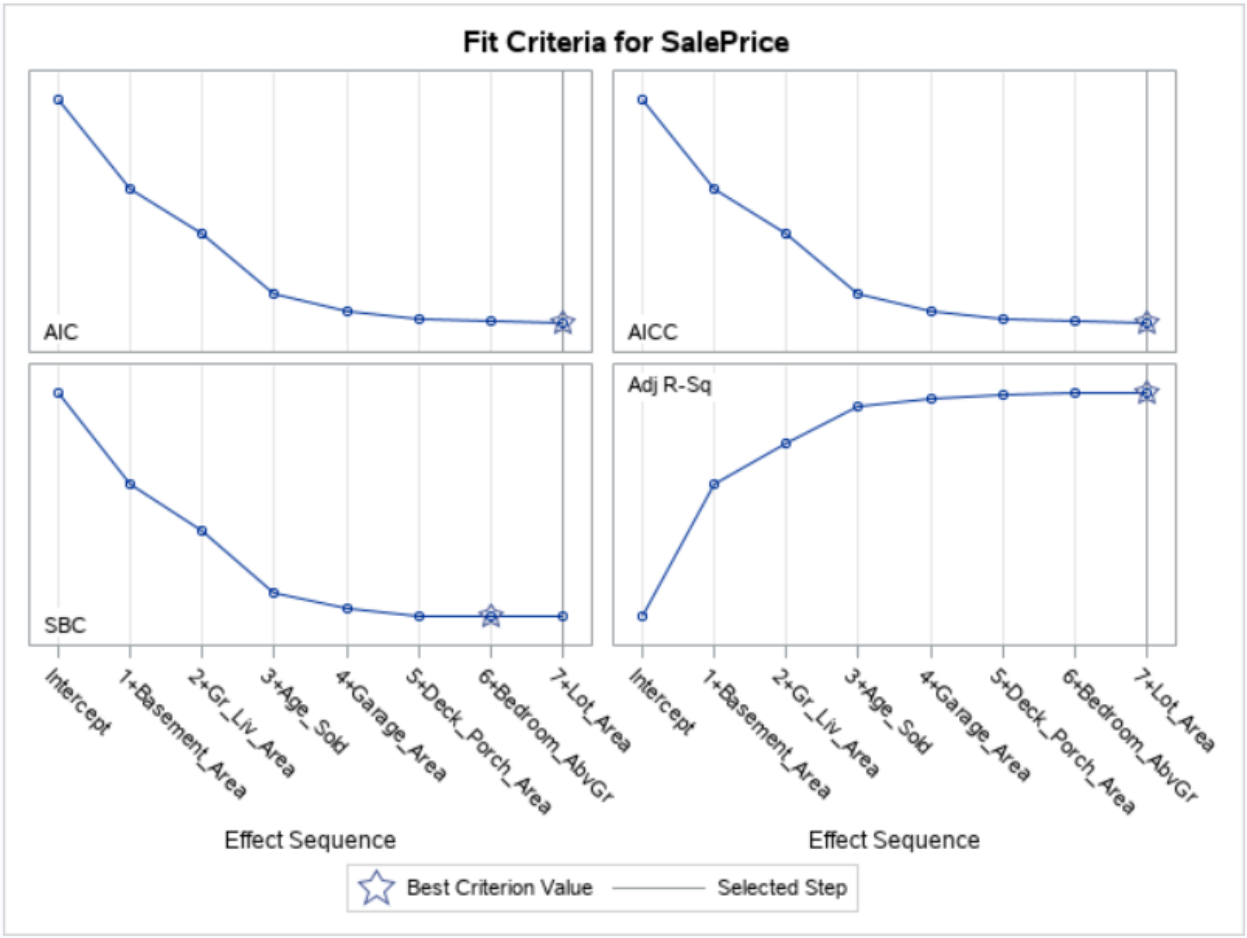
The GLMSELECT Procedure

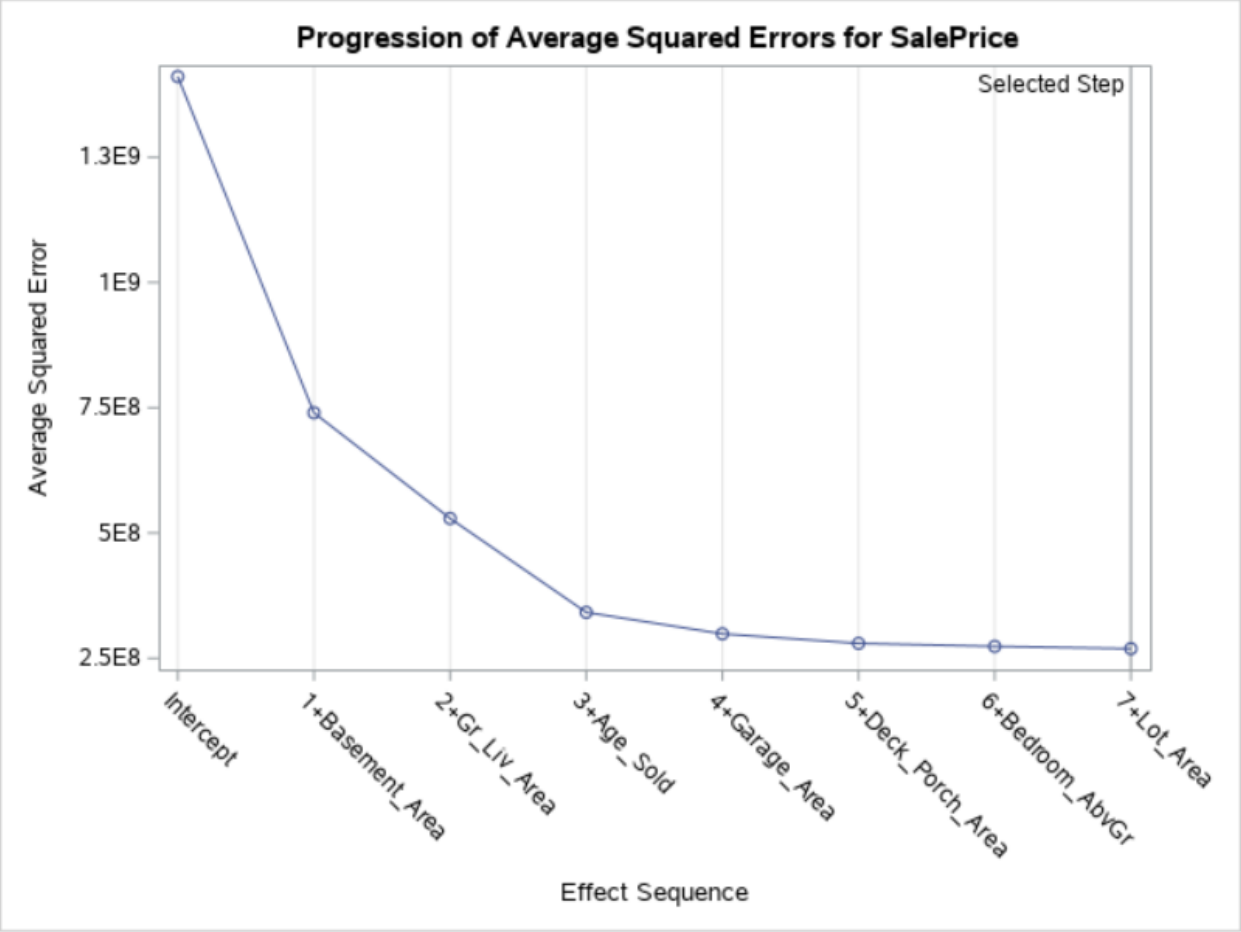
Stepwise Selection Summary					
Step	Effect Entered	Effect Removed	Number Effects In	F Value	Pr > F
0	Intercept		1	0.00	1.0000
1	Basement_Area		2	270.16	<.0001
2	Gr_Liv_Area		3	118.32	<.0001
3	Age_Sold		4	162.37	<.0001
4	Garage_Area		5	42.30	<.0001
5	Deck_Porch_Area		6	19.99	<.0001
6	Bedroom_AbvGr		7	6.41	0.0119
7	Lot_Area		8	4.97	0.0265

Selection stopped because the candidate for entry has SLE > 0.05 and the candidate for removal has SLS < 0.05.

Stop Details					
Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	Total_Bathroom	0.1167	>	0.0500	(SLE)
Removal	Lot_Area	0.0265	<	0.0500	(SLS)


Coefficient Progression for SalePrice
Standardized Coefficient
0.6
0.4
0.2
0.0
-0.2
-0.4
Gr_Liv_Area
Basement_Area
Garage_Area
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Age_Sold
p-value
0.025
0.020
0.015
0.010
0.005
0.000
Selected Step
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
Effect Sequence


Fit Criteria for SalePrice
AIC
AICC
SBC
Adj R-Sq
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
Effect Sequence
Effect Sequence
Best Criterion Value
Selected Step


Progression of Average Squared Errors for SalePrice
Selected Step
Average Squared Error
1.3E9
1E9
7.5E8
5E8
2.5E8
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
Effect Sequence

Stepwise Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 7).

Effects:	Intercept Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	7	3.424508E11	48921543221	176.86
Error	292	80772716963	276618894	
Corrected Total	299	4.232235E11		

Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	47463	5880.674041	8.07
Gr_Liv_Area	1	65.303724	5.436672	12.01
Basement_Area	1	29.849078	3.345400	8.92
Garage_Area	1	36.309606	6.452405	5.63
Deck_Porch_Area	1	32.052554	7.967677	4.02
Lot_Area	1	0.708127	0.317512	2.23
Age_Sold	1	-447.198682	41.019314	-10.90
Bedroom_AbvGr	1	-5042.766498	1687.928168	-2.99

Activity - Optional Stepwise Selection Method Code

In SAS Studio, submit the following code to perform both the forward selection and backward elimination processes.

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
             Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

proc glmselect data=stat1.ameshousing3 plots=all;
   FORWARD: model SalePrice=&interval / selection=forward details=steps select=SL slentry=0.05;
   title "Forward Model Selection for SalePrice - SL 0.05";
run;

proc glmselect data=stat1.ameshousing3 plots=all;
   BACKWARD: model SalePrice=&interval / selection=backward details=steps select=SL slstay=0.05;
   title "Backward Model Selection for SalePrice - SL 0.05";
run;

title;
```

Examine the results.

The final models that are obtained using the SLENTRY=0.05 and SLSTAY=0.05 criteria are displayed for FORWARD, BACKWARD, and STEPWISE. In this instance, all the selected models matched. However, this won't always be the case. When you run stepwise methods on your own data, the methods might select different models.

Also, recall the significance levels that the previous program used for entering the model and staying in the model. If you were to use different significance levels for entering the model and staying in the model, PROC GLMSELECT could produce very different models. The choice of SLENTRY and SLSTAY levels can greatly affect the final models that are selected using stepwise methods.

One last thing to remember is that the stepwise techniques don't take any collinearity in your model into account. Collinearity means that predictor variables in the same model are highly correlated. If collinearity is present in your model, you might want to consider first reducing the collinearity as much as possible and then running stepwise methods on the remaining variables.

```
/*Optional Code that will execute forward and backward selection
  Each with slentry and slstay = 0.05.
*/
proc glmselect data=STAT1.ameshousing3 plots=all;
        FORWARD: model SalePrice = &interval / selection=forward details=steps select=SL
slentry=0.05;
        title "Forward Model Selection for SalePrice - SL 0.05";
run;
```

Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure

Data Set	STAT1.AMESHOUSING3
Dependent Variable	SalePrice
Selection Method	Forward
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	300
Number of Observations Used	300

Dimensions	
Number of Effects	9
Number of Parameters	9

Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Forward Selection: Step 0

Effect Entered: Intercept

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	0	0	.	.
Error	299	4.232235E11	1415463276	
Corrected Total	299	4.232235E11		

Root MSE	37623
Dependent Mean	137525
R-Square	0.0000
Adj R-Sq	0.0000
AIC	6624.21515
AICC	6624.25555
SBC	6325.91893

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	137525	2172.144314	63.31

Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Forward Selection: Step 1

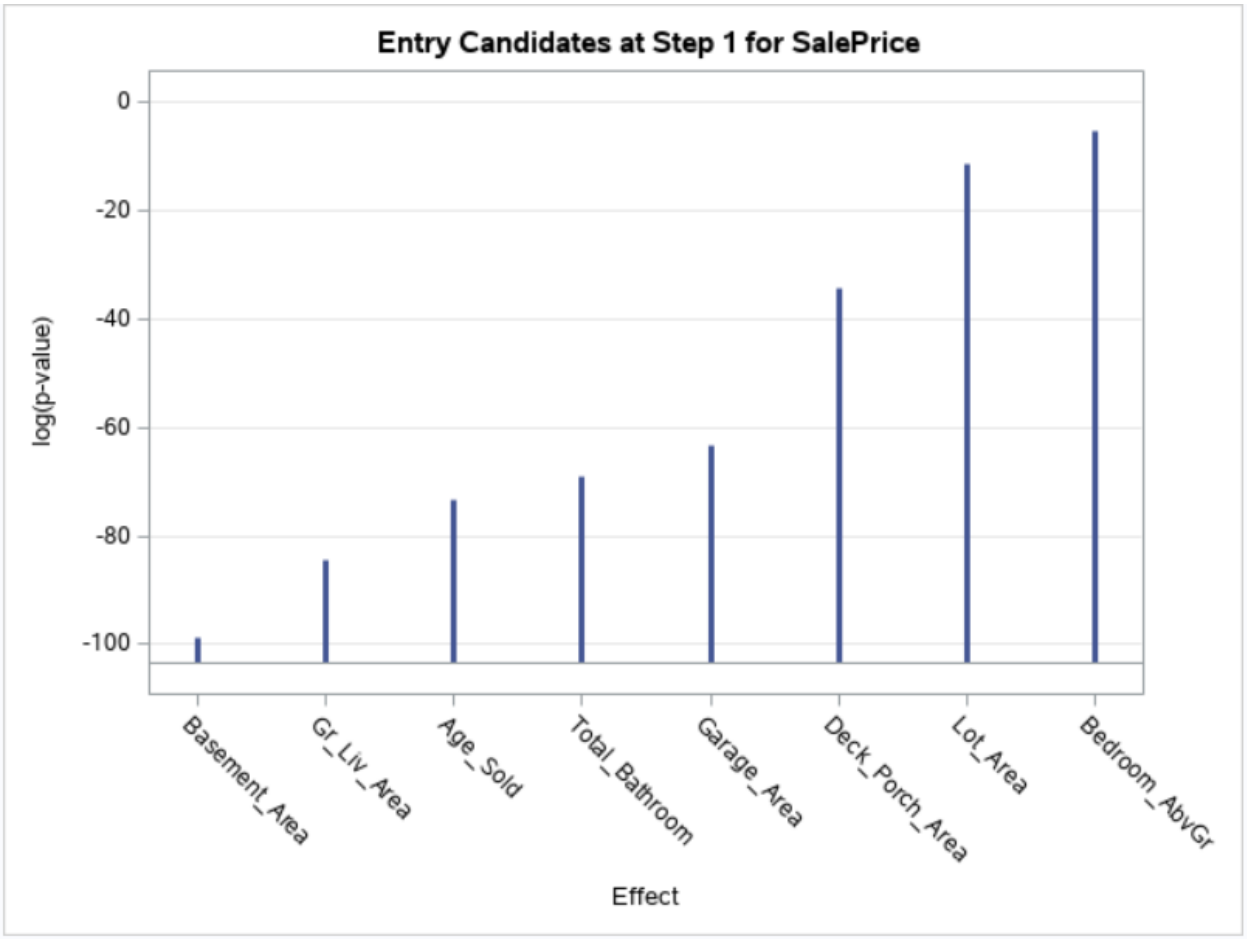
Effect Entered: Basement_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	2.012418E11	2.012418E11	270.16
Error	298	2.219817E11	744904950	
Corrected Total	299	4.232235E11		

Root MSE	27293
Dependent Mean	137525
R-Square	0.4755
Adj R-Sq	0.4737
AIC	6432.62346
AICC	6432.70454
SBC	6138.03102

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	73904	4179.193780	17.68
Basement_Area	1	72.107717	4.387055	16.44

Entry Candidates			
Rank	Effect	Log pValue	Pr > F
1	Basement_Area	-98.8577	<.0001
2	Gr_Liv_Area	-84.6132	<.0001
3	Age_Sold	-73.5219	<.0001
4	Total_Bathroom	-69.1880	<.0001
5	Garage_Area	-63.3558	<.0001
6	Deck_Porch_Area	-34.3105	<.0001
7	Lot_Area	-11.6303	<.0001
8	Bedroom_AbvGr	-5.5339	0.0040


Entry Candidates at Step 1 for SalePrice
0
-20
-40
-60
-80
-100
log(p-value)
Basement_Area
Gr_Liv_Area
Age_Sold
Total_Bathroom
Garage_Area
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Effect

Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Forward Selection: Step 2

Effect Entered: Gr_Liv_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	2	2.64483E11	1.322415E11	247.42
Error	297	1.587405E11	534479711	
Corrected Total	299	4.232235E11		

Root MSE	23119
Dependent Mean	137525
R-Square	0.6249
Adj R-Sq	0.6224
AIC	6334.02620
AICC	6334.16179
SBC	6043.13755

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	12664	6650.339855	1.90
Gr_Liv_Area	1	69.606974	6.399091	10.88
Basement_Area	1	52.309702	4.137885	12.64

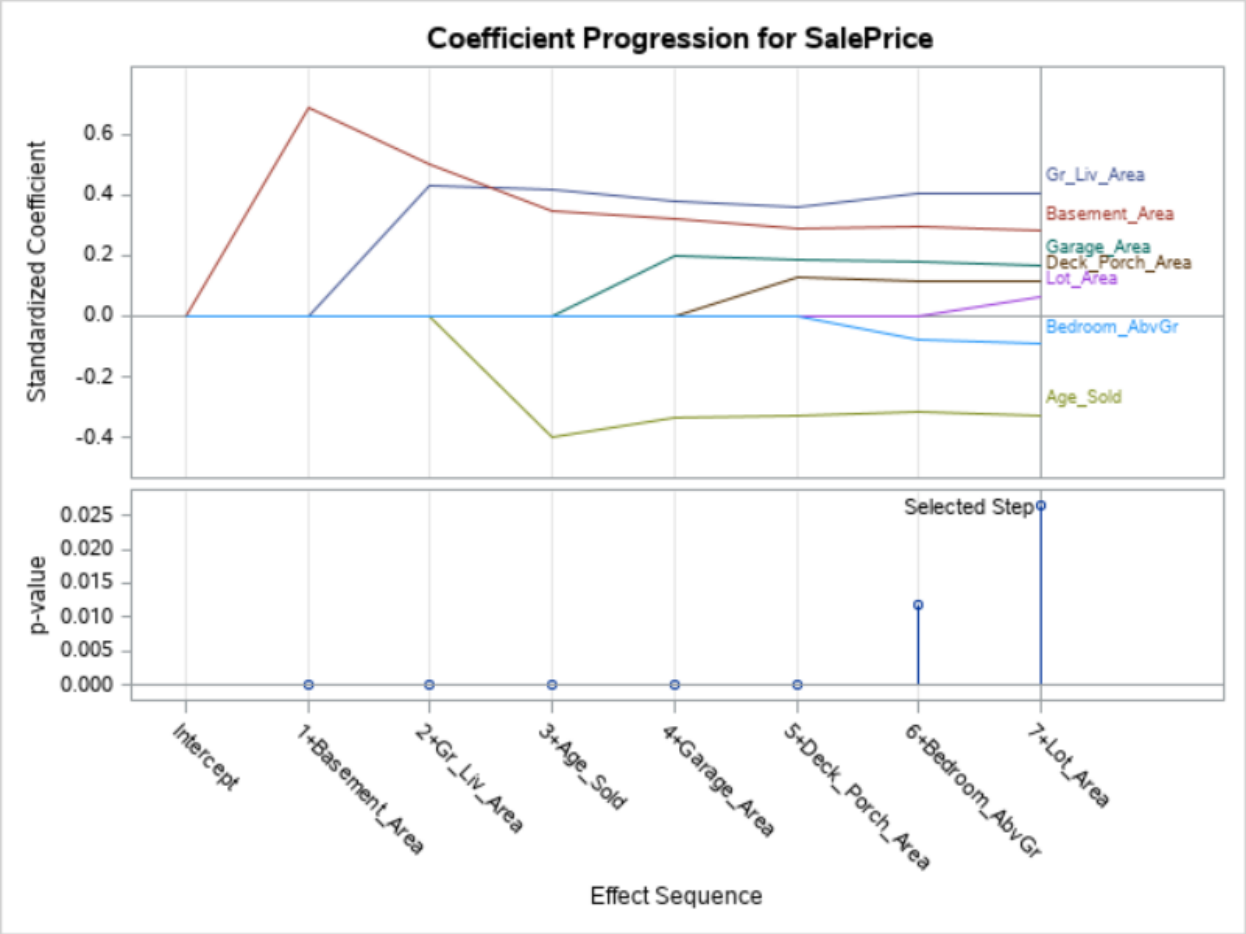
Forward Model Selection for SalePrice - SL 0.05

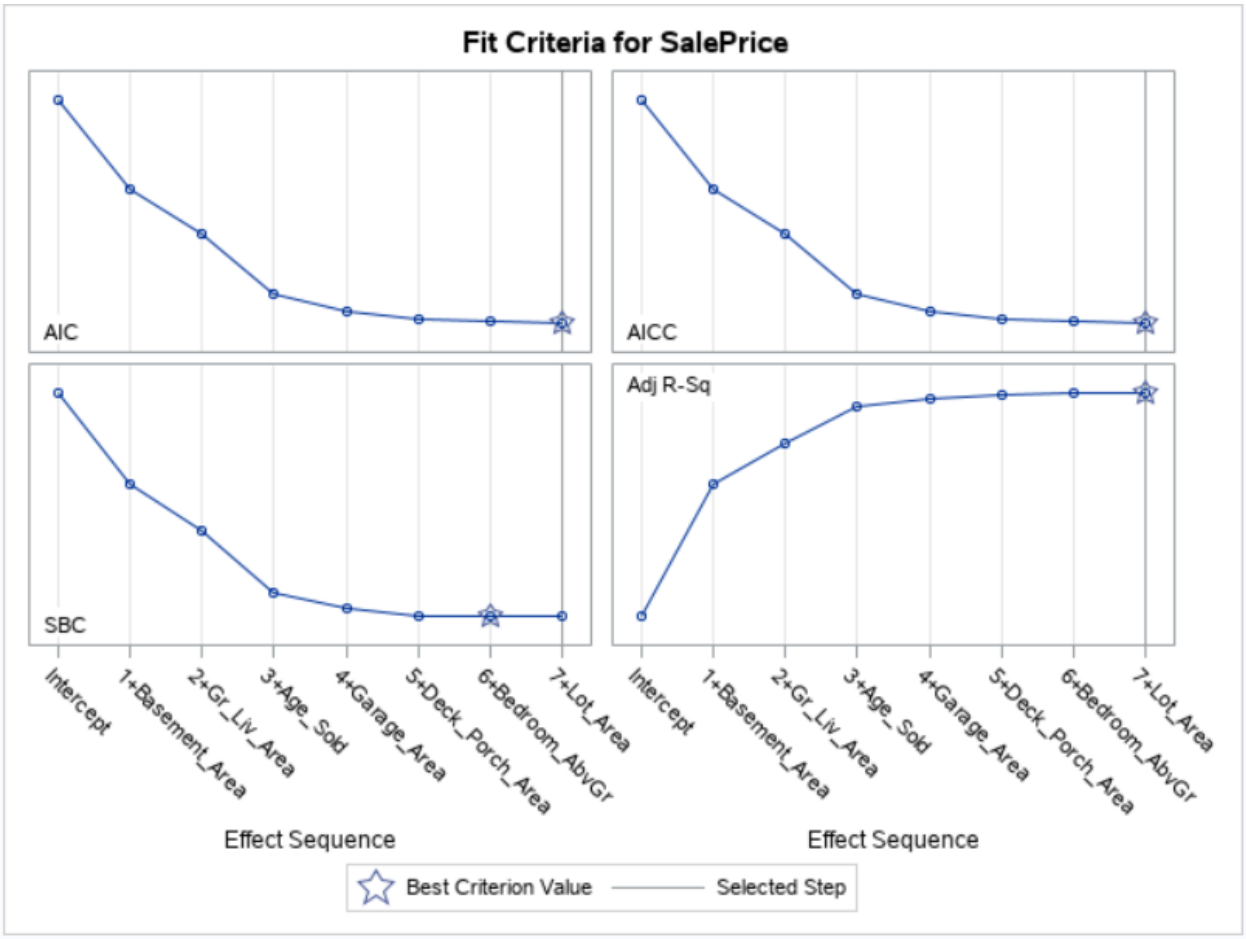
The GLMSELECT Procedure

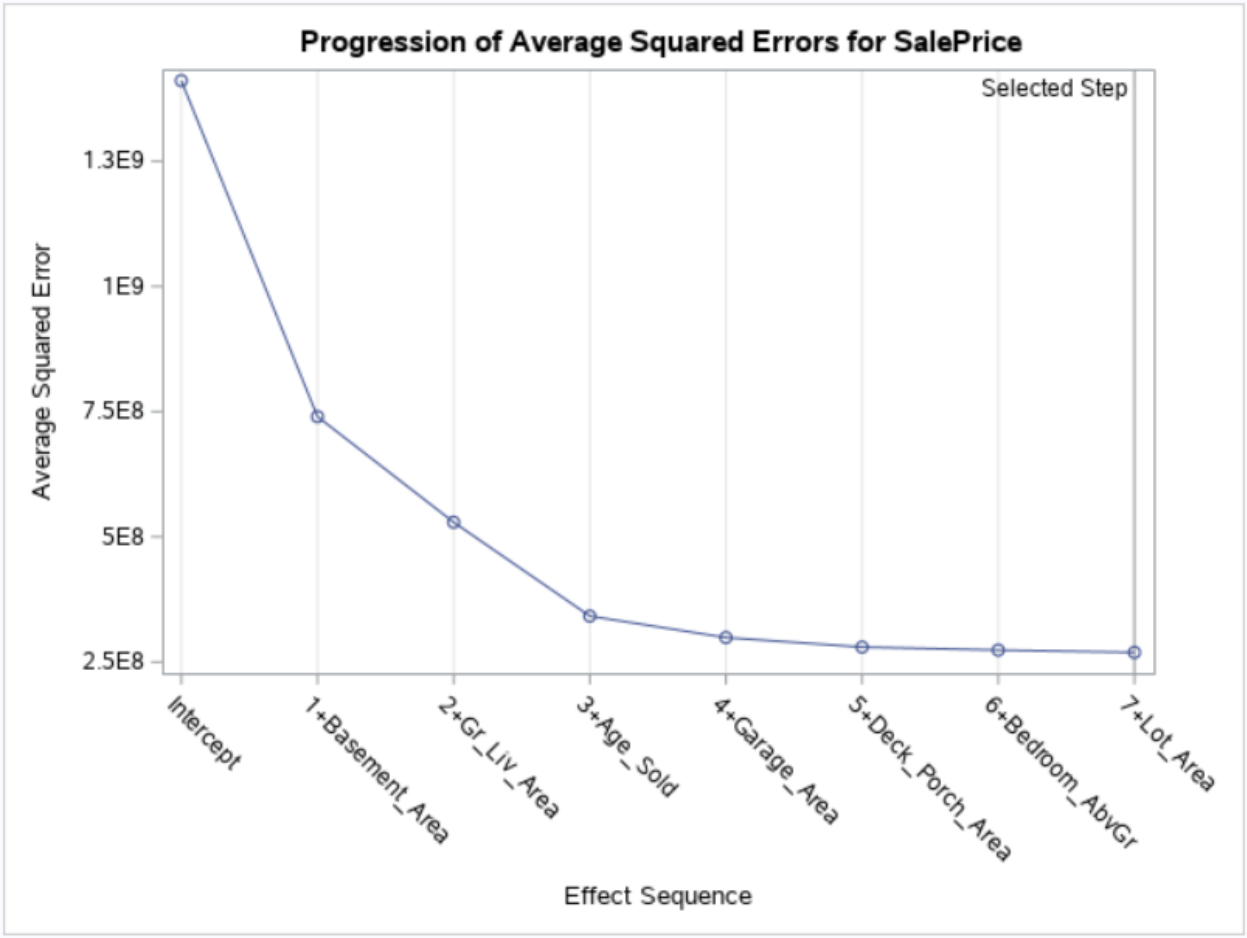
Forward Selection Summary				
Step	Effect Entered	Number Effects In	F Value	Pr > F
0	Intercept	1	0.00	1.0000
1	Basement_Area	2	270.16	<.0001
2	Gr_Liv_Area	3	118.32	<.0001
3	Age_Sold	4	162.37	<.0001
4	Garage_Area	5	42.30	<.0001
5	Deck_Porch_Area	6	19.99	<.0001
6	Bedroom_AbvGr	7	6.41	0.0119
7	Lot_Area	8	4.97	0.0265

Selection stopped as the candidate for entry has SLE > 0.05.

Stop Details					
Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	Total_Bathroom	0.1167	>	0.0500	(SLE)


Coefficient Progression for SalePrice
Standardized Coefficient
0.6
0.4
0.2
0.0
-0.2
-0.4
Gr_Liv_Area
Basement_Area
Garage_Area
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Age_Sold
p-value
0.025
0.020
0.015
0.010
0.005
0.000
Selected Step
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
Effect Sequence


Fit Criteria for SalePrice
AIC
AICC
SBC
Adj R-Sq
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
Effect Sequence
Effect Sequence
Best Criterion Value
Selected Step


Progression of Average Squared Errors for SalePrice
Selected Step
Average Squared Error
1.3E9
1E9
7.5E8
5E8
2.5E8
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
Effect Sequence

Forward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 7).

Effects:	Intercept Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	7	3.424508E11	48921543221	176.86
Error	292	80772716963	276618894	
Corrected Total	299	4.232235E11		

Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	47463	5880.674041	8.07
Gr_Liv_Area	1	65.303724	5.436672	12.01
Basement_Area	1	29.849078	3.345400	8.92
Garage_Area	1	36.309606	6.452405	5.63
Deck_Porch_Area	1	32.052554	7.967677	4.02
Lot_Area	1	0.708127	0.317512	2.23
Age_Sold	1	-447.198682	41.019314	-10.90
Bedroom_AbvGr	1	-5042.766498	1687.928168	-2.99

```
proc glmselect data=STAT1.ameshousing3 plots=all;

        BACKWARD: model SalePrice = &interval / selection=backward details=steps select=SL
slstay=0.05;

        title "Backward Model Selection for SalePrice - SL 0.05";

run;
```

Backward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure

Data Set	STAT1.AMESHOUSING3
Dependent Variable	SalePrice
Selection Method	Backward
Select Criterion	Significance Level
Stop Criterion	Significance Level
Stay Significance Level (SLS)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	300
Number of Observations Used	300

Dimensions	
Number of Effects	9
Number of Parameters	9

Backward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Backward Selection: Step 0

Full Least Squares Model

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	8	3.431321E11	42891512314	155.84
Error	291	80091420996	275228251	
Corrected Total	299	4.232235E11		

Root MSE	16590
Dependent Mean	137525
R-Square	0.8108
Adj R-Sq	0.8056
AIC	6140.79563
AICC	6141.55688
SBC	5872.12967

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	44347	6191.271944	7.16
Gr_Liv_Area	1	63.197764	5.585739	11.31
Basement_Area	1	28.692184	3.417034	8.40
Garage_Area	1	35.754191	6.445840	5.55
Deck_Porch_Area	1	31.370539	7.959436	3.94
Lot_Area	1	0.699495	0.316761	2.21
Age_Sold	1	-420.815037	44.219144	-9.52
Bedroom_AbvGr	1	-4834.848748	1688.858227	-2.86
Total_Bathroom	1	3022.124723	1920.839066	1.57

Backward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Backward Selection: Step 1

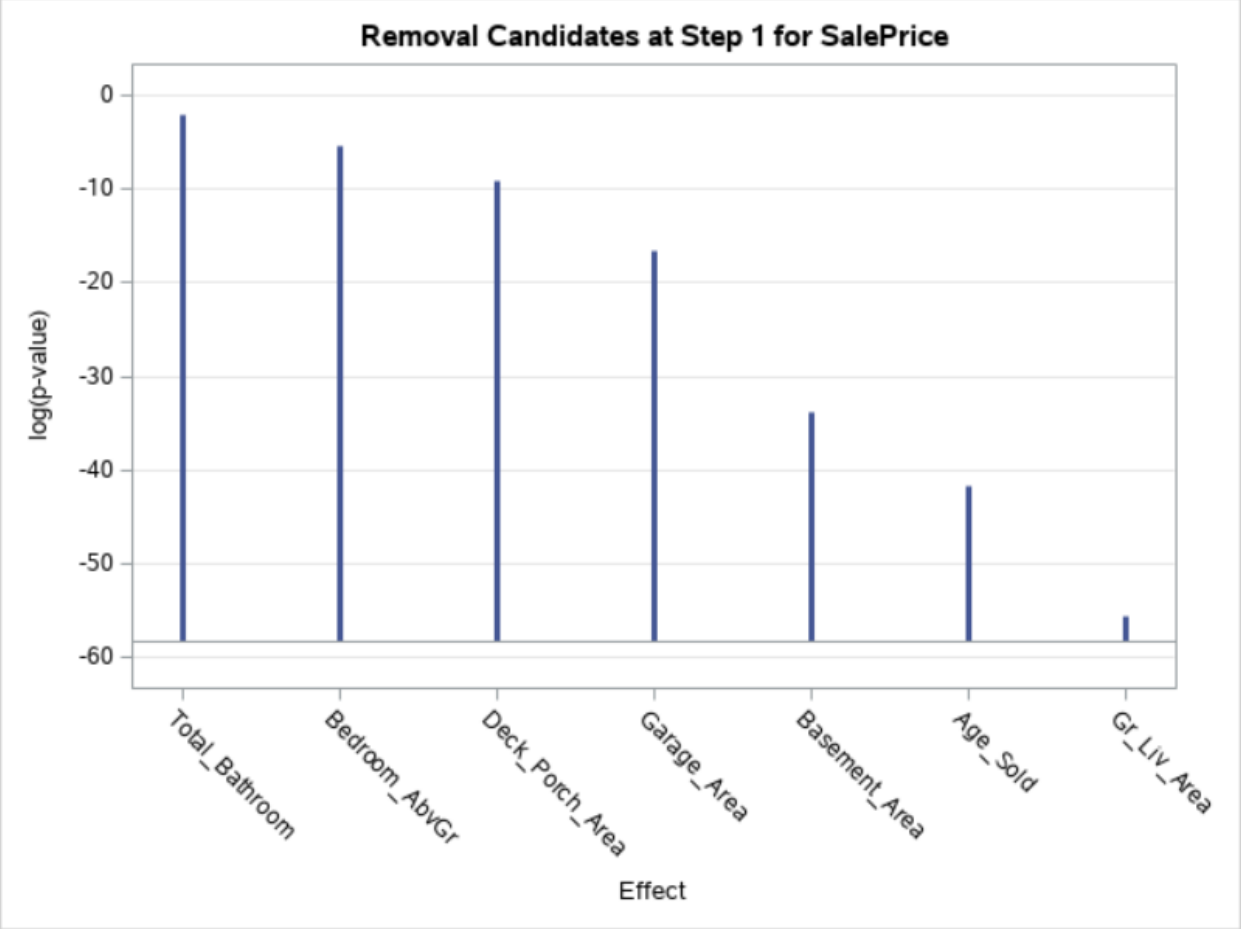
Effect Removed: Total_Bathroom

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	7	3.424508E11	48921543221	176.86
Error	292	80772716963	276618894	
Corrected Total	299	4.232235E11		

Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	47463	5880.674041	8.07
Gr_Liv_Area	1	65.303724	5.436672	12.01
Basement_Area	1	29.849078	3.345400	8.92
Garage_Area	1	36.309606	6.452405	5.63
Deck_Porch_Area	1	32.052554	7.967677	4.02
Lot_Area	1	0.708127	0.317512	2.23
Age_Sold	1	-447.198682	41.019314	-10.90
Bedroom_AbvGr	1	-5042.766498	1687.928168	-2.99

Removal Candidates			
Rank	Effect	Log pValue	Pr > F
1	Total_Bathroom	-2.1479	0.1167
2	Bedroom_AbvGr	-5.4027	0.0045
3	Deck_Porch_Area	-9.1949	0.0001
4	Garage_Area	-16.5434	<.0001
5	Basement_Area	-33.8268	<.0001
6	Age_Sold	-41.7794	<.0001
7	Gr_Liv_Area	-55.5230	<.0001



Backward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure

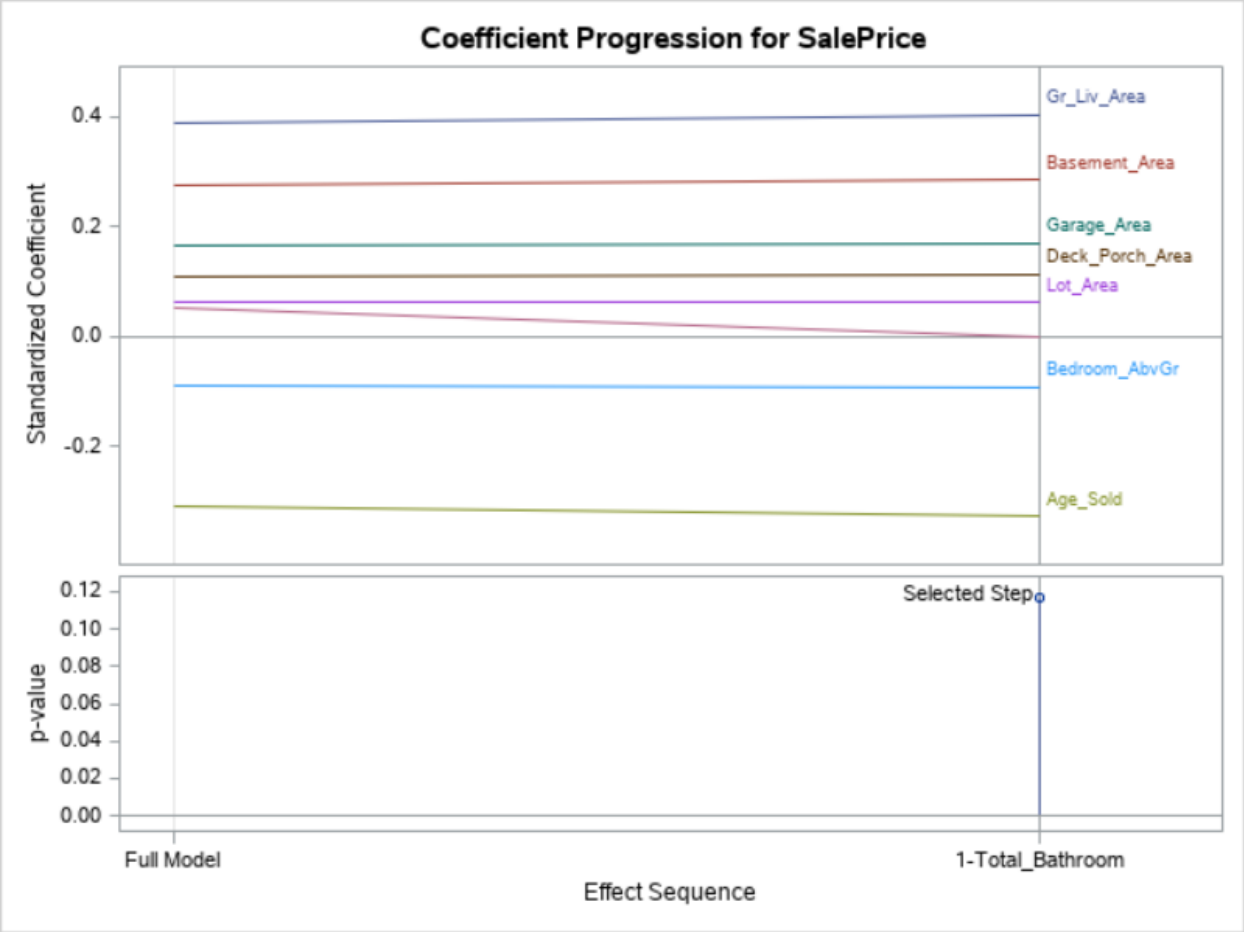
Backward Selection Summary

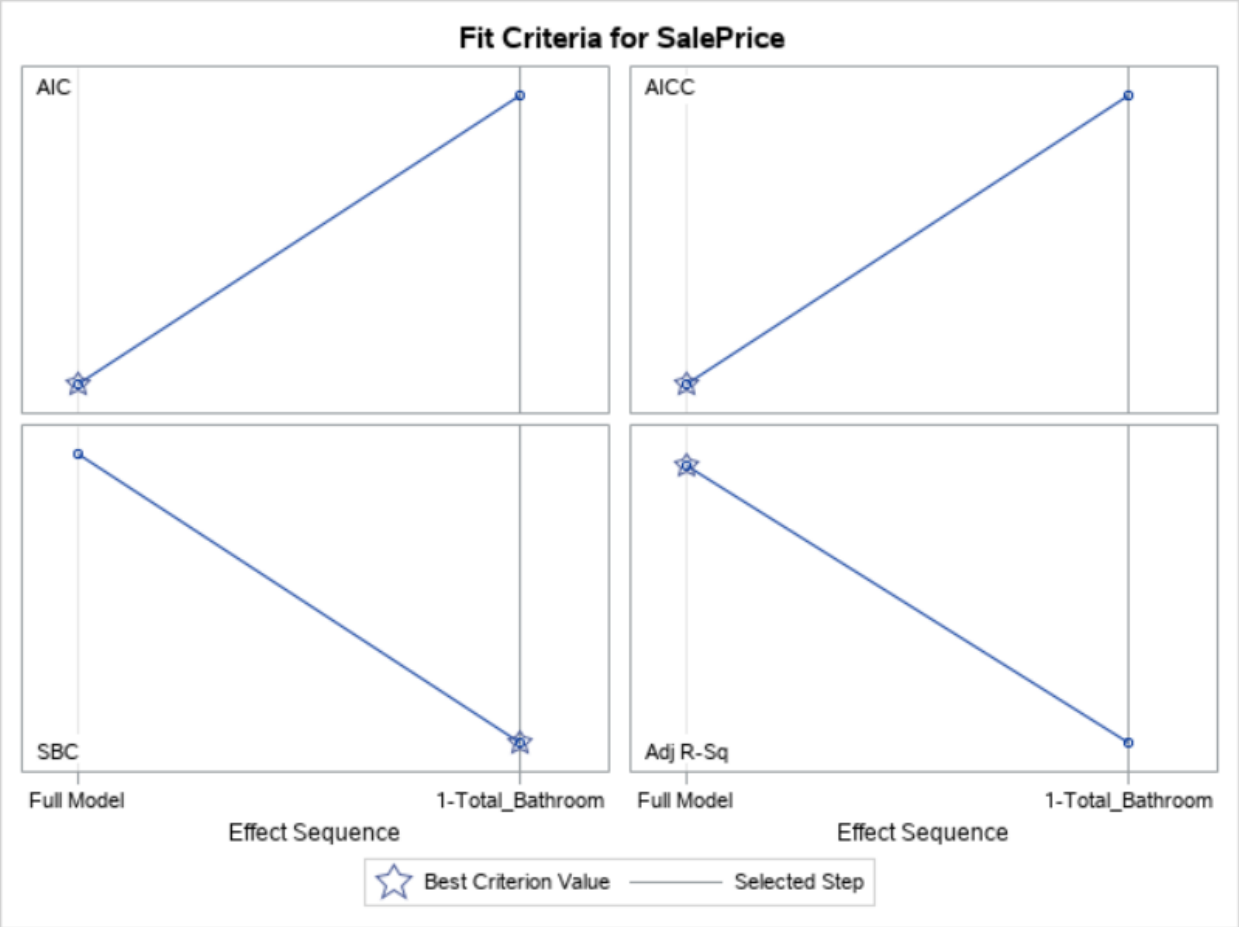
Step	Effect Removed	Number Effects In	F Value	Pr > F
0		9		
1	Total_Bathroom	8	2.48	0.1167

Selection stopped because the next candidate for removal has SLS < 0.05.

Stop Details

Candidate For	Effect	Candidate Significance		Compare Significance	
Removal	Lot_Area	0.0265	<	0.0500	(SLS)


Coefficient Progression for SalePrice
Standardized Coefficient
0.4
0.2
0.0
-0.2
Gr_Liv_Area
Basement_Area
Garage_Area
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Age_Sold
p-value
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Selected Step
Full Model
1-Total_Bathroom
Effect Sequence


Fit Criteria for SalePrice
AIC
AICC
SBC
Adj R-Sq
Full Model
1-Total_Bathroom
Full Model
1-Total_Bathroom
Effect Sequence
Effect Sequence
Best Criterion Value
Selected Step


Progression of Average Squared Errors for SalePrice
Selected Step
2.7E8
2.7E8
2.7E8
2.7E8
2.7E8
Average Squared Error
Full Model
1-Total_Bathroom
Effect Sequence

Backward Model Selection for SalePrice - SL 0.05

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 1).

Effects:	Intercept Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr

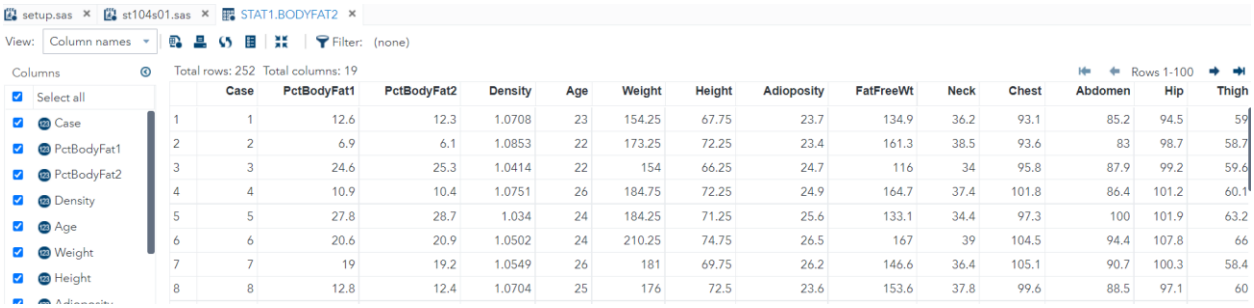
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	7	3.424508E11	48921543221	176.86
Error	292	80772716963	276618894	
Corrected Total	299	4.232235E11		

Root MSE	16632
Dependent Mean	137525
R-Square	0.8091
Adj R-Sq	0.8046
AIC	6141.33678
AICC	6141.95747
SBC	5868.96704

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	47463	5880.674041	8.07
Gr_Liv_Area	1	65.303724	5.436672	12.01
Basement_Area	1	29.849078	3.345400	8.92
Garage_Area	1	36.309606	6.452405	5.63
Deck_Porch_Area	1	32.052554	7.967677	4.02
Lot_Area	1	0.708127	0.317512	2.23
Age_Sold	1	-447.198682	41.019314	-10.90
Bedroom_AbvGr	1	-5042.766498	1687.928168	-2.99

If you specify the same significance levels for the STEPWISE, BACKWARD, and FORWARD selection methods, do all three methods result in the same final model?

No because the techniques for selecting or eliminating variables differ between the three selection methods, they don't always produce the same final model.



	Case	PctBodyFat1	PctBodyFat2	Density	Age	Weight	Height	Adioposity	FatFreeWt	Neck	Chest	Abdomen	Hip	Thigh
1	1	12.6	12.3	1.0708	23	154.25	67.75	23.7	134.9	36.2	93.1	85.2	94.5	59
2	2	6.9	6.1	1.0853	22	173.25	72.25	23.4	161.3	38.5	93.6	83	98.7	58.7
3	3	24.6	25.3	1.0414	22	154	66.25	24.7	116	34	95.8	87.9	99.2	59.6
4	4	10.9	10.4	1.0751	26	184.75	72.25	24.9	164.7	37.4	101.8	86.4	101.2	60.1
5	5	27.8	28.7	1.034	24	184.25	71.25	25.6	133.1	34.4	97.3	100	101.9	63.2
6	6	20.6	20.9	1.0502	24	210.25	74.75	26.5	167	39	104.5	94.4	107.8	66
7	7	19	19.2	1.0549	26	181	69.75	26.2	146.6	36.4	105.1	90.7	100.3	58.4
8	8	12.8	12.4	1.0704	25	176	72.5	23.6	153.6	37.8	99.6	88.5	97.1	60

```
/*st104s01.sas*/ /*Part A*/

ods graphics on;

proc glmselect data=STAT1.bodyfat2 plots=all;

        STEPWISESL: model PctBodyFat2 = Age Weight Height Neck Chest Abdomen

                    Hip Thigh Knee Ankle Biceps Forearm Wrist

                    / SELECTION=STEPWISE SELECT=SL;

        title 'SL STEPWISE Selection with PctBodyFat2';

run;
```

SL STEPWISE Selection with PctBodyFat2

The GLMSELECT Procedure

Data Set	STAT1.BODYFAT2
Dependent Variable	PctBodyFat2
Selection Method	Stepwise
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.15
Stay Significance Level (SLS)	0.15
Effect Hierarchy Enforced	None

Number of Observations Read	252
Number of Observations Used	252

Dimensions	
Number of Effects	14
Number of Parameters	14

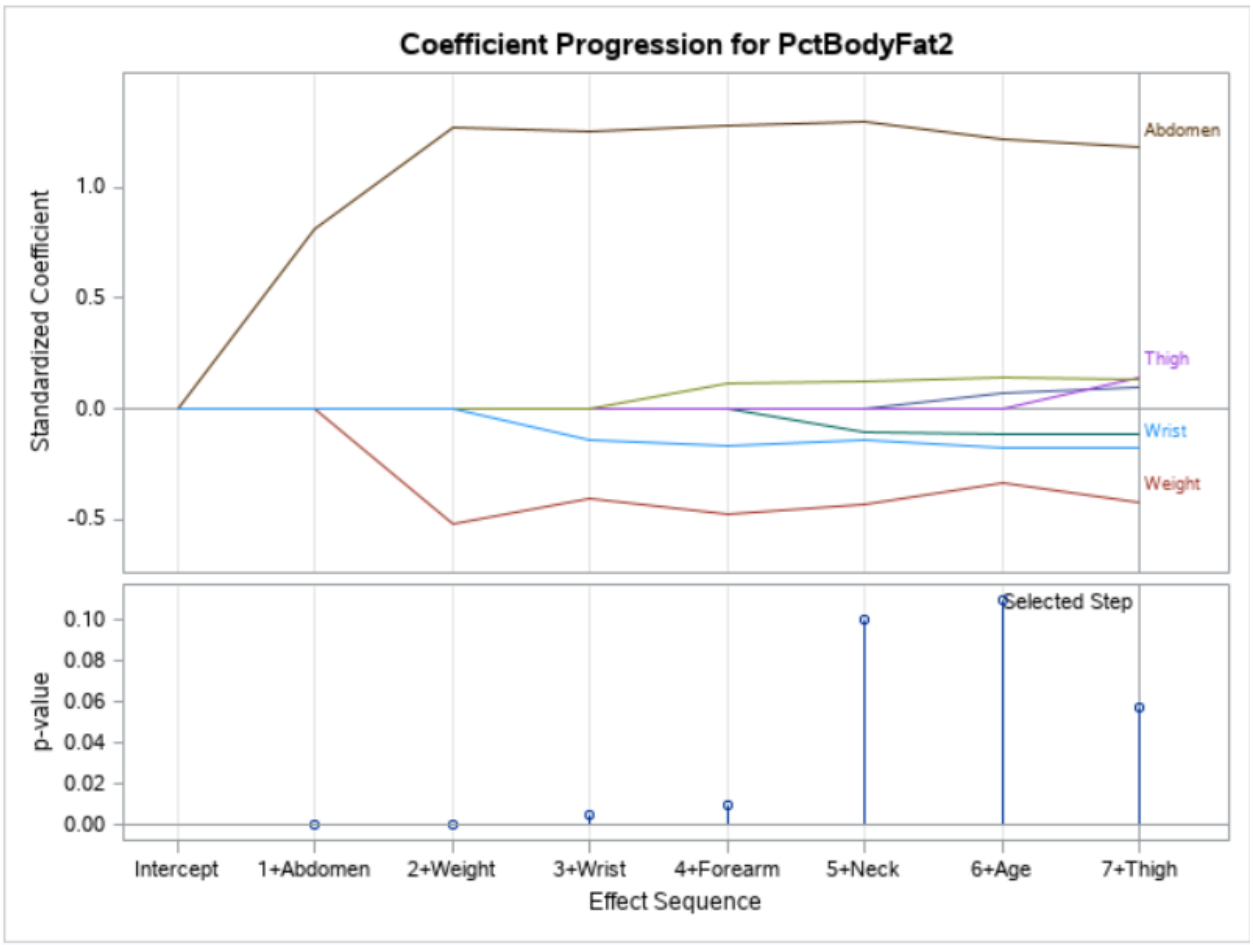
SL STEPWISE Selection with PctBodyFat2

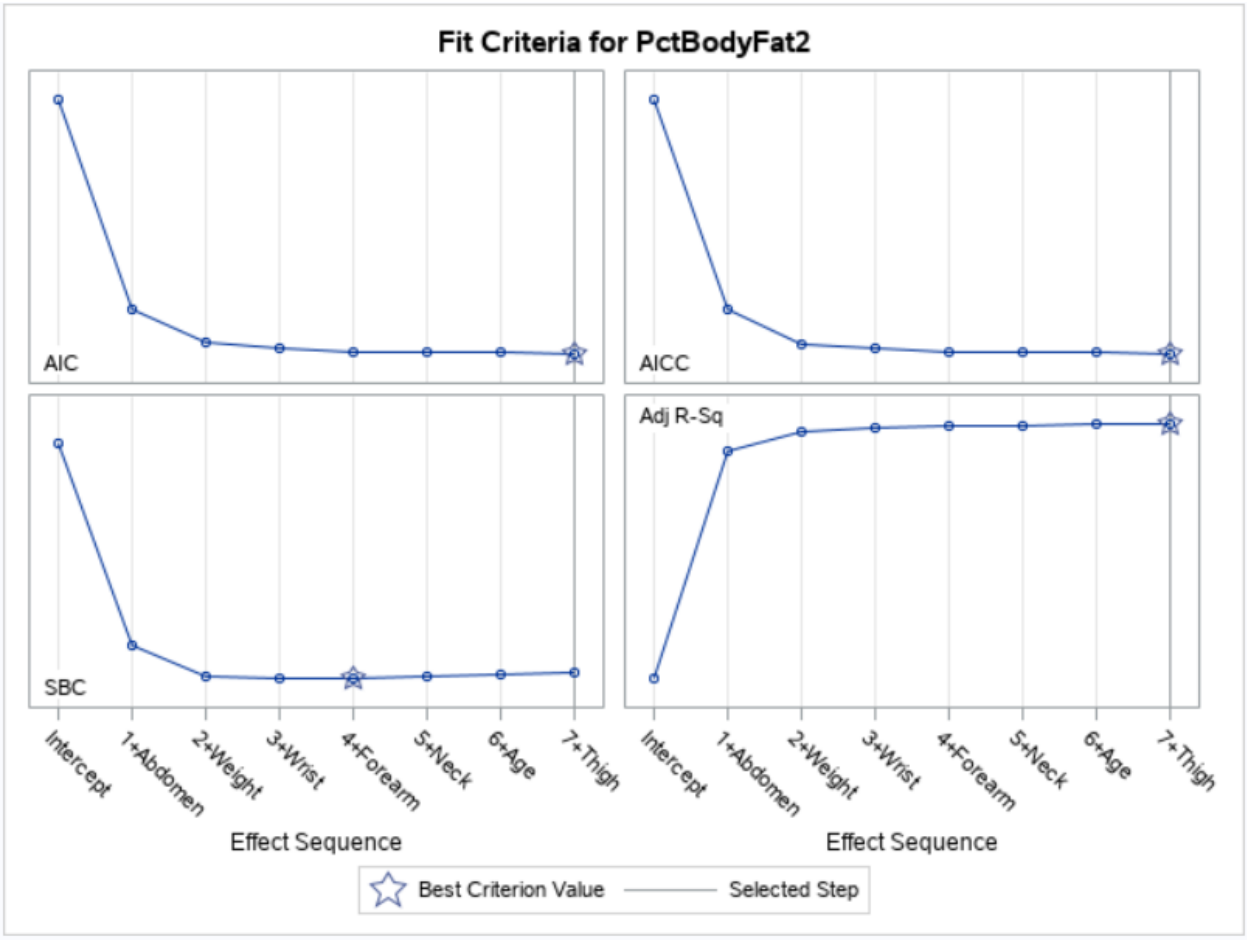
The GLMSELECT Procedure

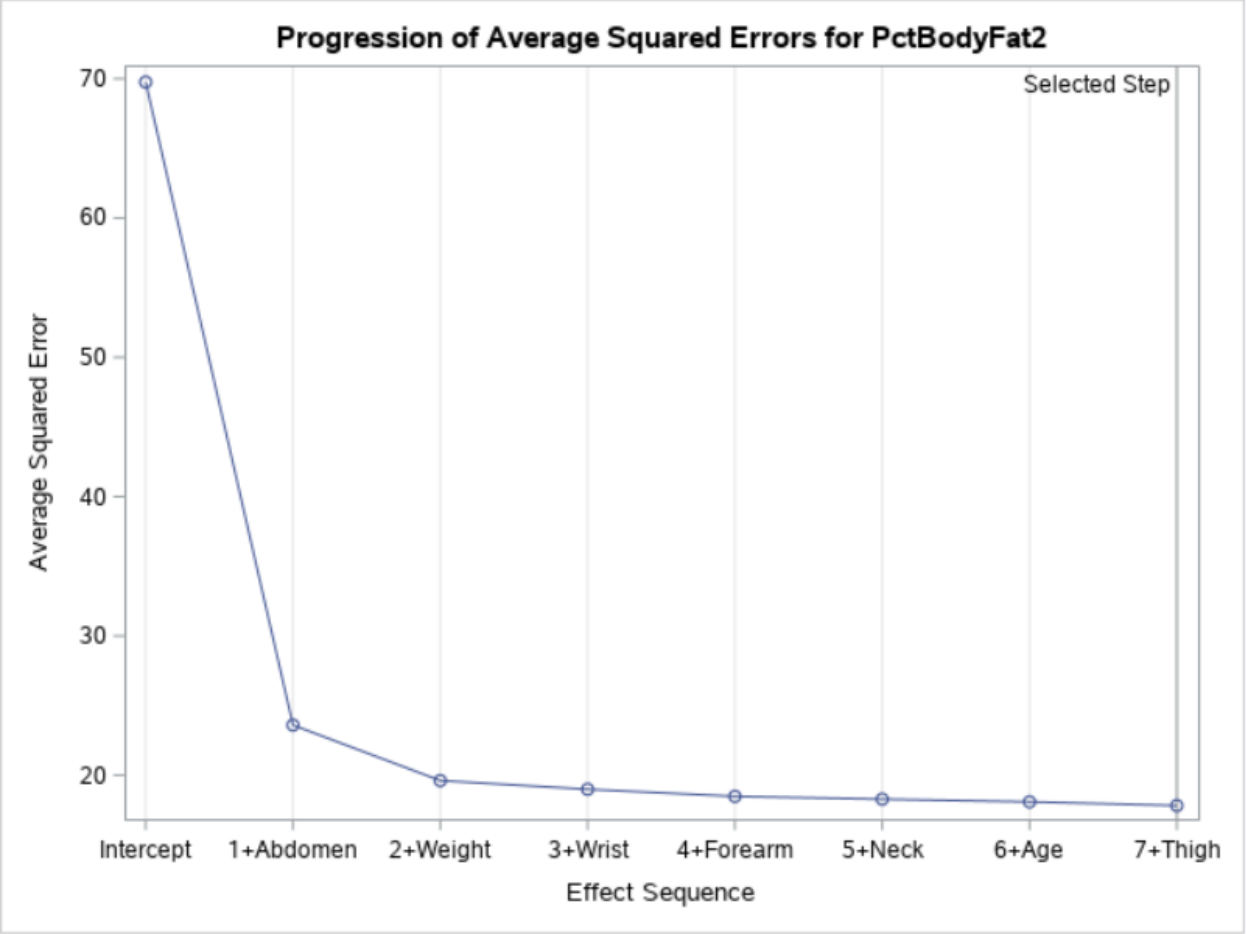
Stepwise Selection Summary					
Step	Effect Entered	Effect Removed	Number Effects In	F Value	Pr > F
0	Intercept		1	0.00	1.0000
1	Abdomen		2	488.93	<.0001
2	Weight		3	50.58	<.0001
3	Wrist		4	8.15	0.0047
4	Forearm		5	6.78	0.0098
5	Neck		6	2.73	0.1000
6	Age		7	2.58	0.1098
7	Thigh		8	3.66	0.0569

Selection stopped because the candidate for entry has SLE > 0.15 and the candidate for removal has SLS < 0.15.

Stop Details					
Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	Hip	0.1594	>	0.1500	(SLE)
Removal	Neck	0.0684	<	0.1500	(SLS)


Coefficient Progression for PctBodyFat2
Standardized Coefficient
1.0
0.5
0.0
-0.5
Abdomen
Thigh
Wrist
Weight
p-value
0.10
0.08
0.06
0.04
0.02
0.00
Selected Step
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
5+Neck
6+Age
7+Thigh
Effect Sequence


Fit Criteria for PctBodyFat2
AIC
AICC
SBC
Adj R-Sq
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
5+Neck
6+Age
7+Thigh
Effect Sequence
Effect Sequence
Best Criterion Value
Selected Step


Progression of Average Squared Errors for PctBodyFat2
Selected Step
Average Squared Error
70
60
50
40
30
20
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
5+Neck
6+Age
7+Thigh
Effect Sequence

SL STEPWISE Selection with PctBodyFat2

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 7).

Effects:	Intercept Age Weight Neck Abdomen Thigh Forearm Wrist

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	7	13087	1869.59160	101.56
Error	244	4491.84861	18.40922	
Corrected Total	251	17579		

Root MSE	4.29060
Dependent Mean	19.15079
R-Square	0.7445
Adj R-Sq	0.7371
AIC	995.90881
AICC	996.65261
SBC	770.14425

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-33.257991	9.006812	-3.69
Age	1	0.068166	0.030792	2.21
Weight	1	-0.119441	0.034025	-3.51
Neck	1	-0.403802	0.220620	-1.83
Abdomen	1	0.917885	0.069499	13.21
Thigh	1	0.221960	0.116013	1.91
Forearm	1	0.553139	0.184788	2.99
Wrist	1	-1.532401	0.510415	-3.00

```
/*st104s01.sas*/ /*Part B*/
proc glmselect data=STAT1.bodyfat2 plots=all;
        FORWARDSL: model PctBodyFat2 = Age Weight Height Neck Chest Abdomen
                Hip Thigh Knee Ankle Biceps Forearm Wrist
                / SELECTION=FORWARD SELECT=SL;
        title 'SL FORWARD Selection with PctBodyFat2';
run;
```

SL FORWARD Selection with PctBodyFat2

The GLMSELECT Procedure

Data Set	STAT1.BODYFAT2
Dependent Variable	PctBodyFat2
Selection Method	Forward
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.5
Effect Hierarchy Enforced	None

Number of Observations Read	252
Number of Observations Used	252

Dimensions	
Number of Effects	14
Number of Parameters	14

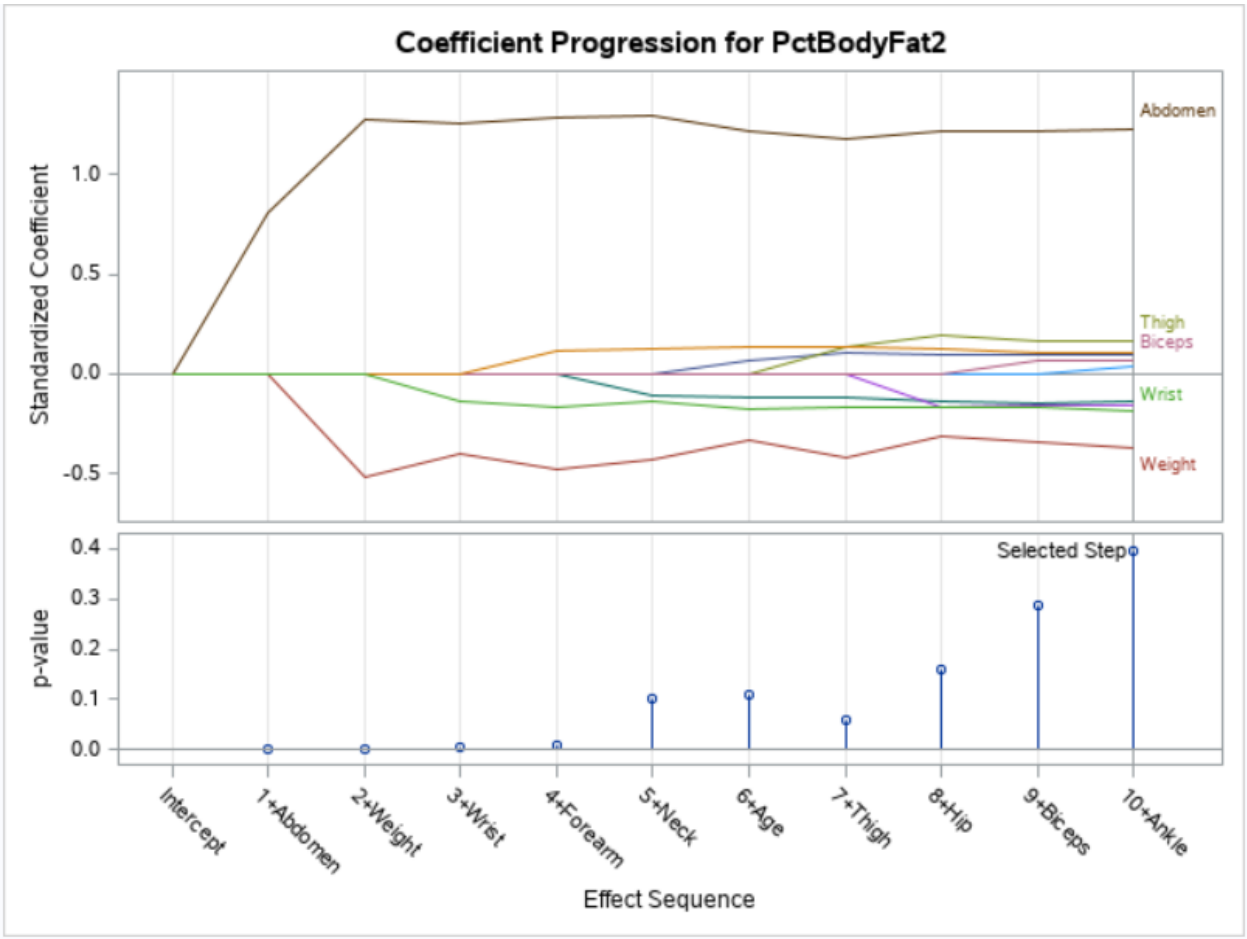
SL FORWARD Selection with PctBodyFat2

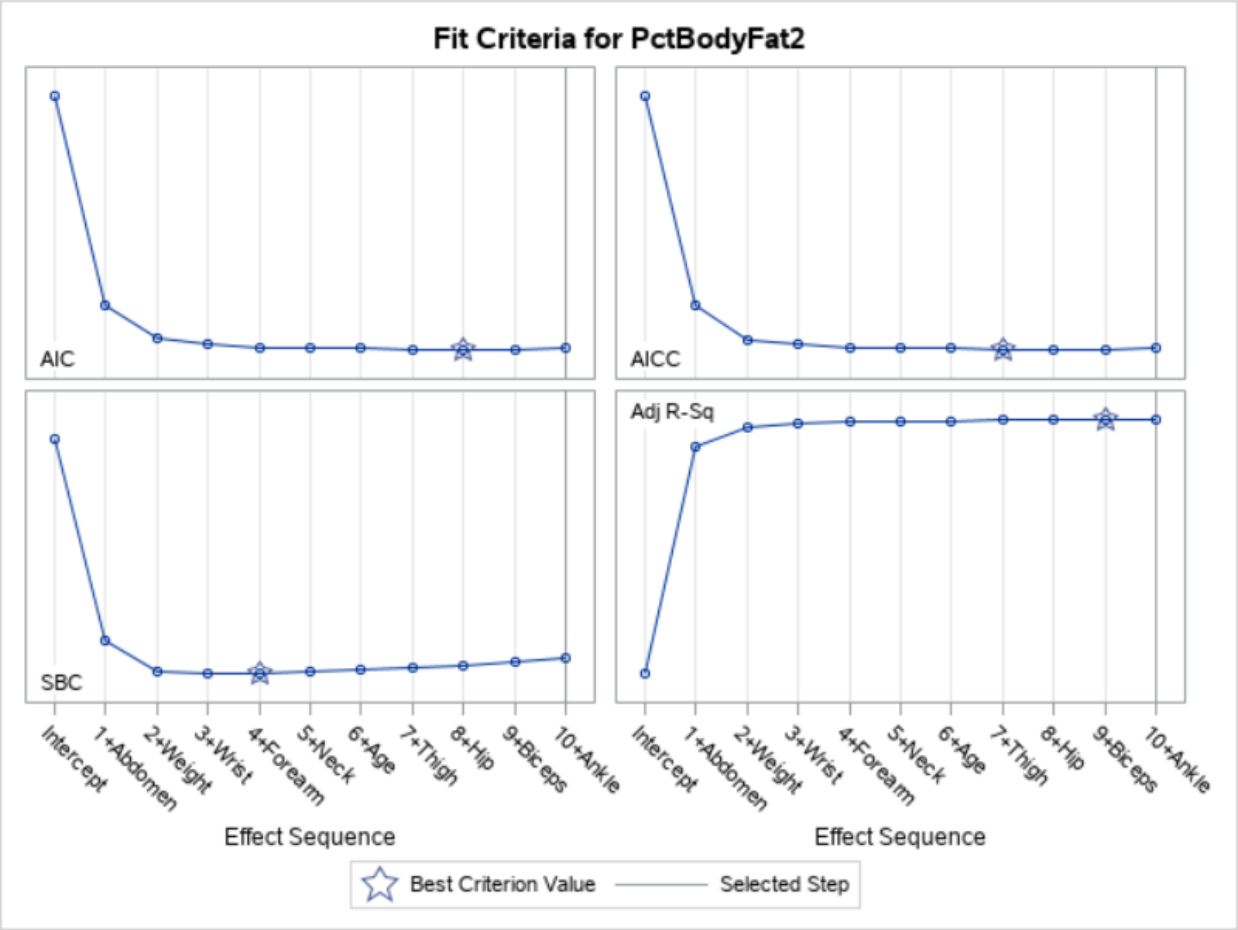
The GLMSELECT Procedure

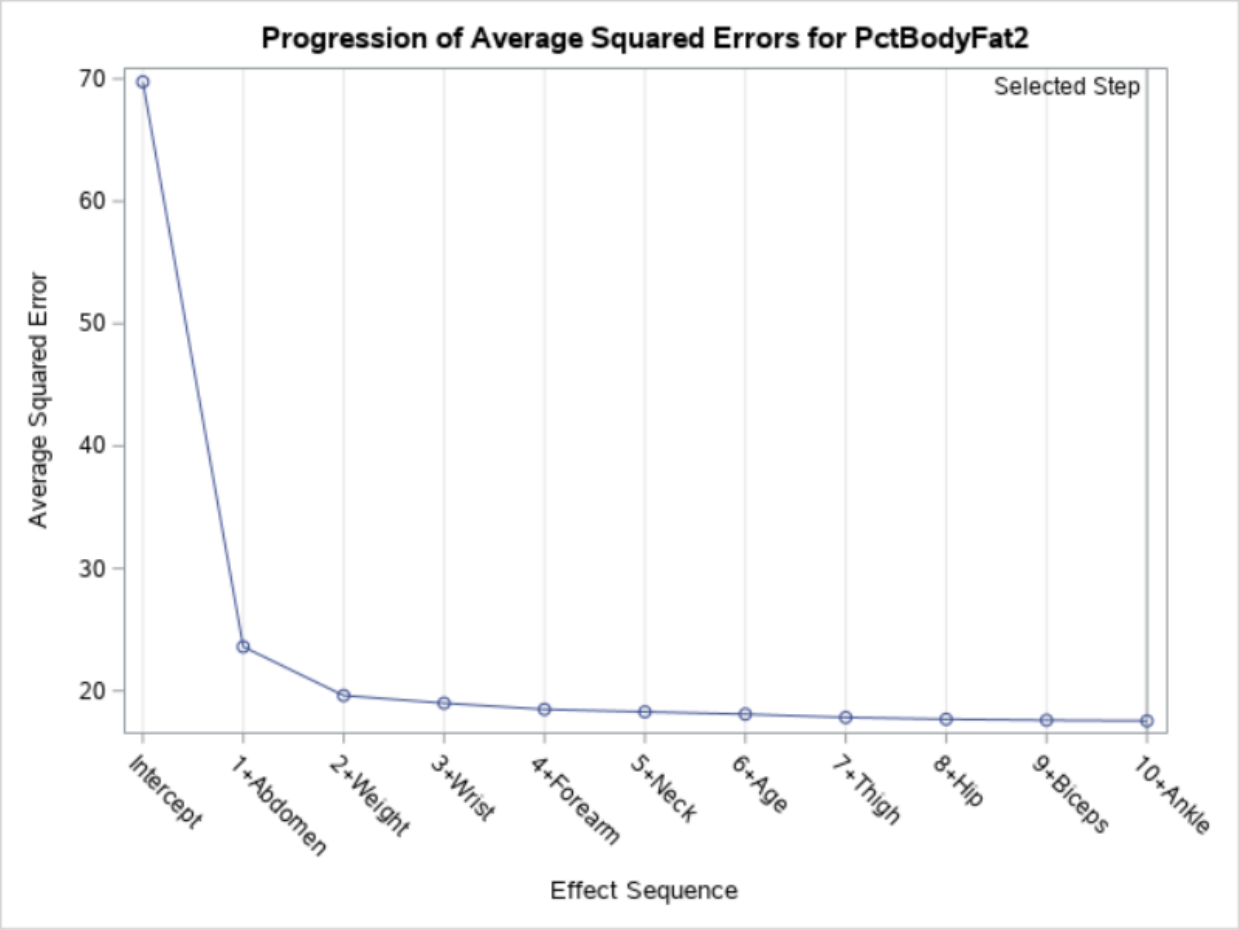
Forward Selection Summary				
Step	Effect Entered	Number Effects In	F Value	Pr > F
0	Intercept	1	0.00	1.0000
1	Abdomen	2	488.93	<.0001
2	Weight	3	50.58	<.0001
3	Wrist	4	8.15	0.0047
4	Forearm	5	6.78	0.0098
5	Neck	6	2.73	0.1000
6	Age	7	2.58	0.1098
7	Thigh	8	3.66	0.0569
8	Hip	9	1.99	0.1594
9	Biceps	10	1.13	0.2888
10	Ankle	11	0.72	0.3957

Selection stopped as the candidate for entry has SLE > 0.5.

Stop Details					
Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	Height	0.8473	>	0.5000	(SLE)


Coefficient Progression for PctBodyFat2
Standardized Coefficient
1.0
0.5
0.0
-0.5
Abdomen
Thigh
Biceps
Wrist
Weight
p-value
0.4
0.3
0.2
0.1
0.0
Selected Step
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
5+Neck
6+Age
7+Thigh
8+Hip
9+Biceps
10+Ankle
Effect Sequence


Fit Criteria for PctBodyFat2
AIC
AICC
SBC
Adj R-Sq
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
5+Neck
6+Age
7+Thigh
8+Hip
9+Biceps
10+Ankle
Effect Sequence
Effect Sequence
Best Criterion Value
Selected Step


Progression of Average Squared Errors for PctBodyFat2
Selected Step
Average Squared Error
70
60
50
40
30
20
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
5+Neck
6+Age
7+Thigh
8+Hip
9+Biceps
10+Ankle
Effect Sequence

SL FORWARD Selection with PctBodyFat2

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 10).

Effects:	Intercept Age Weight Neck Abdomen Hip Thigh Ankle Biceps Forearm Wrist

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	10	13158	1315.76595	71.72
Error	241	4421.33035	18.34577	
Corrected Total	251	17579		

Root MSE	4.28320
Dependent Mean	19.15079
R-Square	0.7485
Adj R-Sq	0.7381
AIC	997.92124
AICC	999.22668
SBC	782.74496

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-25.999624	12.153156	-2.14
Age	1	0.065093	0.030919	2.11
Weight	1	-0.107396	0.042068	-2.55
Neck	1	-0.467490	0.228115	-2.05
Abdomen	1	0.957721	0.072760	13.16
Hip	1	-0.179124	0.139083	-1.29
Thigh	1	0.259259	0.133892	1.94
Ankle	1	0.184526	0.216864	0.85
Biceps	1	0.186171	0.168580	1.10
Forearm	1	0.453031	0.195932	2.31
Wrist	1	-1.656662	0.527061	-3.14

```
/*st104s01.sas*/ /*Part C*/
proc glmselect data=STAT1.bodyfat2 plots=all;
        FORWARDSL: model PctBodyFat2 = Age Weight Height Neck Chest Abdomen
                Hip Thigh Knee Ankle Biceps Forearm Wrist
                / SELECTION=FORWARD SELECT=SL SLENTRY=0.05;
        title 'SL FORWARD (0.05) Selection with PctBodyFat2';
run;
```

SL FORWARD (0.05) Selection with PctBodyFat2

The GLMSELECT Procedure

Data Set	STAT1.BODYFAT2
Dependent Variable	PctBodyFat2
Selection Method	Forward
Select Criterion	Significance Level
Stop Criterion	Significance Level
Entry Significance Level (SLE)	0.05
Effect Hierarchy Enforced	None

Number of Observations Read	252
Number of Observations Used	252

Dimensions	
Number of Effects	14
Number of Parameters	14

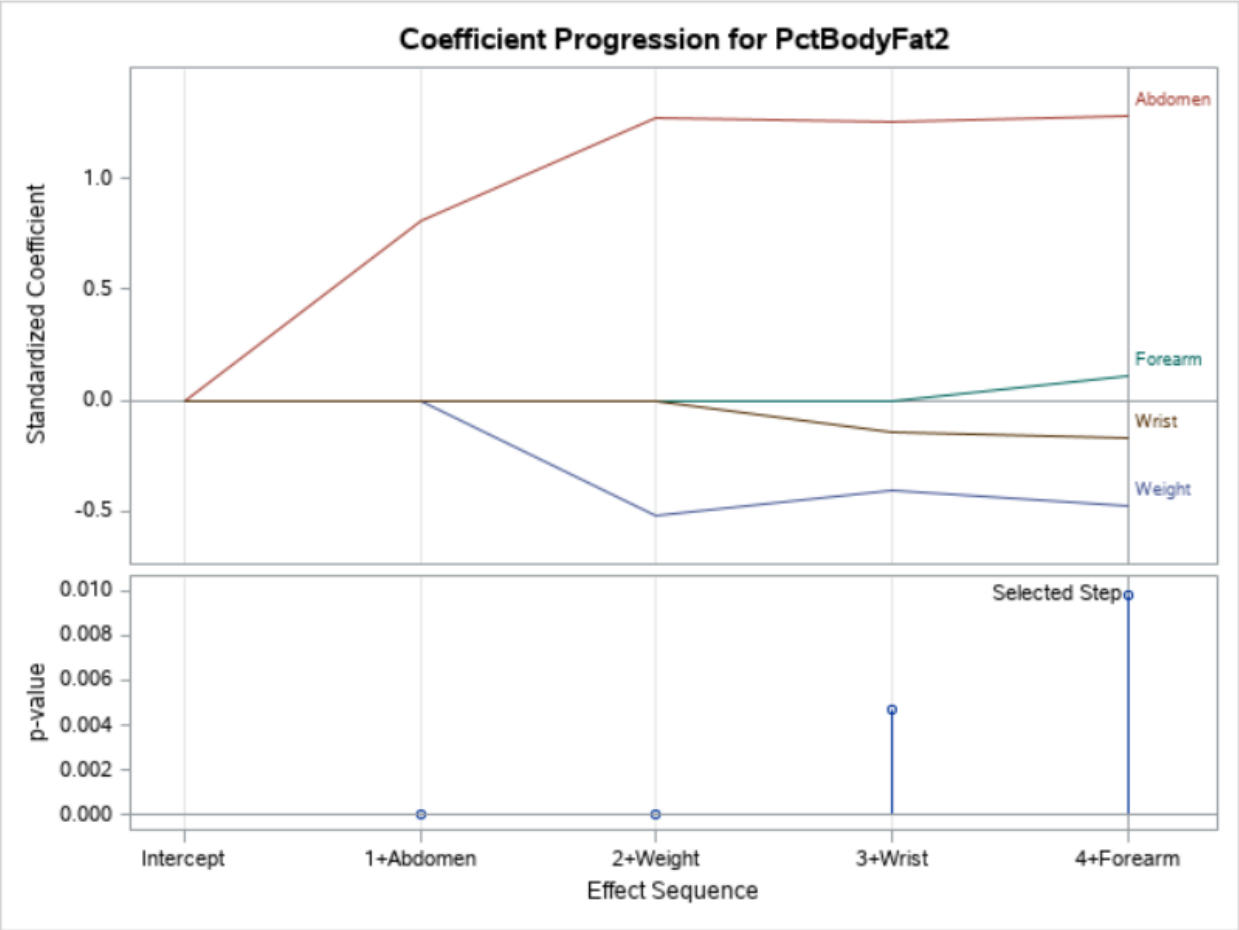
SL FORWARD (0.05) Selection with PctBodyFat2

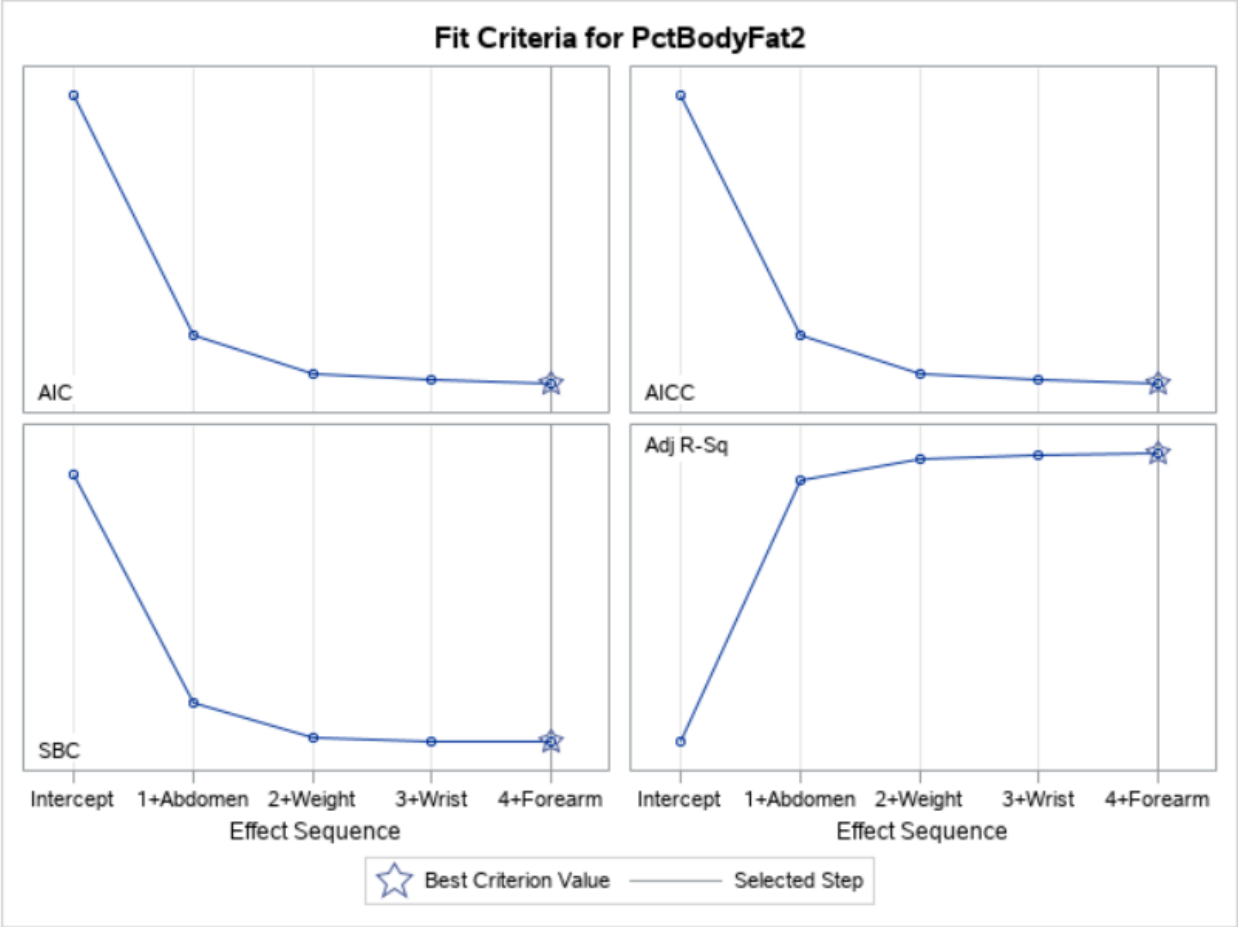
The GLMSELECT Procedure

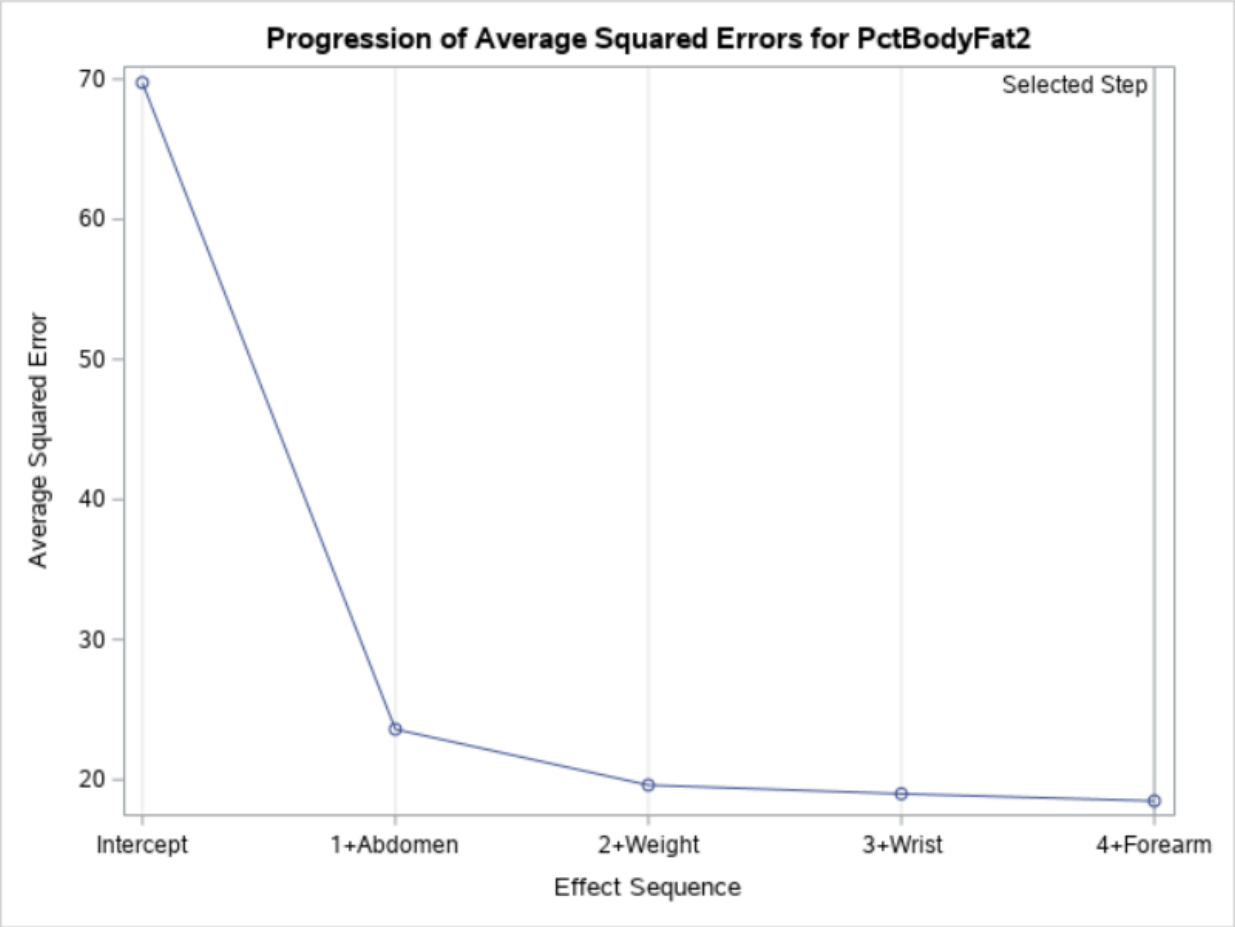
Forward Selection Summary				
Step	Effect Entered	Number Effects In	F Value	Pr > F
0	Intercept	1	0.00	1.0000
1	Abdomen	2	488.93	<.0001
2	Weight	3	50.58	<.0001
3	Wrist	4	8.15	0.0047
4	Forearm	5	6.78	0.0098

Selection stopped as the candidate for entry has SLE > 0.05.

Stop Details					
Candidate For	Effect	Candidate Significance		Compare Significance	
Entry	Neck	0.1000	>	0.0500	(SLE)


Coefficient Progression for PctBodyFat2
Standardized Coefficient
1.0
0.5
0.0
-0.5
Abdomen
Forearm
Wrist
Weight
p-value
0.010
0.008
0.006
0.004
0.002
0.000
Selected Step
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
Effect Sequence


Fit Criteria for PctBodyFat2
AIC
AICC
SBC
Adj R-Sq
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
Effect Sequence
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
Effect Sequence
Best Criterion Value
Selected Step


Progression of Average Squared Errors for PctBodyFat2
Selected Step
Average Squared Error
70
60
50
40
30
20
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
Effect Sequence

SL FORWARD (0.05) Selection with PctBodyFat2

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 4).

Effects:	Intercept Weight Abdomen Forearm Wrist

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	4	12921	3230.18852	171.28
Error	247	4658.23577	18.85925	
Corrected Total	251	17579		

Root MSE	4.34272
Dependent Mean	19.15079
R-Square	0.7350
Adj R-Sq	0.7307
AIC	999.07467
AICC	999.41753
SBC	762.72182

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-34.854074	7.245005	-4.81
Weight	1	-0.135631	0.024748	-5.48
Abdomen	1	0.995751	0.056066	17.76
Forearm	1	0.472928	0.181661	2.60
Wrist	1	-1.505562	0.442666	-3.40

Practice - Using PROC GLMSELECT to Perform Stepwise Selection

Question 1

Use the **stat1.bodyfat2** data set to identify a set of "best" models. Use significance-level model selection techniques.

1. With the SELECTION=STEPWISE option, use SELECT=SL in PROC GLMSELECT to identify a set of candidate models that predict **PctBodyFat2** as a function of the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**. Use the default values for SLENTRY= and SLSTAY=.
2. Submit the code. What do you notice about the results?

Solution code:

```
/*st104s01.sas*/ /*Part A*/

ods graphics on;
proc glmselect data=stat1.bodyfat2 plots=all;
   STEPWISESL: model PctBodyFat2=Age Weight Height Neck Chest
                                 Abdomen Hip Thigh Knee Ankle Biceps
                                 Forearm Wrist
                                 / SELECTION=STEPWISE SELECT=SL;
   title 'SL STEPWISE Selection with PctBodyFat2';
run;
```

In the results, notice the following:

- Selection stopped because the candidate for entry has SLE > 0.15 and the candidate for removal has SLS < 0.15.
- The stepwise selection process, using significance level, seems to select an eight-effect model (including the intercept).
- The Coefficient panel shows that the standardized coefficients do not vary greatly when additional effects are added to the model.
- The Fit panel indicates that the best model, according to AIC, AICC, and adjusted R-square, is the final model viewed during the selection process. SBC shows a minimum at step four.
- The parameter estimates from the selected model are presented in the Parameter Estimates table.

Question 2
Modify the code to specify the forward selection process (FORWARD). Submit the code. What do you notice about the results?

Solution code:

```
/*st104s01.sas*/ /*Part B*/

proc glmselect data=stat1.bodyfat2 plots=all;
   FORWARDSL: model PctBodyFat2=Age Weight Height
                                Neck Chest Abdomen Hip Thigh
                                Knee Ankle Biceps Forearm Wrist
                                / SELECTION=FORWARD SELECT=SL;
   title 'SL FORWARD Selection with PctBodyFat2';
run;
```

In the results, notice the following:

- Selection stopped as the candidate for entry has SLE > 0.5.
- The forward selection process, using significance level, seems to select an 11-effect model (including the intercept).
- The Coefficient panel shows that the standardized coefficients do not vary greatly when additional effects are added to the model.
- The Fit panel indicates that the best models, according to AIC, AICC, adjusted R-square, and SBC, are at various steps in the selection progression.
- The parameter estimates from the selected model are presented in the Parameter Estimates table.

Question 3

Modify the code to use a significance-level-for-entry criterion of 0.05, instead of the default SLENTRY= value, 0.50. Submit the code, and view the results.

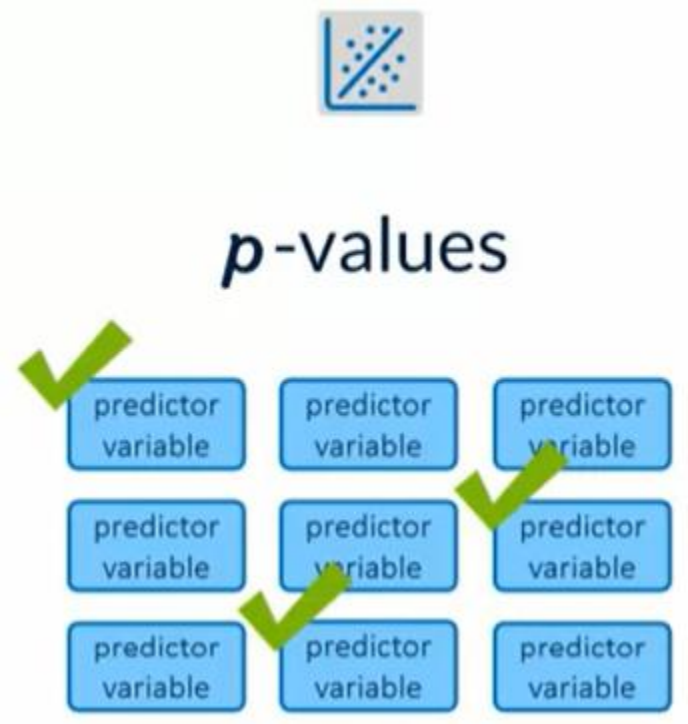
How many variables would result from a model using forward selection and a significance-level-for-entry criterion of 0.05?

Solution code:

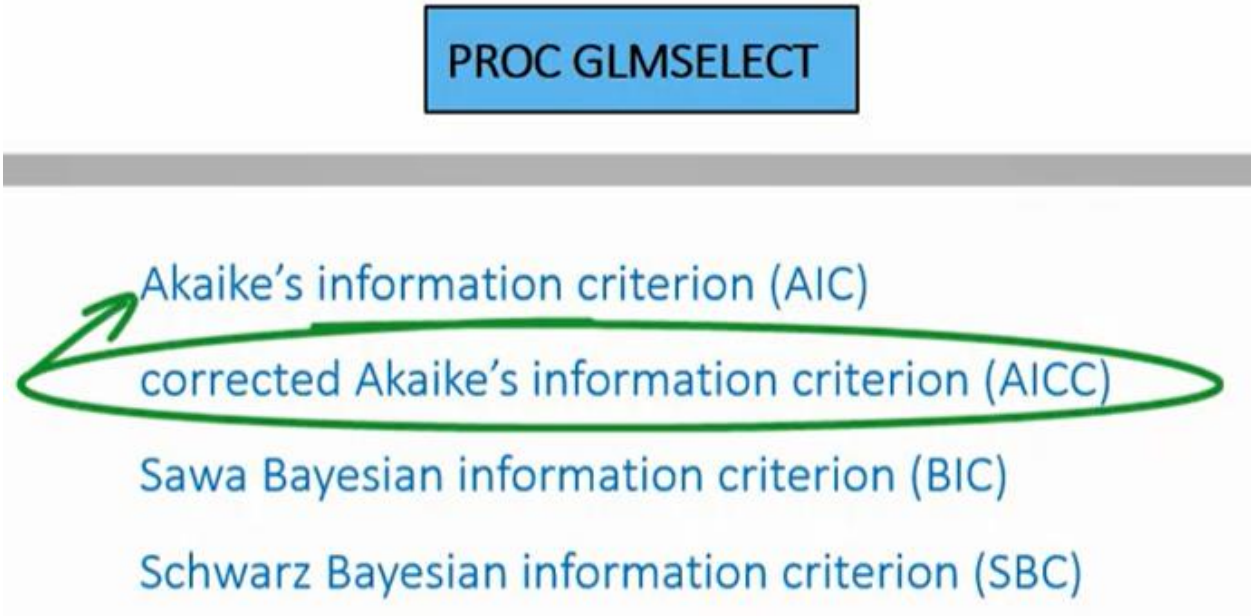
```
/*st104s01.sas*/ /*Part C*/

proc glmselect data=stat1.bodyfat2 plots=all;
   FORWARDSL: model PctBodyFat2=Age Weight Height
                                Neck Chest Abdomen Hip Thigh
                                Knee Ankle Biceps Forearm Wrist
                                / SELECTION=FORWARD SELECT=SL
                                SLENTRY=0.05;
    title 'SL FORWARD (0.05) Selection with PctBodyFat2';
run;
```

The results show that when the value of SLENTRY= is changed from the default to 0.05, the number of effects in the selected model is reduced to five (including the intercept).

Information Criterion and Other Selection Options
Scenario



Information Criteria



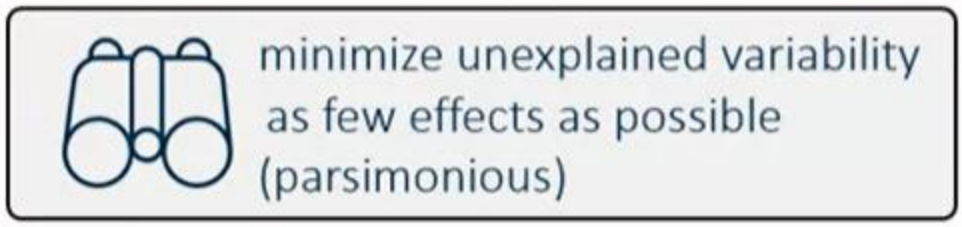
PROC GLMSELECT

Akaike's information criterion (AIC)

corrected Akaike's information criterion (AICC)

Sawa Bayesian information criterion (BIC)

Schwarz Bayesian information criterion (SBC)

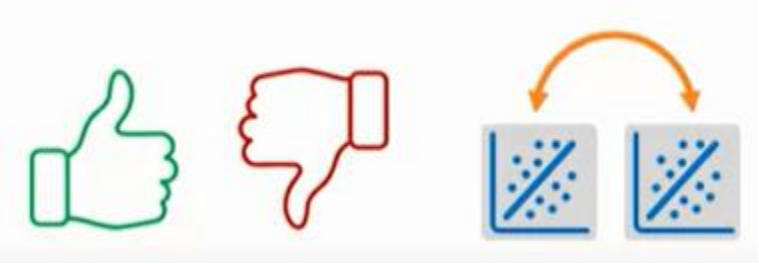


Akaike's information criterion (AIC)

corrected Akaike's information criterion (AICC)

Sawa Bayesian information criterion (BIC)

Schwarz Bayesian information criterion (SBC)



$$n\log\left(\frac{SSE}{n}\right)$$

information criterion	penalty component
AIC	$2p+n+2$
AICC	$\frac{n(n+p)}{n-p-2}$
BIC	$2(p+2)q-2q^2$
SBC	$p\log(n)$

Information Criteria Penalty Components

Beyond significance level, there are several statistics, referred to as information criteria, that can be used to evaluate competing models as well as direct the selection process within the GLMSELECT procedure. Each criterion searches for a model that will minimize the unexplained variability using as few effects within the model as possible. The model with as few effects as possible is referred to as the most parsimonious model. The calculation for each information criterion begins with

$$n\log(\frac{SSE}{n})$$

It then invokes a penalty representing the complexity of the model. The table below shows the penalty for each criterion, where n is the number of observations, p is the number of parameters including the intercept, and

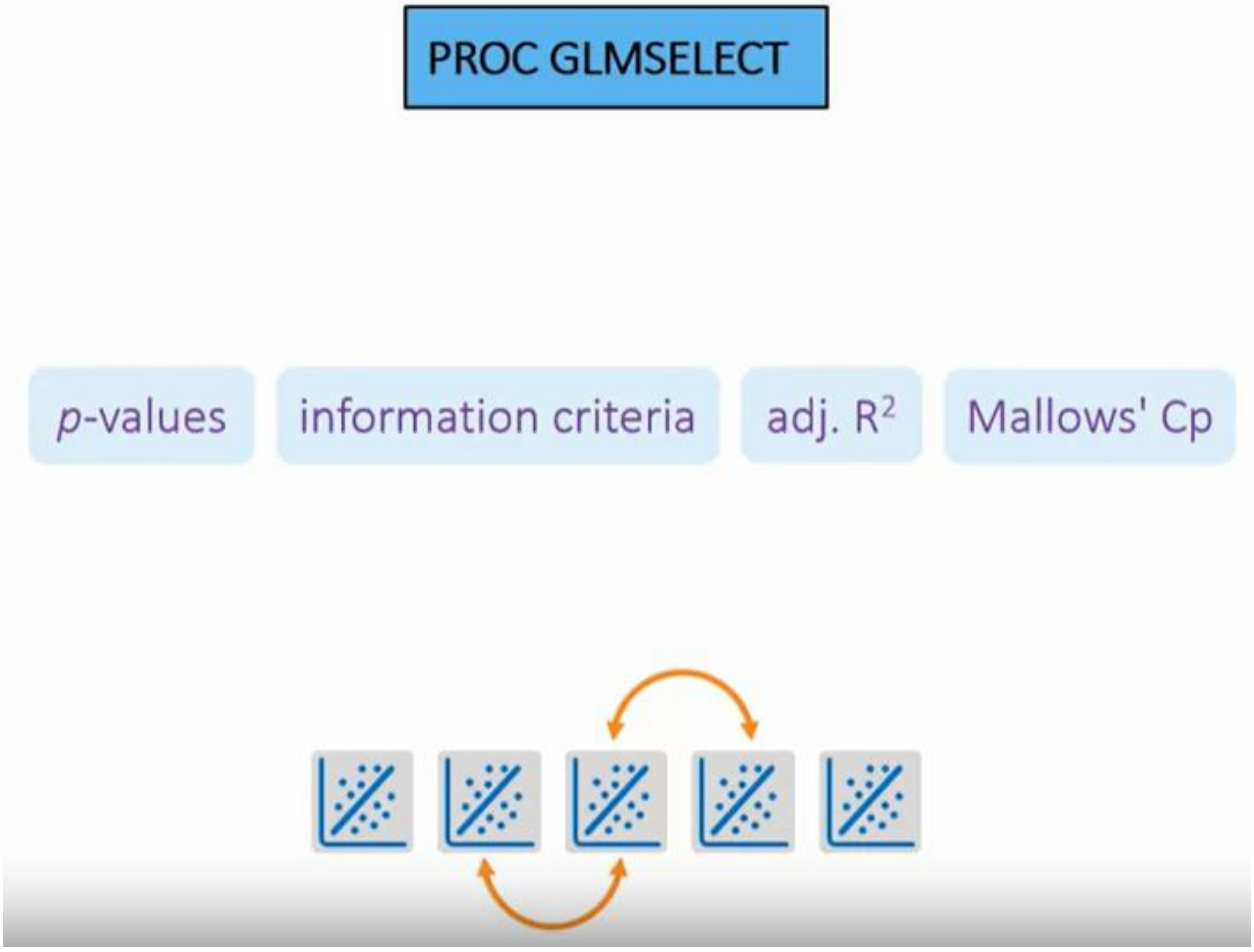
$$\widehat{\sigma}^2$$

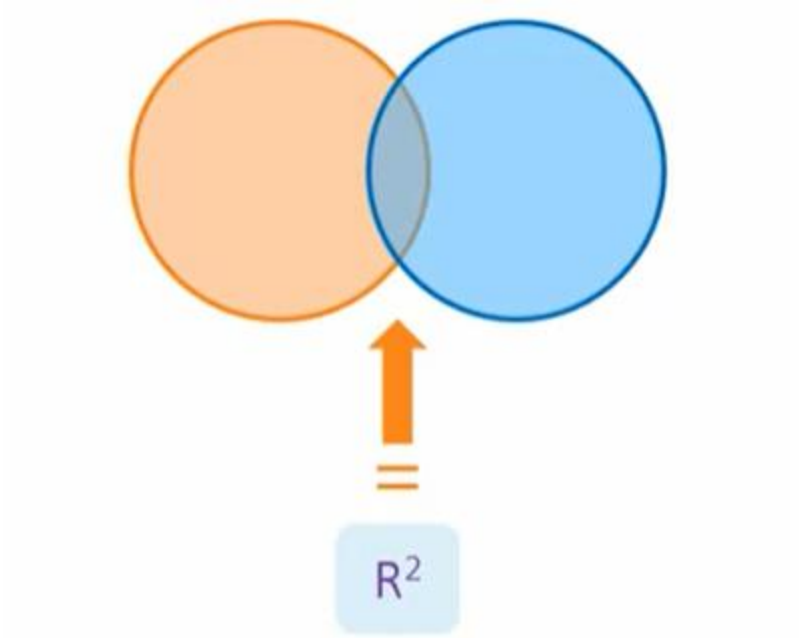
is the estimate of pure error variance from fitting the full model. For each information criterion, smaller is better. Note: In the BIC penalty,

$$q = \frac{n\widehat{\sigma}^2}{SSE}$$

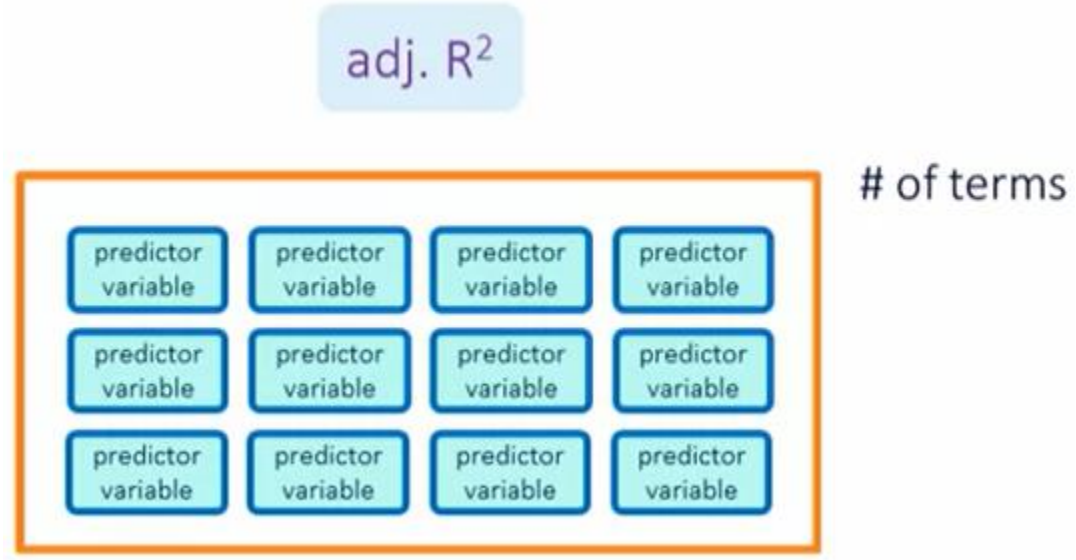
Information Criterion	Penalty Component
AIC	$2p + n + 2$
AICC	$\frac{n(n+p)}{n-p-2}$
BIC	$2(p + 2)q - 2q^2$
SBC	$p\log(n)$

Adjusted R-Square and Mallow's Cp



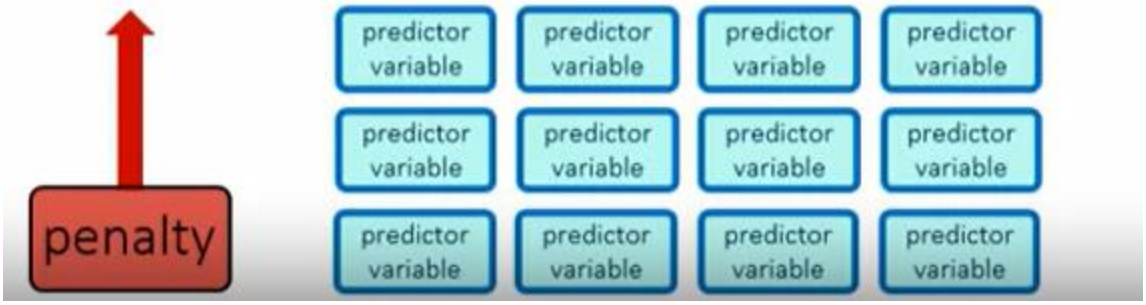

R^2


predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable


adj. R^2
# of terms
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable

$$R^2_{ADJ} = 1 - \frac{(n-i)(1-R^2)}{n-p}$$

adj. R^2



Demo Performing Model Selection Using PROC GLMSELECT

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
        Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st104d02.sas*/
ods graphics on;
proc glmselect data=STAT1.ameshousing3 plots=all;
    STEPWISEAIC: model SalePrice = &interval / selection=stepwise details=steps select=AIC;
    title "Stepwise Model Selection for SalePrice - AIC";
run;

proc glmselect data=STAT1.ameshousing3 plots=all;
    STEPWISEBIC: model SalePrice = &interval / selection=stepwise details=steps select=BIC;
    title "Stepwise Model Selection for SalePrice - BIC";
run;

proc glmselect data=STAT1.ameshousing3 plots=all;
    STEPWISEAICC: model SalePrice = &interval / selection=stepwise details=steps select=AICC;
    title "Stepwise Model Selection for SalePrice - AICC";
run;

proc glmselect data=STAT1.ameshousing3 plots=all;
    STEPWISESBC: model SalePrice = &interval / selection=stepwise details=steps select=SBC;
    title "Stepwise Model Selection for SalePrice - SBC";
run;
```

PROC GLMSELECT DATA=*SAS-data-set <options>*;
*<label:>***MODEL** *dependent=regressors< / options>*;
RUN;

Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure
Stepwise Selection: Step 0

Effect Entered: Intercept

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	0	0	.	.
Error	299	4.232235E11	1415463276	
Corrected Total	299	4.232235E11		

Root MSE	37623
Dependent Mean	137525
R-Square	0.0000
Adj R-Sq	0.0000
AIC	6624.21515
AICC	6624.25555
SBC	6325.91893

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	137525	2172.144314	63.31

Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure
Stepwise Selection: Step 1

Effect Entered: Basement_Area

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	1	2.012418E11	2.012418E11	270.16
Error	298	2.219817E11	744904950	
Corrected Total	299	4.232235E11		

Root MSE	27293
Dependent Mean	137625
R-Square	0.4755
Adj R-Sq	0.4737
AIC	6432.62346
AICC	6432.70454
SBC	6138.03102

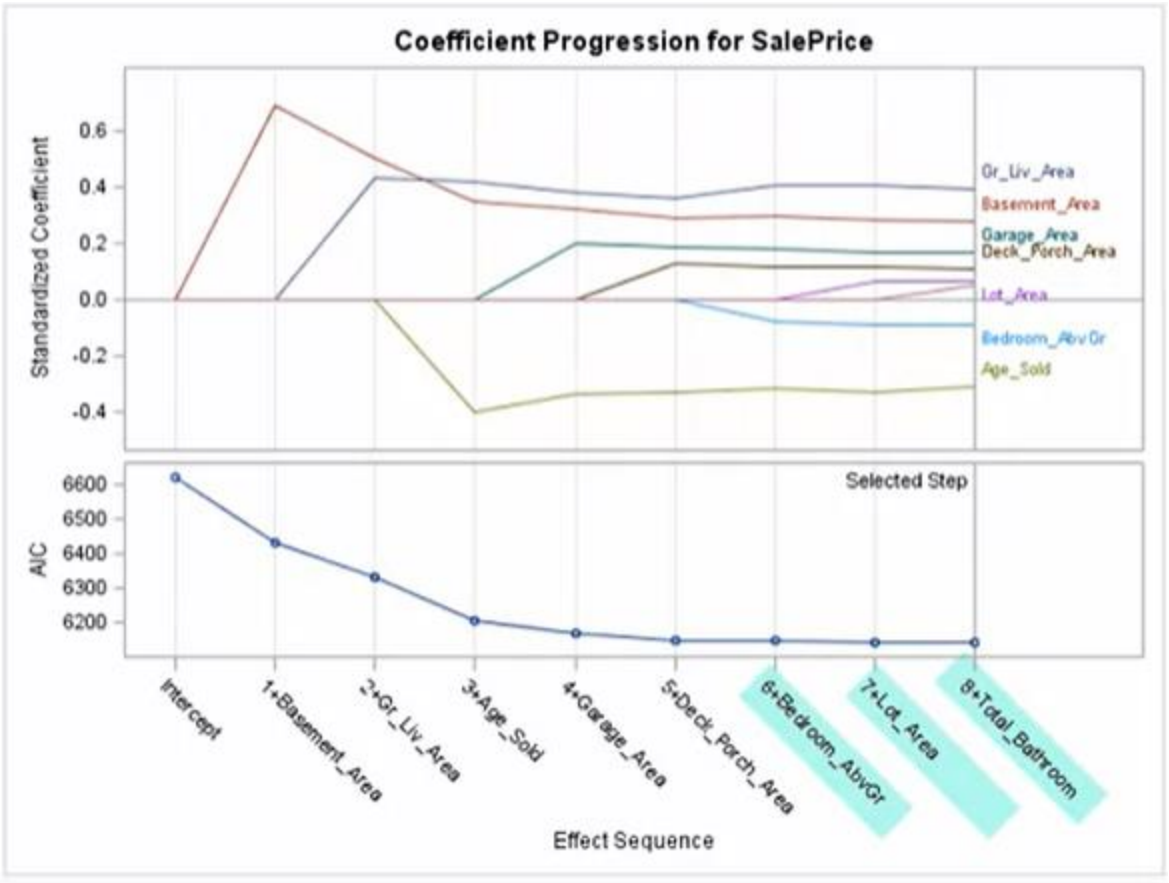
Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	73904	4179.193780	17.68
Basement_Area	1	72.107717	4.387055	16.44

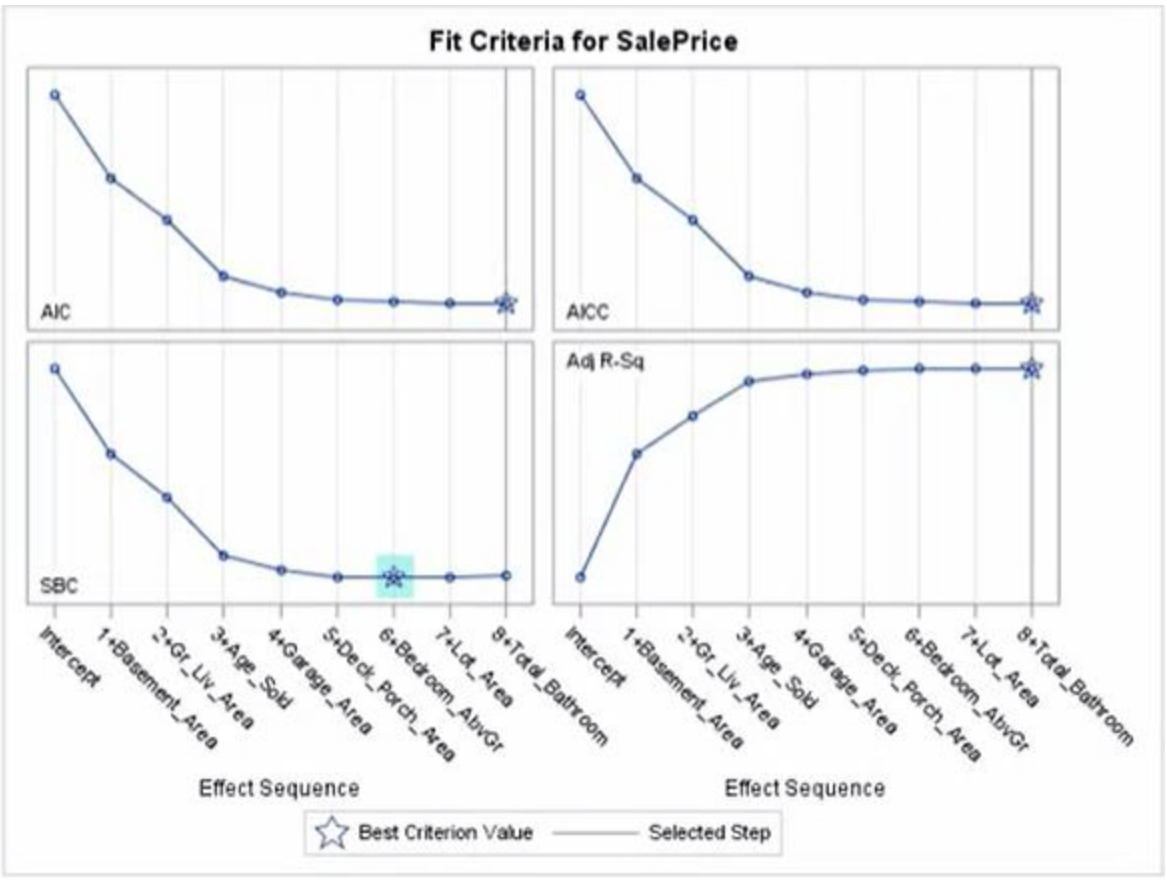
Stepwise Model Selection for SalePrice - AIC

The GLMSELECT Procedure

Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	AIC
0	Intercept		1	6624.2151
1	Basement_Area		2	6432.6235
2	Gr_Liv_Area		3	6334.0262
3	Age_Sold		4	6204.8293
4	Garage_Area		5	6166.6273
5	Deck_Porch_Area		6	6148.8927
6	Bedroom_AbvGr		7	6144.4040
7	Lot_Area		8	6141.3368
8	Total_Bathroom		9	6140.7956*
* Optimal Value of Criterion				

Selection stopped because all effects are in the final model.


Coefficient Progression for SalePrice
Standardized Coefficient
0.6
0.4
0.2
0.0
-0.2
-0.4
Gr_Liv_Area
Basement_Area
Garage_Area
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Age_Sold
Selected Step
AIC
6600
6500
6400
6300
6200
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
8+Total_Bathroom
Effect Sequence



Stepwise Model Selection for SalePrice - BIC

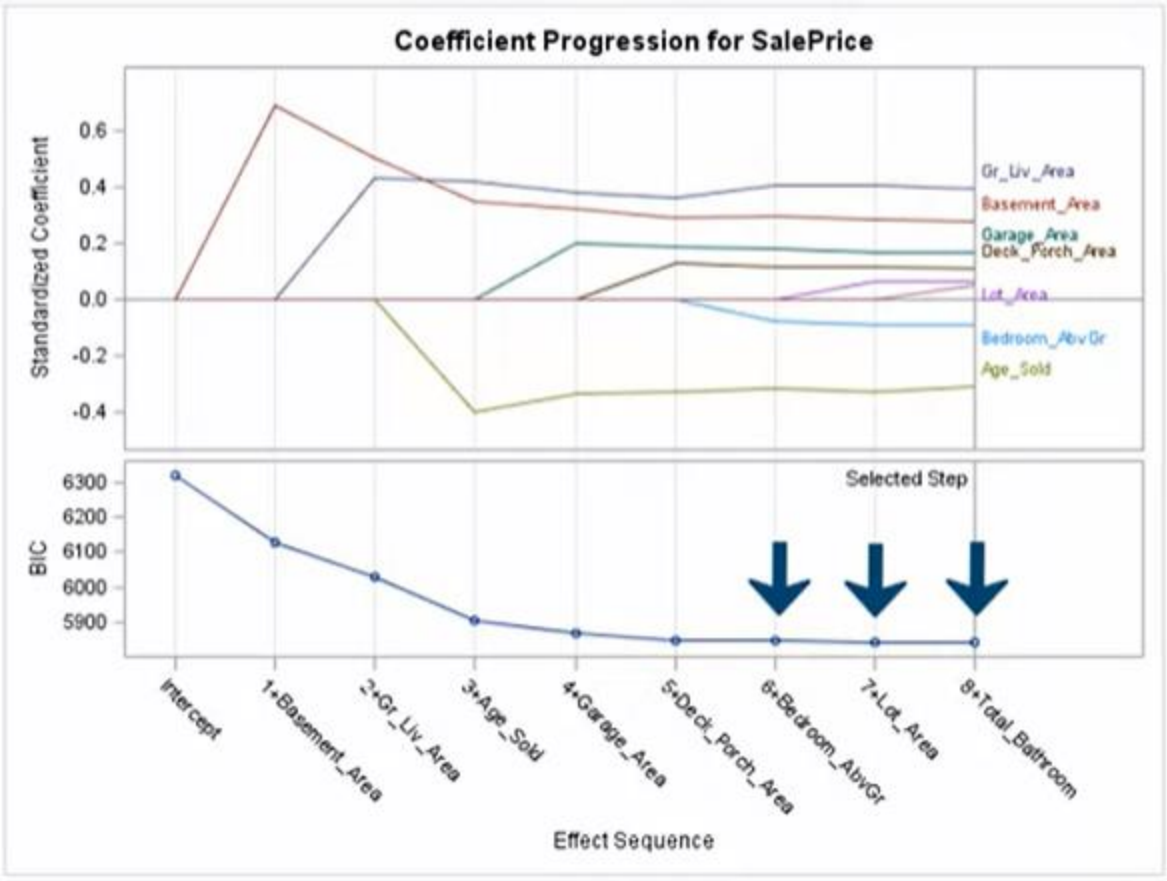
The GLMSELECT Procedure

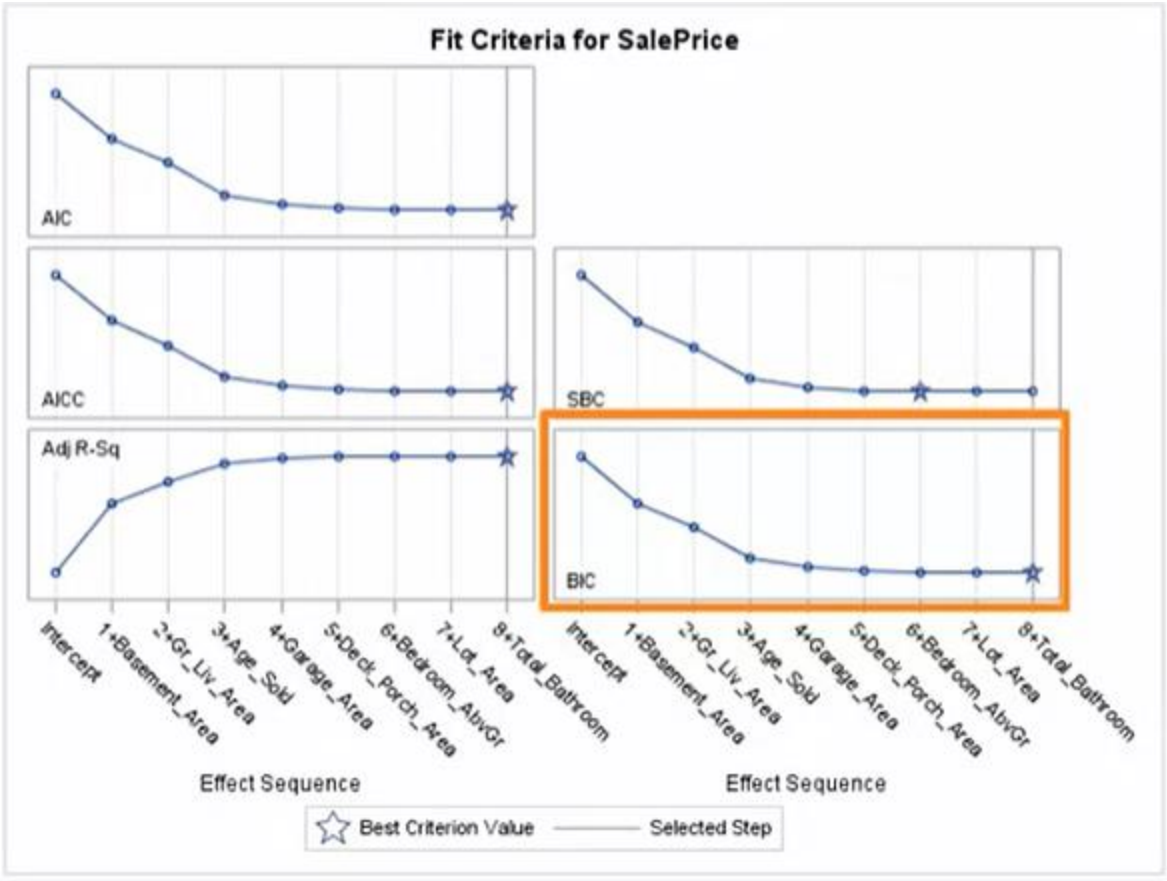
Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	BIC
0	Intercept		1	6321.3096
1	Basement_Area		2	6129.3224
2	Gr_Liv_Area		3	6030.6866
3	Age_Sold		4	5903.1974
4	Garage_Area		5	5865.8247
5	Deck_Porch_Area		6	5848.6954
6	Bedroom_AbvGr		7	5844.4755
7	Lot_Area		8	5841.6915
8	Total_Bathroom		9	5841.3504*

* Optimal Value of Criterion

Selection stopped because all effects are in the final model.


Coefficient Progression for SalePrice
Standardized Coefficient
0.6
0.4
0.2
0.0
-0.2
-0.4
Gr_Liv_Area
Basement_Area
Garage_Area
Deck_Porch_Area
Lot_Area
Bedroom_AbvGr
Age_Sold
Selected Step
BIC
6300
6200
6100
6000
5900
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
7+Lot_Area
8+Total_Bathroom
Effect Sequence



Stepwise Model Selection for SalePrice - SBC

The GLMSELECT Procedure

Stepwise Selection Summary

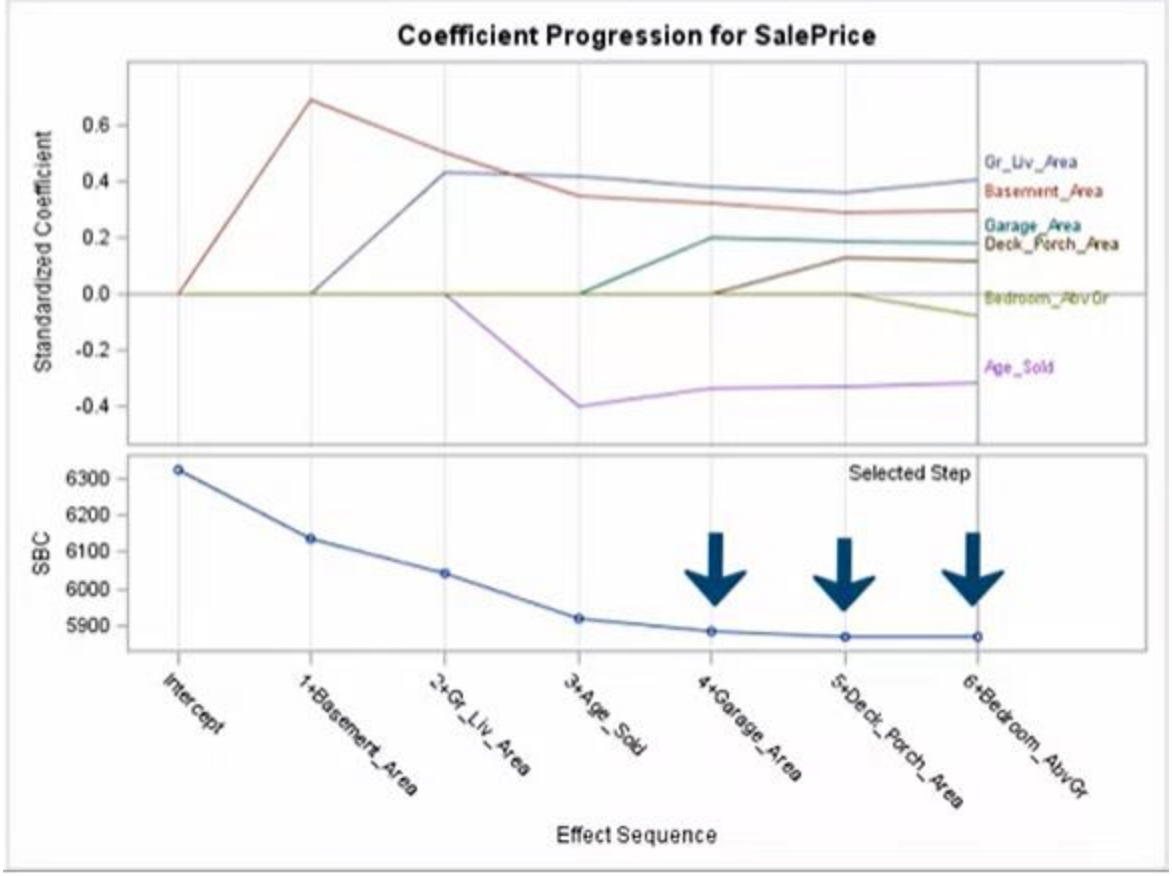
Step	Effect Entered	Effect Removed	Number Effects In	SBC
0	Intercept		1	6325.9189
1	Basement_Area		2	6138.0310
2	Gr_Liv_Area		3	6043.1375
3	Age_Sold		4	5917.6444
4	Garage_Area		5	5883.1463
5	Deck_Porch_Area		6	5869.1154
6	Bedroom_AbvGr		7	5868.3305*

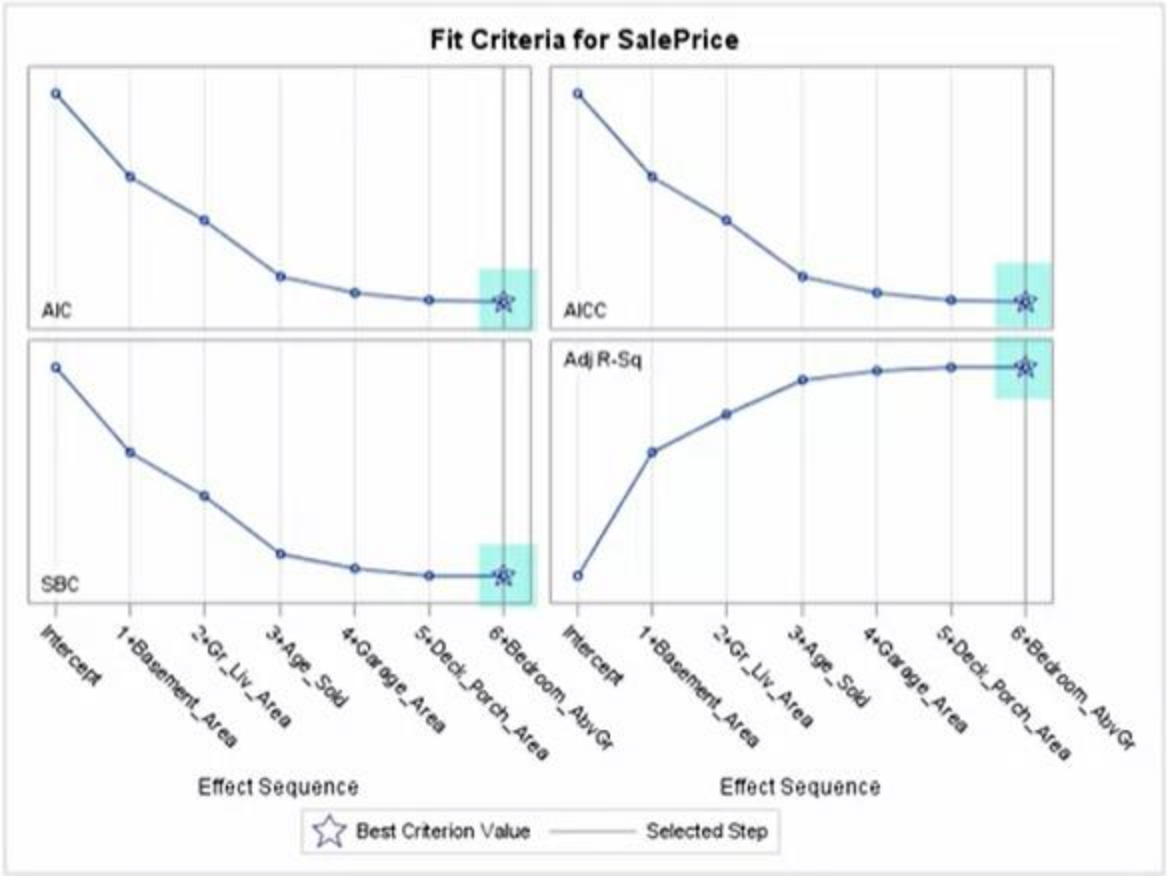
* Optimal Value of Criterion

Selection stopped at a local minimum of the SBC criterion.

Stop Details

Candidate For	Effect	Candidate SBC		Compare SBC
Entry	Lot_Area	5868.9670	>	5868.3305
Removal	Bedroom_AbvGr	5869.1154	>	5868.3305


Coefficient Progression for SalePrice
Standardized Coefficient
0.6
0.4
0.2
0.0
-0.2
-0.4
Gr_Liv_Area
Basement_Area
Garage_Area
Deck_Porch_Area
Bedroom_AbvGr
Age_Sold
SBC
6300
6200
6100
6000
5900
Selected Step
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Deck_Porch_Area
6+Bedroom_AbvGr
Effect Sequence

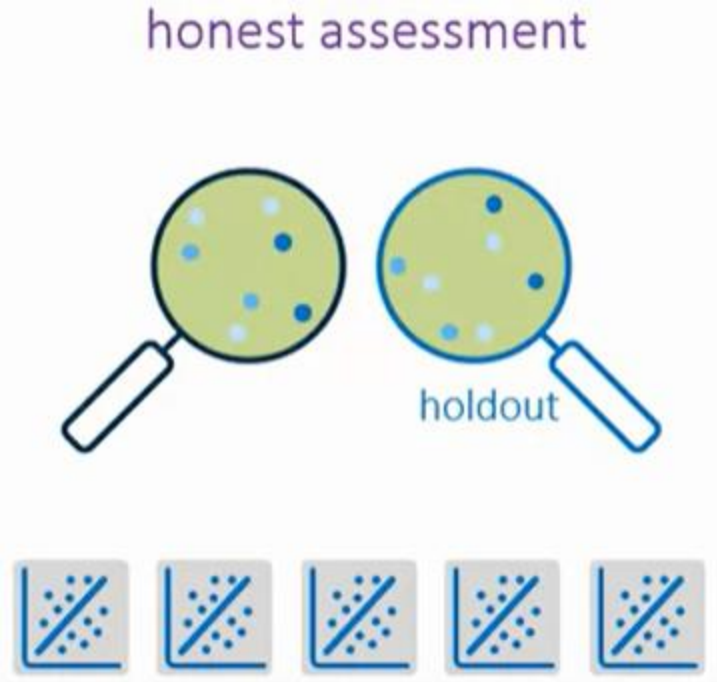


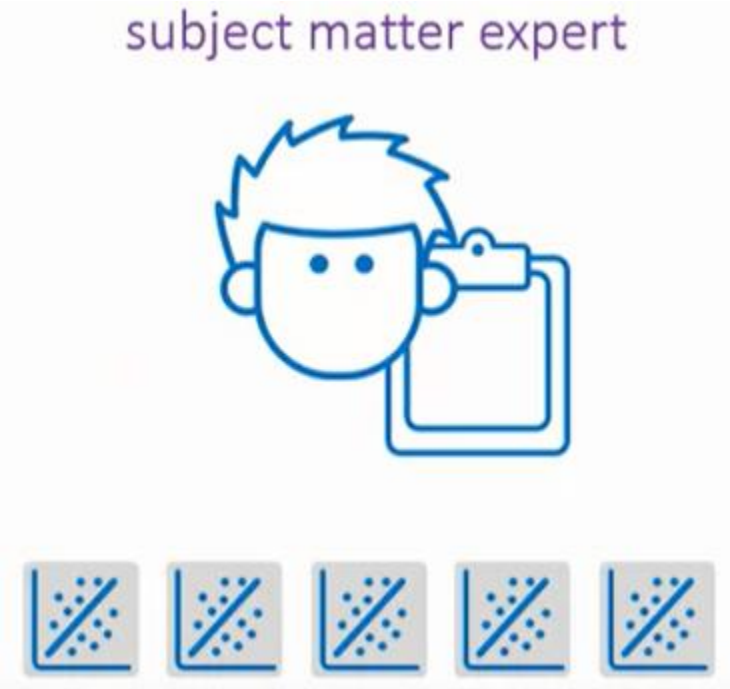
model-building strategies:

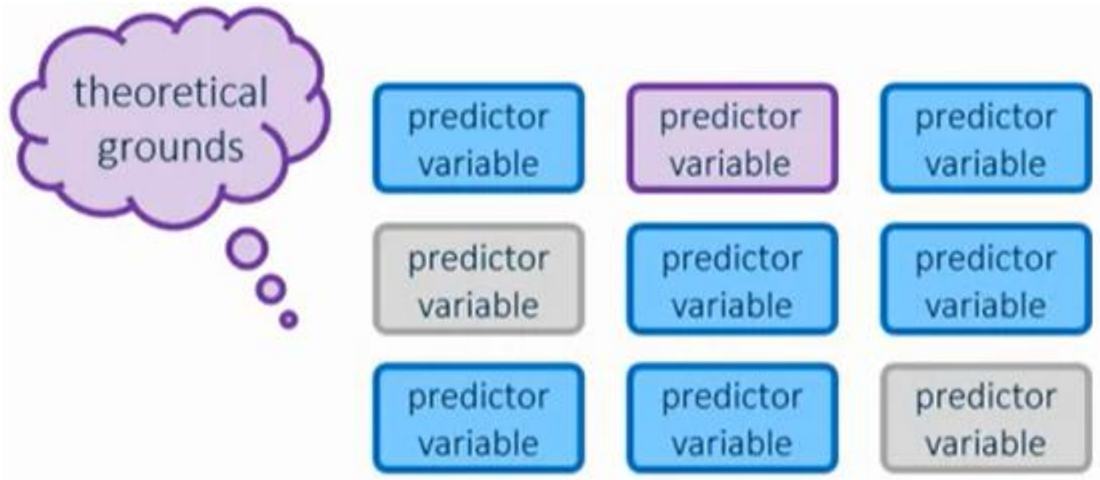
AIC, BIC, and AICC → model with all eight effects

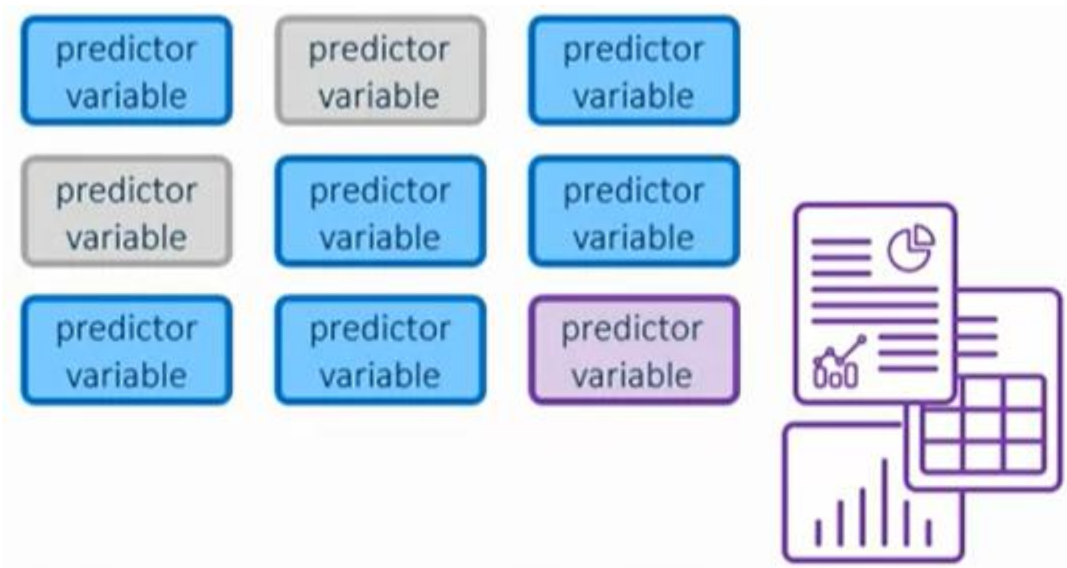
SBC → model with six effects

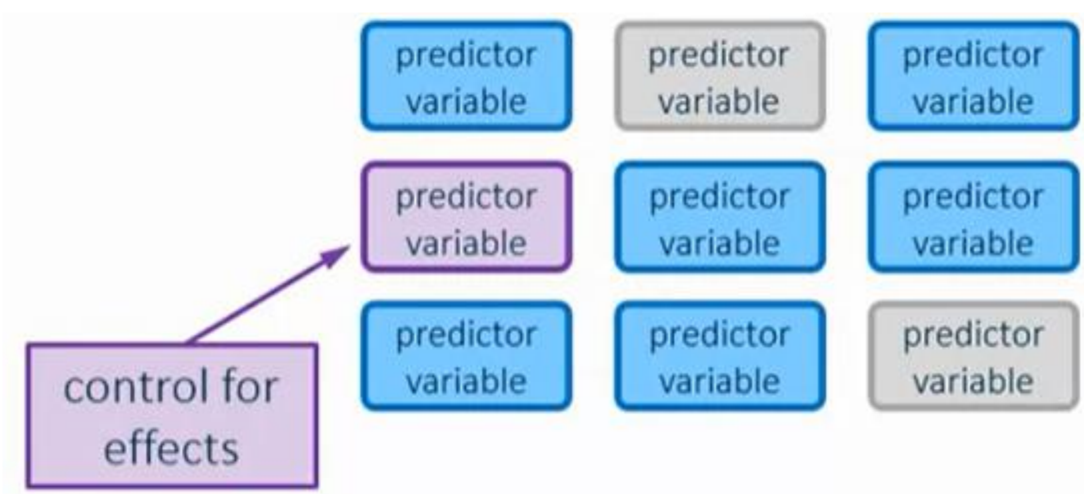
Stepwise, backward, and forward → same model with seven effects
(SLENTRY=0.05 and SLSTAY=0.05)

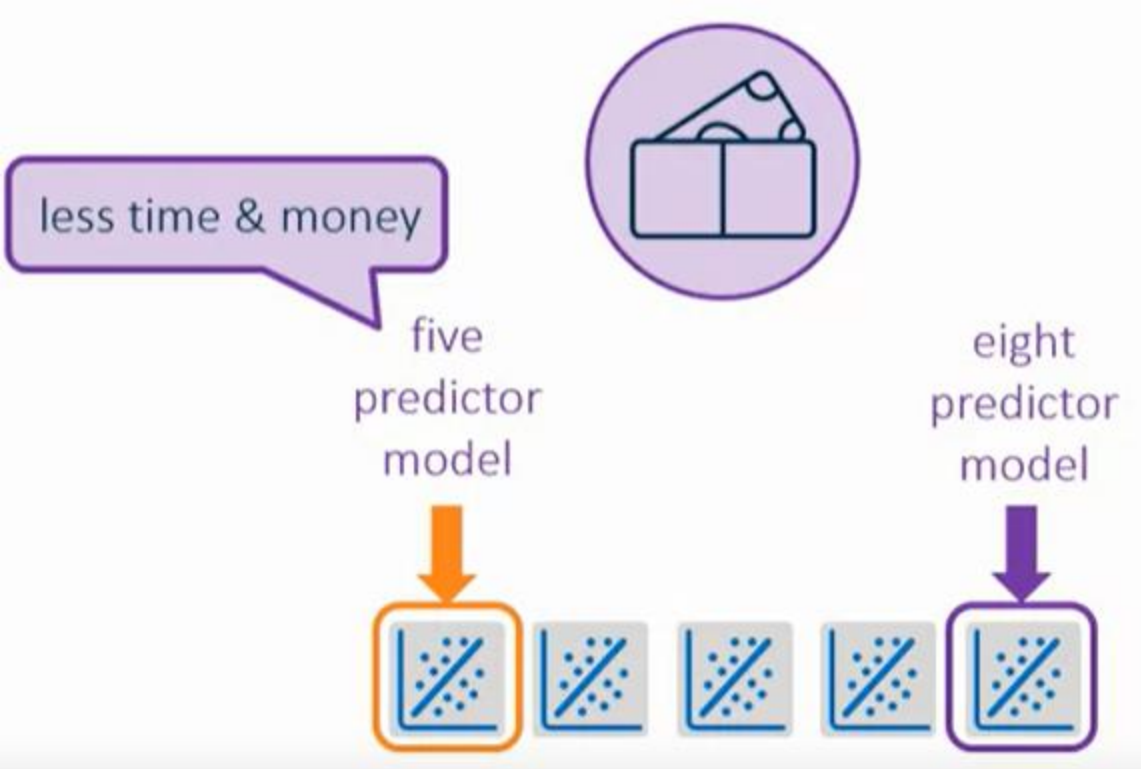

honest assessment
holdout


subject matter expert


theoretical grounds
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable


predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable


predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
predictor variable
control for effects




```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
      Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom ;

/*st104d02.sas*/
ods graphics on;
proc glmselect data=STAT1.ameshousing3 plots=all;
      STEPWISEAIC: model SalePrice = &interval / selection=stepwise details=steps select=AIC;
      title "Stepwise Model Selection for SalePrice - AIC";
```

```
run;

proc glmselect data=STAT1.ameshousing3 plots=all;
        STEPWISEBIC: model SalePrice = &interval / selection=stepwise details=steps select=BIC;
        title "Stepwise Model Selection for SalePrice - BIC";
run;

proc glmselect data=STAT1.ameshousing3 plots=all;
        STEPWISEAICC: model SalePrice = &interval / selection=stepwise details=steps select=AICC;
        title "Stepwise Model Selection for SalePrice - AICC";
run;

proc glmselect data=STAT1.ameshousing3 plots=all;
        STEPWISESBC: model SalePrice = &interval / selection=stepwise details=steps select=SBC;
        title "Stepwise Model Selection for SalePrice - SBC";
run;
```

```
/*st104s02.sas*/ /*Part A*/
ods graphics on;
proc glmselect data=STAT1.bodyfat2 plots=all;
        STEPWISESBC: model PctBodyFat2 = Age Weight Height Neck Chest Abdomen
                Hip Thigh Knee Ankle Biceps Forearm Wrist
                / SELECTION=STEPWISE SELECT=SBC;
        title 'SBC STEPWISE Selection with PctBodyFat2';
run;
```

SBC STEPWISE Selection with PctBodyFat2

The GLMSELECT Procedure

Data Set	STAT1.BODYFAT2
Dependent Variable	PctBodyFat2
Selection Method	Stepwise
Select Criterion	SBC
Stop Criterion	SBC
Effect Hierarchy Enforced	None

Number of Observations Read	252
Number of Observations Used	252

Dimensions	
Number of Effects	14
Number of Parameters	14

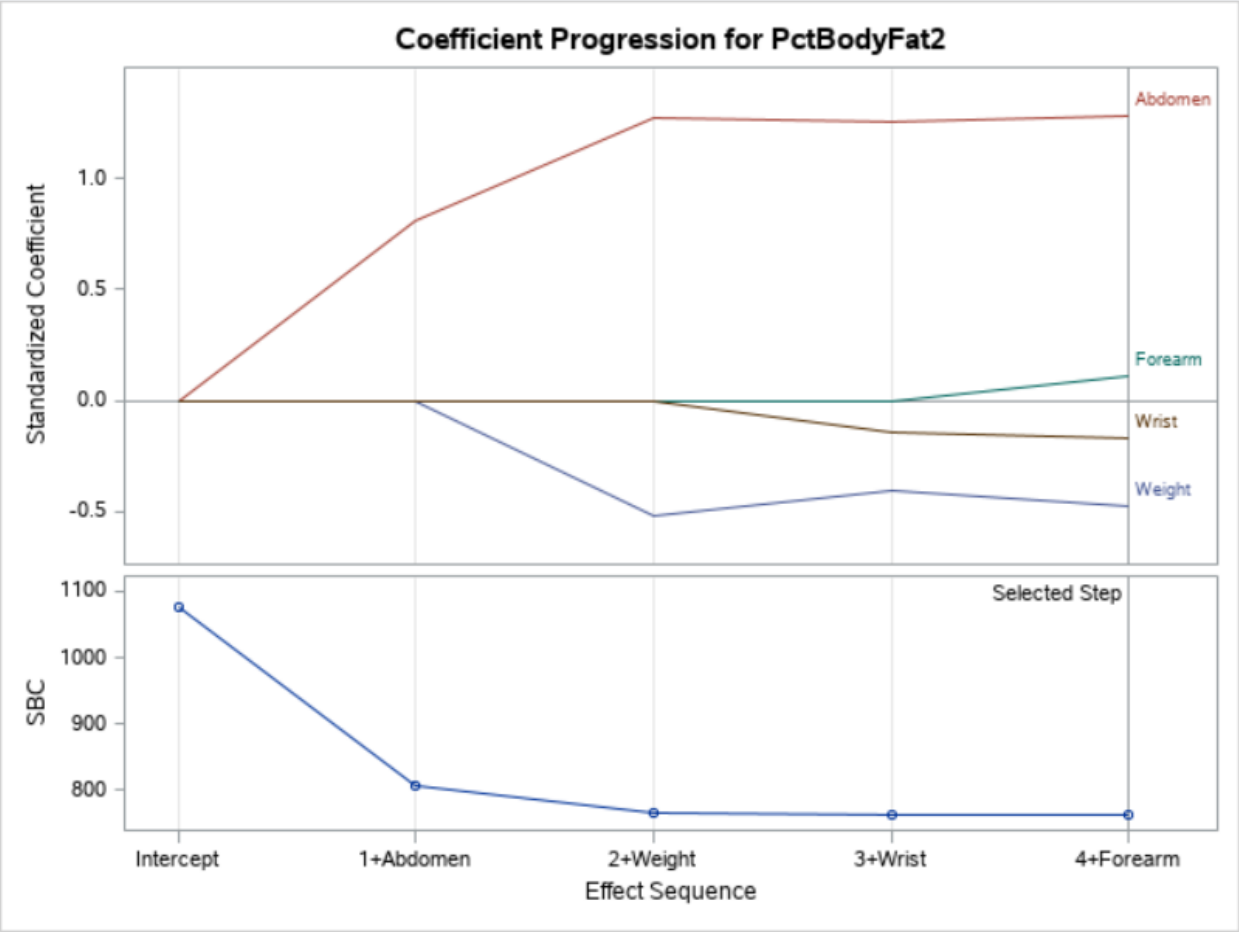
SBC STEPWISE Selection with PctBodyFat2

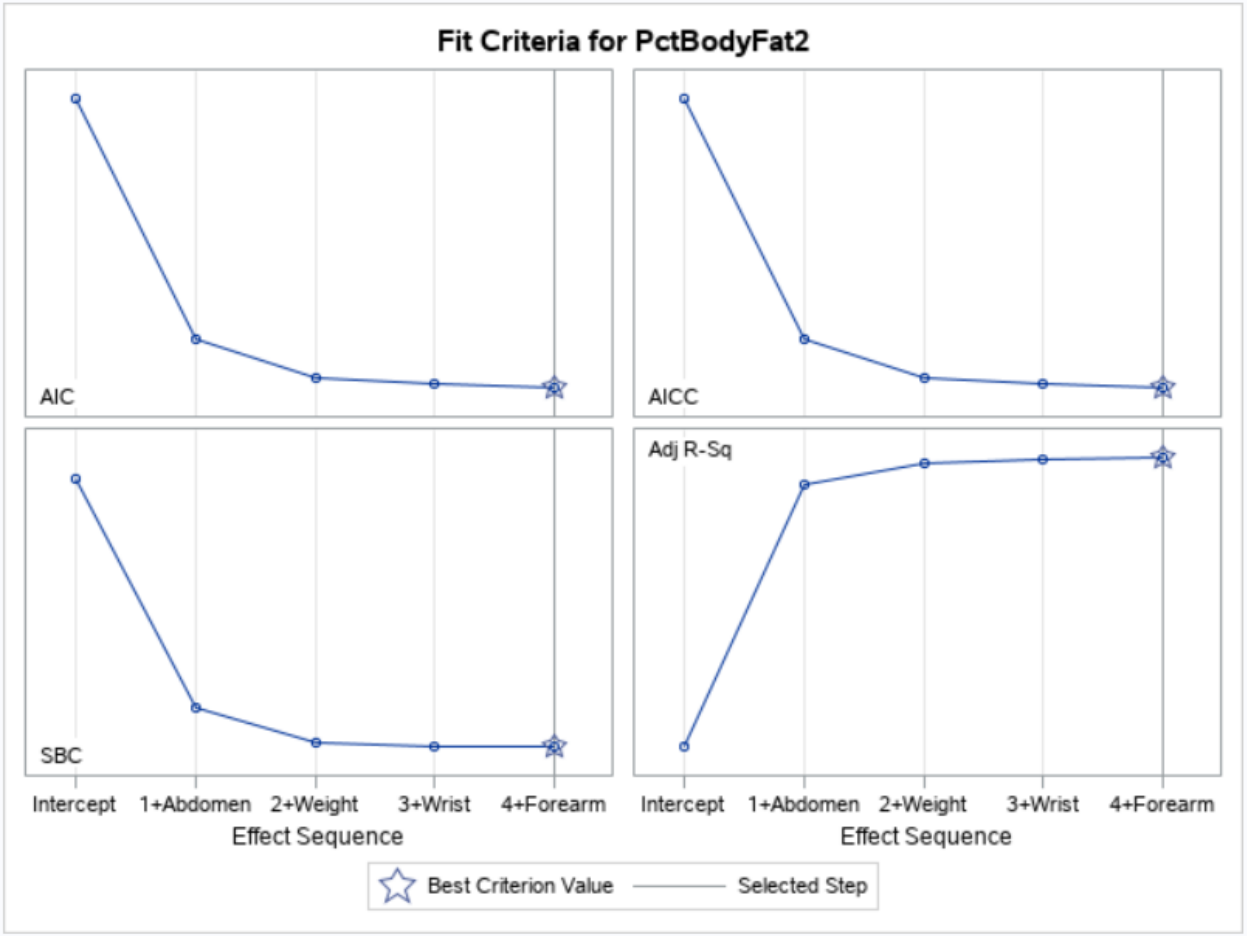
The GLMSELECT Procedure

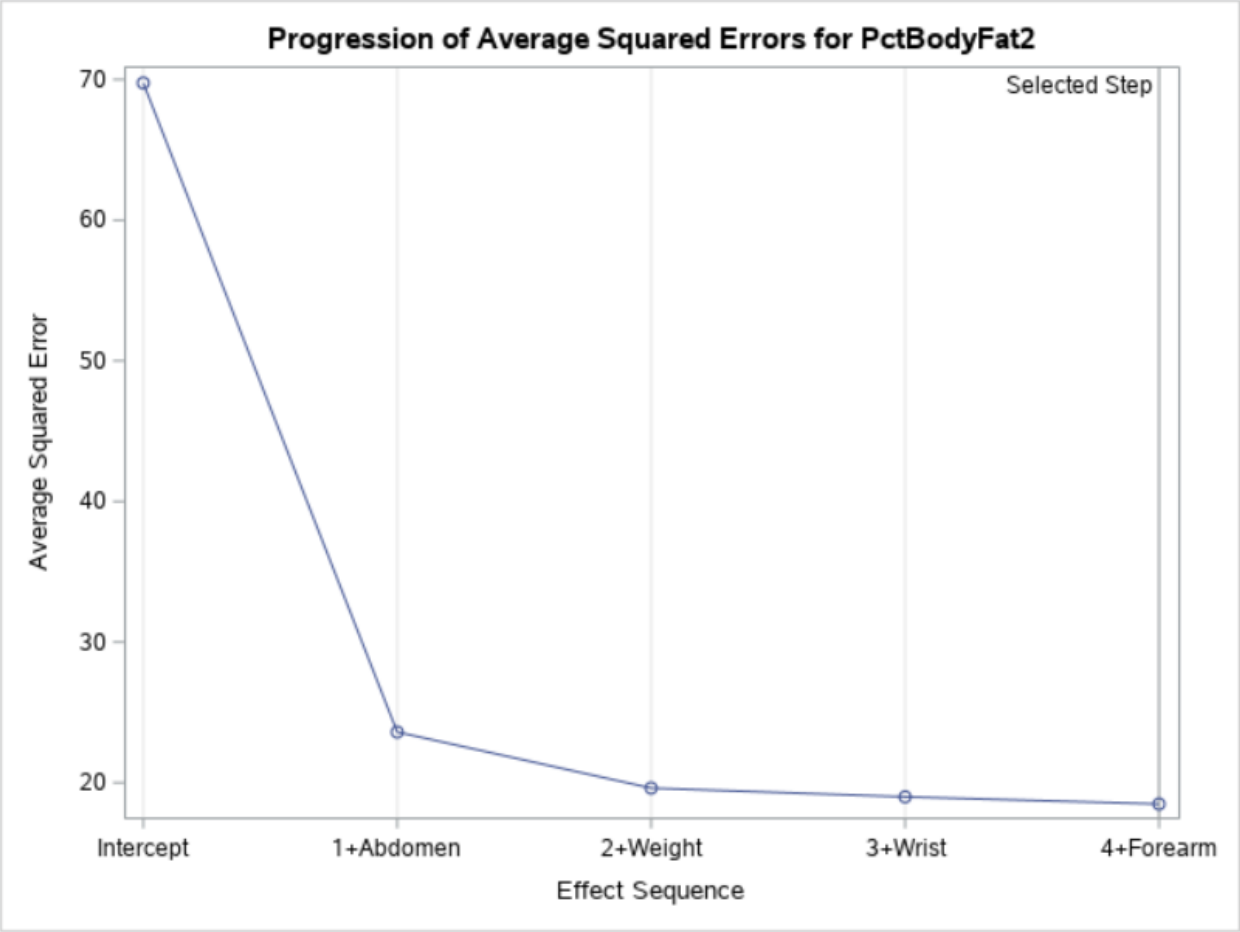
Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	SBC
0	Intercept		1	1075.2771
1	Abdomen		2	807.7042
2	Weight		3	766.6280
3	Wrist		4	764.0139
4	Forearm		5	762.7218*
* Optimal Value of Criterion				

Selection stopped at a local minimum of the SBC criterion.

Stop Details				
Candidate For	Effect	Candidate SBC		Compare SBC
Entry	Neck	765.4734	>	762.7218
Removal	Forearm	764.0139	>	762.7218


Coefficient Progression for PctBodyFat2
Standardized Coefficient
1.0
0.5
0.0
-0.5
Abdomen
Forearm
Wrist
Weight
SBC
1100
1000
900
800
Selected Step
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
Effect Sequence


Fit Criteria for PctBodyFat2
AIC
AICC
SBC
Adj R-Sq
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
Effect Sequence
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
Effect Sequence
Best Criterion Value
Selected Step


Progression of Average Squared Errors for PctBodyFat2
Selected Step
Average Squared Error
70
60
50
40
30
20
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
Effect Sequence

SBC STEPWISE Selection with PctBodyFat2

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 4).

Effects:	Intercept Weight Abdomen Forearm Wrist

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	4	12921	3230.18852	171.28
Error	247	4658.23577	18.85925	
Corrected Total	251	17579		

Root MSE	4.34272
Dependent Mean	19.15079
R-Square	0.7350
Adj R-Sq	0.7307
AIC	999.07467
AICC	999.41753
SBC	762.72182

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-34.854074	7.245005	-4.81
Weight	1	-0.135631	0.024748	-5.48
Abdomen	1	0.995751	0.056066	17.76
Forearm	1	0.472928	0.181661	2.60

```
/*st104s02.sas*/ /*Part B*/
proc glmselect data=STAT1.bodyfat2 plots=all;
        STEPWISEAIC: model PctBodyFat2 = Age Weight Height Neck Chest Abdomen
                Hip Thigh Knee Ankle Biceps Forearm Wrist
                / SELECTION=STEPWISE SELECT=AIC;
        title 'AIC STEPWISE Selection with PctBodyFat2';
run;
```

AIC STEPWISE Selection with PctBodyFat2

The GLMSELECT Procedure

Data Set	STAT1.BODYFAT2
Dependent Variable	PctBodyFat2
Selection Method	Stepwise
Select Criterion	AIC
Stop Criterion	AIC
Effect Hierarchy Enforced	None

Number of Observations Read	252
Number of Observations Used	252

Dimensions	
Number of Effects	14
Number of Parameters	14

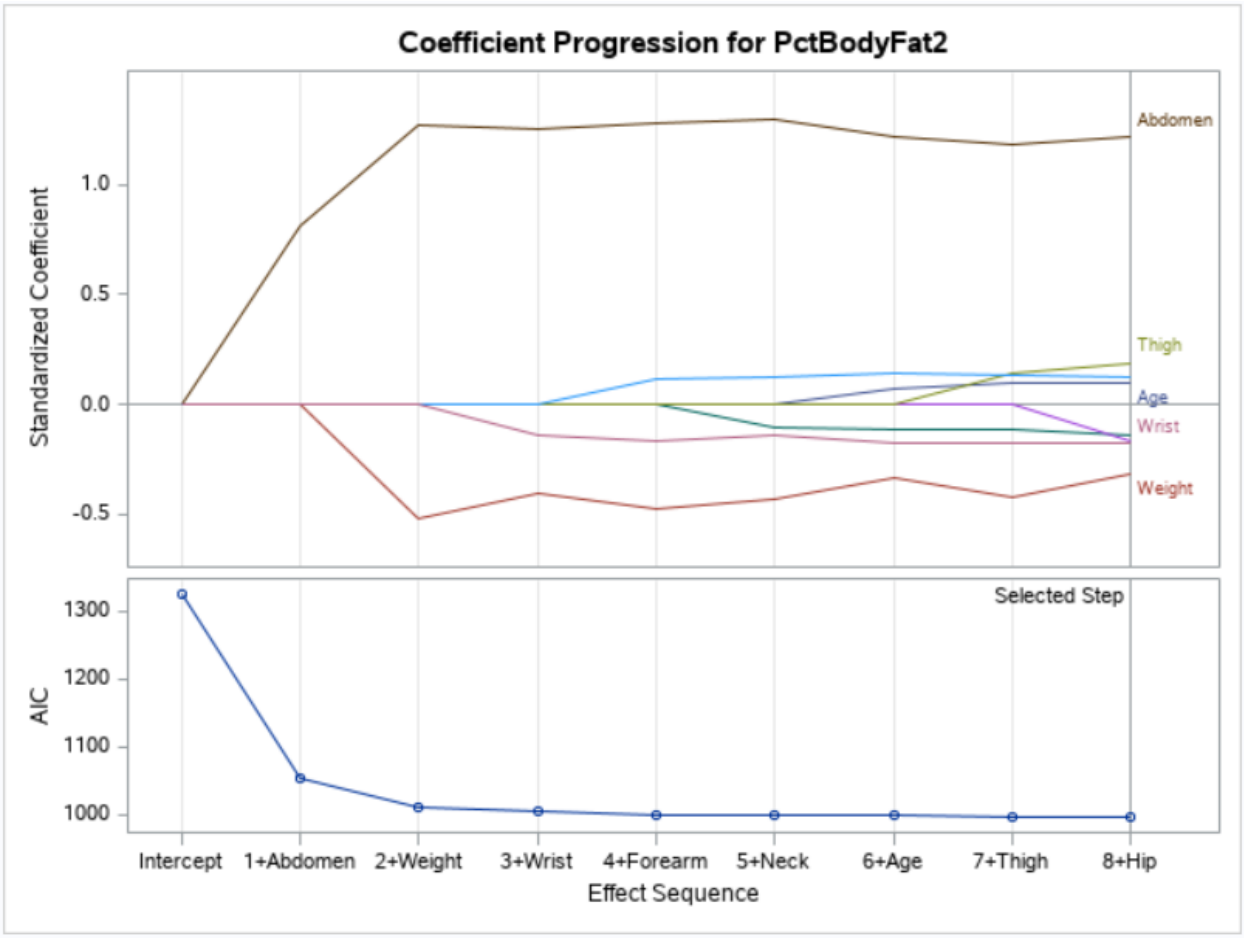
AIC STEPWISE Selection with PctBodyFat2

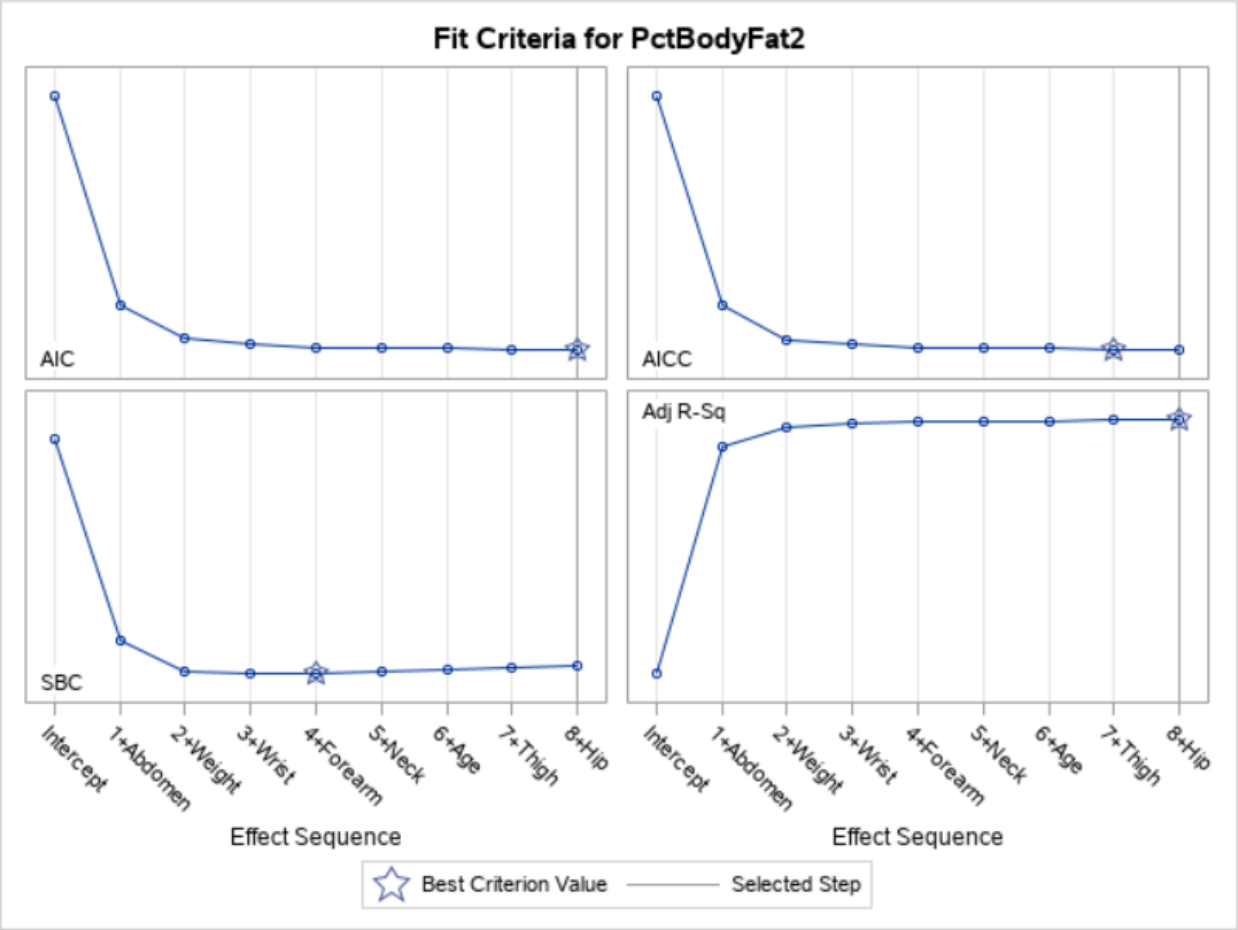
The GLMSELECT Procedure

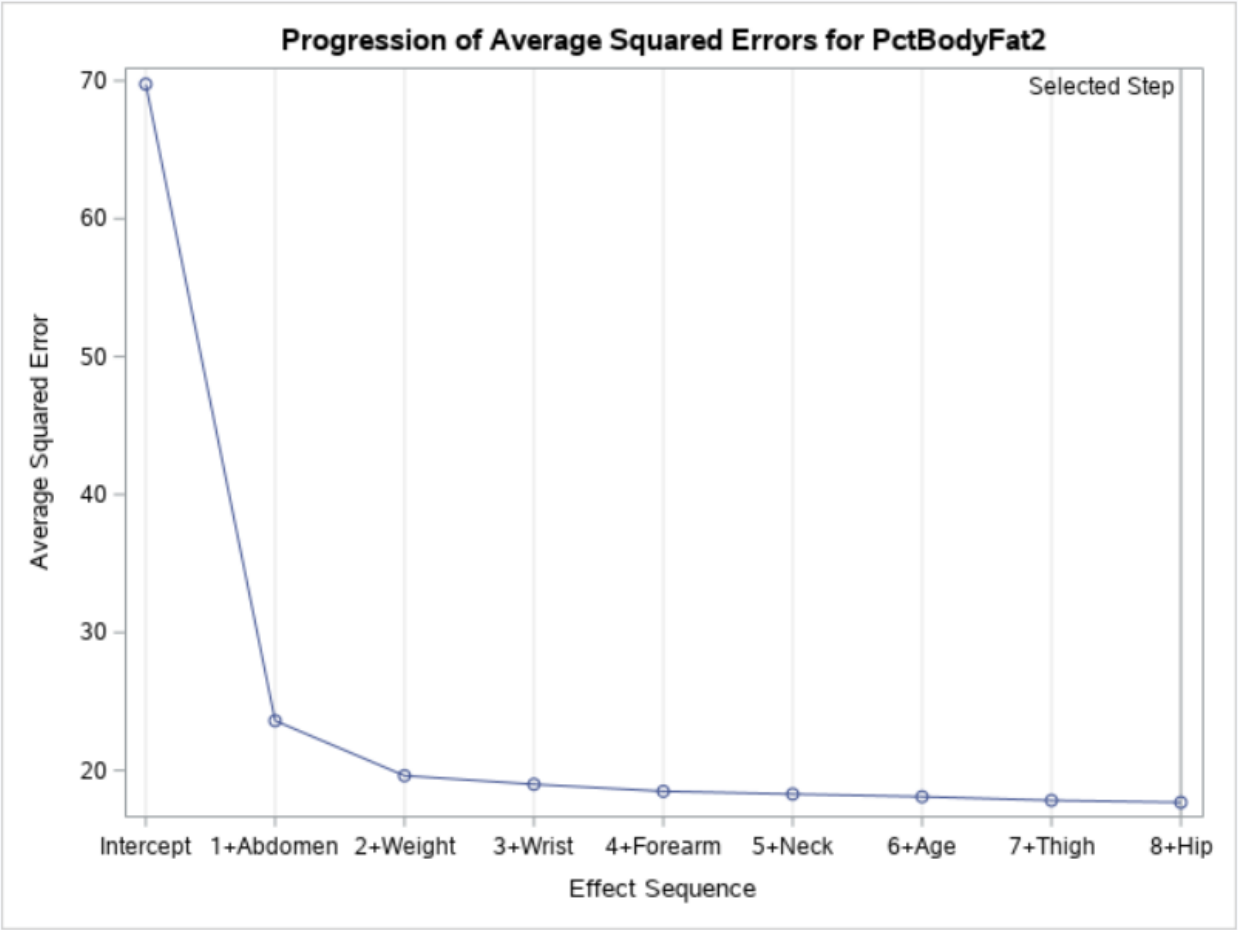
Stepwise Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	AIC
0	Intercept		1	1325.7477
1	Abdomen		2	1054.6453
2	Weight		3	1010.0398
3	Wrist		4	1003.8962
4	Forearm		5	999.0747
5	Neck		6	998.2968
6	Age		7	997.6612
7	Thigh		8	995.9088
8	Hip		9	995.8514*
* Optimal Value of Criterion				

Selection stopped at a local minimum of the AIC criterion.

Stop Details				
Candidate For	Effect	Candidate AIC		Compare AIC
Entry	Biceps	996.6772	>	995.8514
Removal	Hip	995.9088	>	995.8514


Coefficient Progression for PctBodyFat2
Standardized Coefficient
1.0
0.5
0.0
-0.5
Abdomen
Thigh
Age
Wrist
Weight
Selected Step
AIC
1300
1200
1100
1000
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
5+Neck
6+Age
7+Thigh
8+Hip
Effect Sequence


Fit Criteria for PctBodyFat2
AIC
AICC
SBC
Adj R-Sq
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
5+Neck
6+Age
7+Thigh
8+Hip
Effect Sequence
Effect Sequence
Best Criterion Value
Selected Step


Progression of Average Squared Errors for PctBodyFat2
Selected Step
Average Squared Error
70
60
50
40
30
20
Intercept
1+Abdomen
2+Weight
3+Wrist
4+Forearm
5+Neck
6+Age
7+Thigh
8+Hip
Effect Sequence

AIC STEPWISE Selection with PctBodyFat2

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 8).

Effects:	Intercept Age Weight Neck Abdomen Hip Thigh Forearm Wrist

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	8	13124	1640.45820	89.47
Error	243	4455.32427	18.33467	
Corrected Total	251	17579		

Root MSE	4.28190
Dependent Mean	19.15079
R-Square	0.7466
Adj R-Sq	0.7382
AIC	995.85136
AICC	996.76422
SBC	773.61622

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-22.656373	11.713855	-1.93
Age	1	0.065780	0.030776	2.14
Weight	1	-0.089853	0.039906	-2.25
Neck	1	-0.466558	0.224617	-2.08
Abdomen	1	0.944815	0.071934	13.13

Practice - Using PROC GLMSELECT to Perform Other Model Selection Techniques

Question 1

Use the **stat1.bodyfat2** data set to identify a set of "best" models using other model selection techniques.

1. With the SELECTION=STEPWISE option, use SELECT=SBC in PROC GLMSELECT to identify a set of candidate models that predict **PctBodyFat2** as a function of the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**.
2. Submit the code.

What do you notice about the results?
The results to show 5 variables to be used (including intercept)

Solution code:

```
/*st104s02.sas*/ /*Part A*/

ods graphics on;
proc glmselect data=STAT1.bodyfat2 plots=all;
   STEPWISESBC: model PctBodyFat2 = Age Weight Height Neck Chest Abdomen
                      Hip Thigh Knee Ankle Biceps Forearm Wrist
                      / SELECTION=STEPWISE SELECT=SBC;
   title 'SBC STEPWISE Selection with PctBodyFat2';
run;
```

In the results, notice the following:

- The stepwise selection process, using SELECT=SBC, seems to select a five-effect model (including the intercept).
- The Coefficient panel shows that the standardized coefficients do not vary greatly when additional effects are added to the model.
- The Fit panel indicates that the best model, according to AIC, AICC, adjusted R-square, and SBC, is the final model viewed during the selection process. Remember that this statement compares only the models that were viewed in these steps of the selection process.
- The parameter estimates from the selected model are presented in the Parameter Estimates table.

Question 2
Modify the code to specify **SELECT=AIC**. Submit the code and view the results. How many effects are in the selected model?
9 effects (including intercept)

Solution code:

```
/*st104s02.sas*/ /*Part B*/

proc glmselect data=stat1.bodyfat2 plots=all;
   STEPWISEAIC: model PctBodyFat2=Age Weight Height
                      Neck Chest Abdomen Hip Thigh
                      Knee Ankle Biceps Forearm Wrist
                      / SELECTION=STEPWISE SELECT=AIC;
   title 'AIC STEPWISE Selection with PctBodyFat2';
run;
quit;
title;
```

Using SELECT=AIC, the selected model contains nine effects (including the intercept).