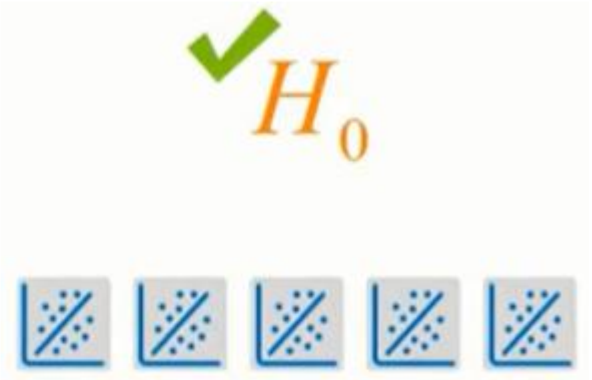
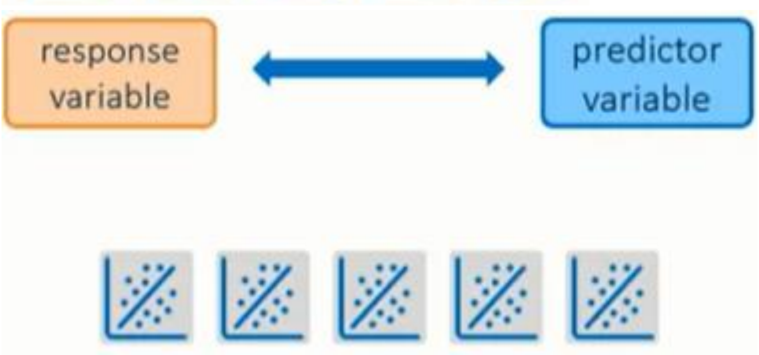
SBA Statistical Business Analyst with SAS

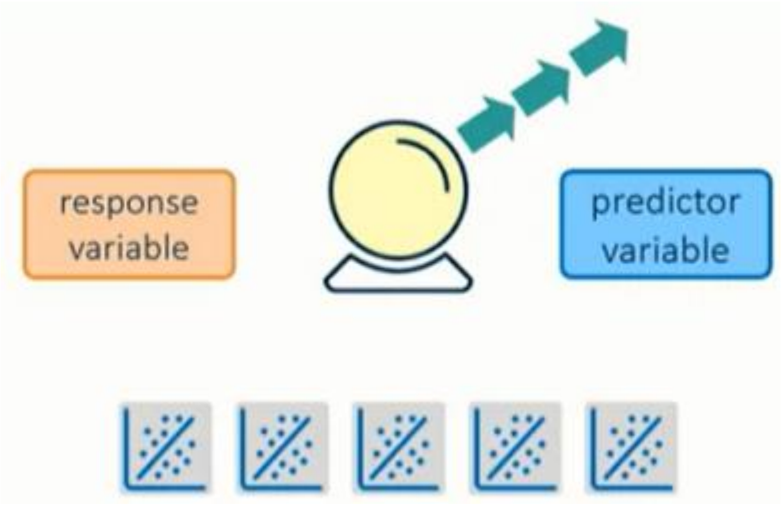
SBA2 Regression Modeling Fundamentals

SBA202B Predictive Modeling and Scoring Predictive Models

Overview







SALE
historical customer data

SALE
response to promotion

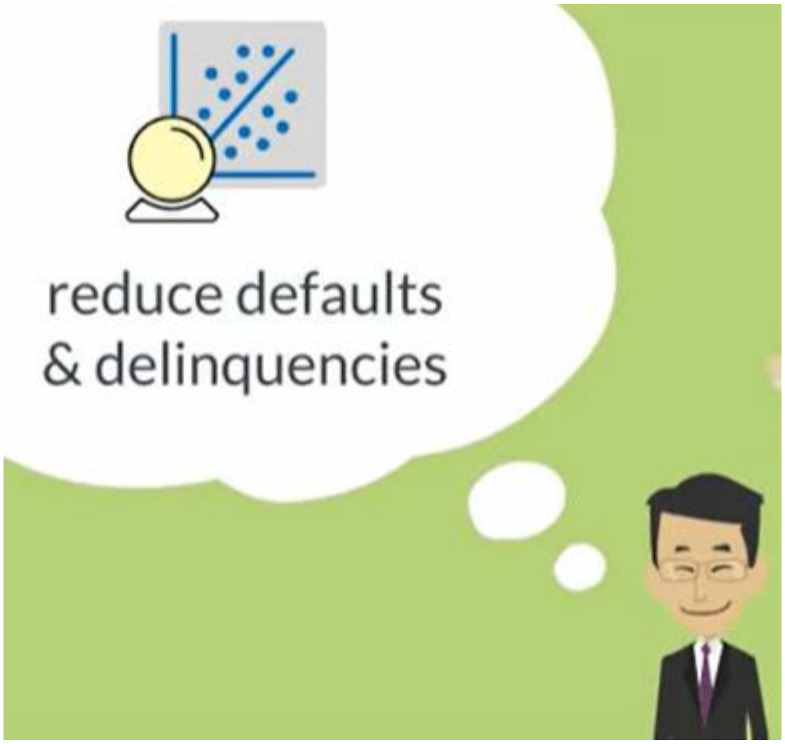
CALENDAR
S M T W T F S
1 2 3 4
5 6 7 8 9 10 11
12 13 14 15 16 17 18
19 20 21 22 23 24 25
26 27 28 29 30

Loyalty!

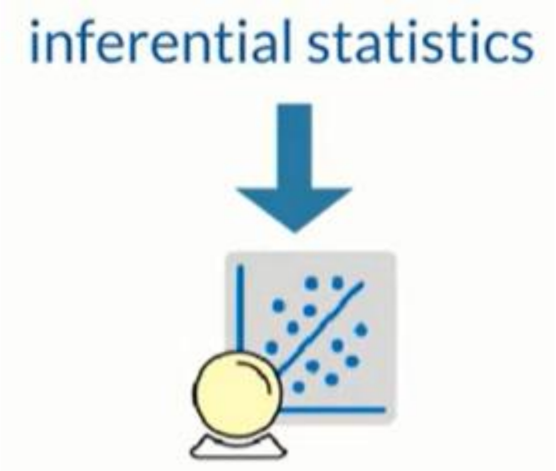
CREDIT CARD
2739 4909 8964 1234
2739 4909 8964 1234

income
employment status
credit rating

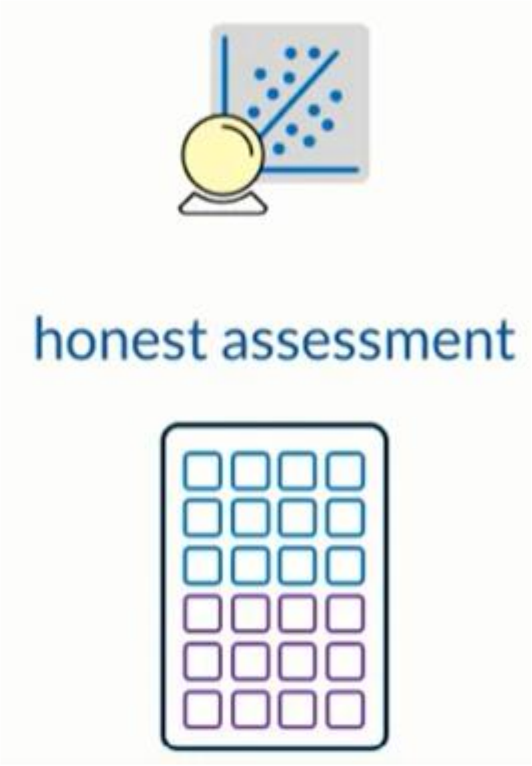
debt status

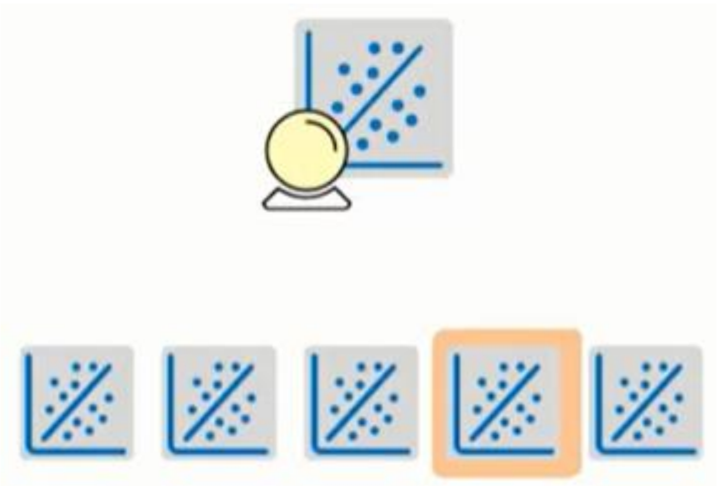


reduce defaults & delinquencies



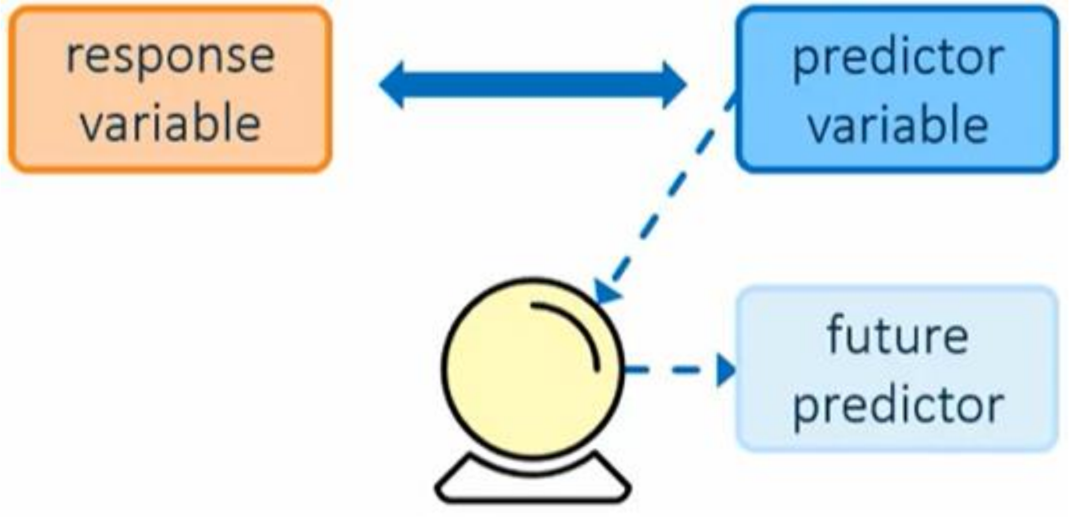
inferential statistics

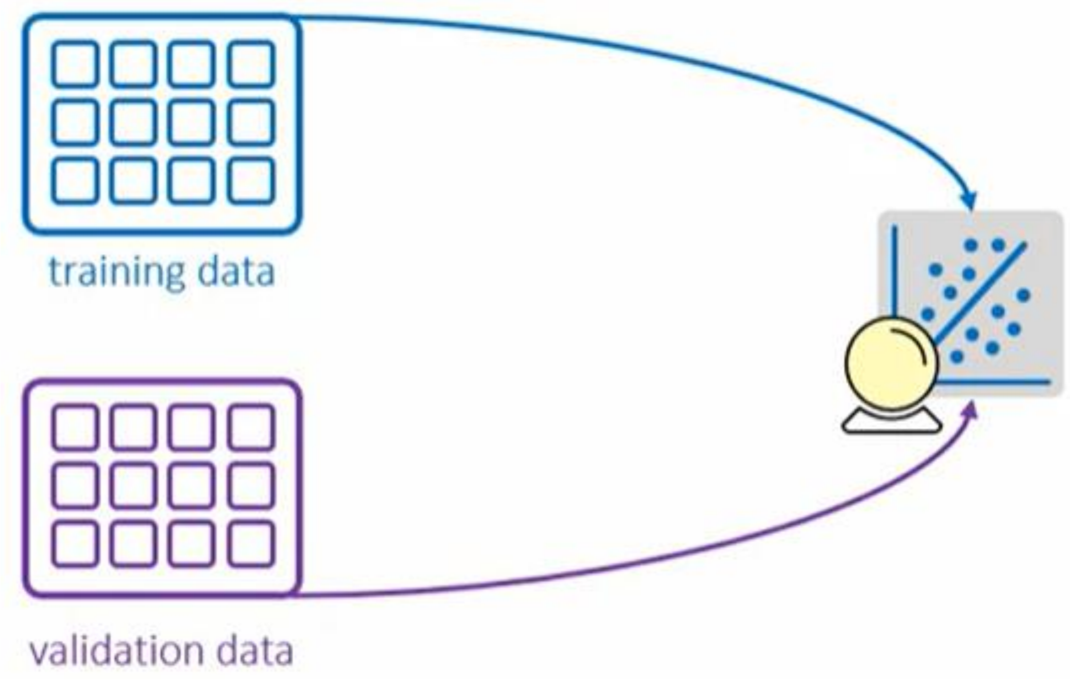

honest assessment

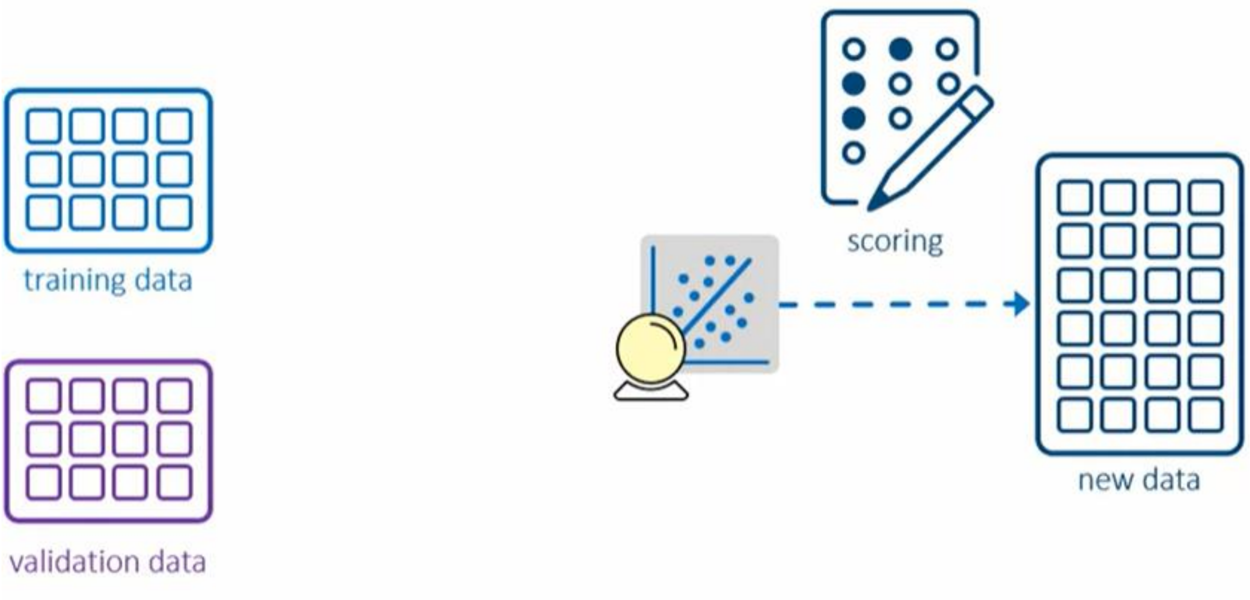


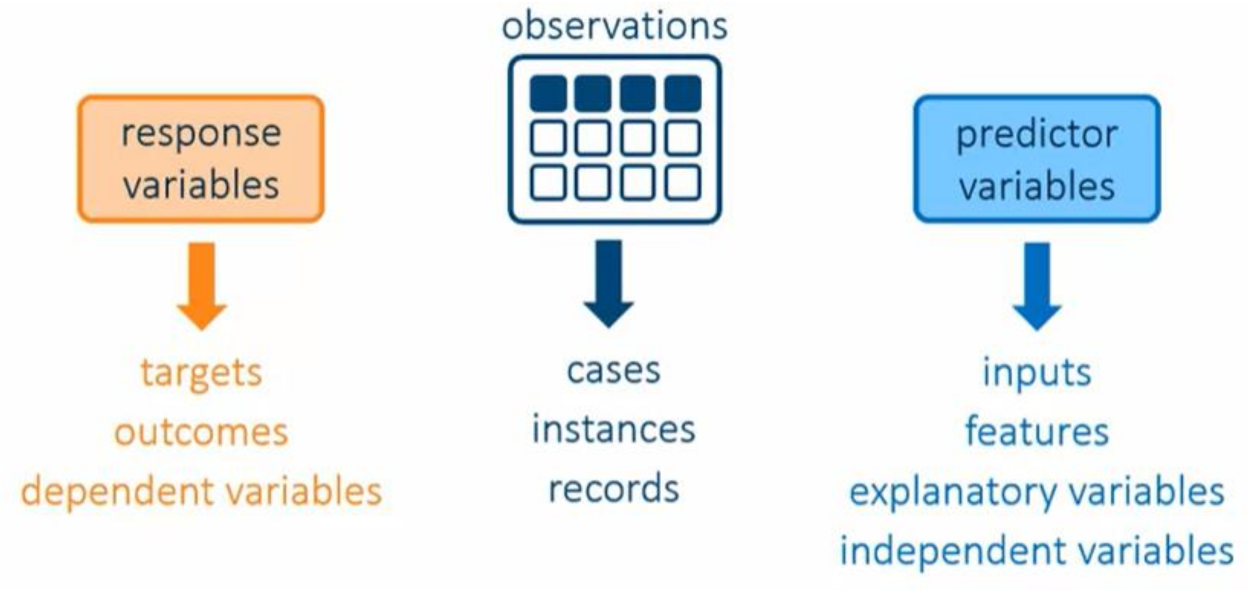
Brief Introduction to Predictive Modeling Scenario

Predictive Modeling Terminology









observations

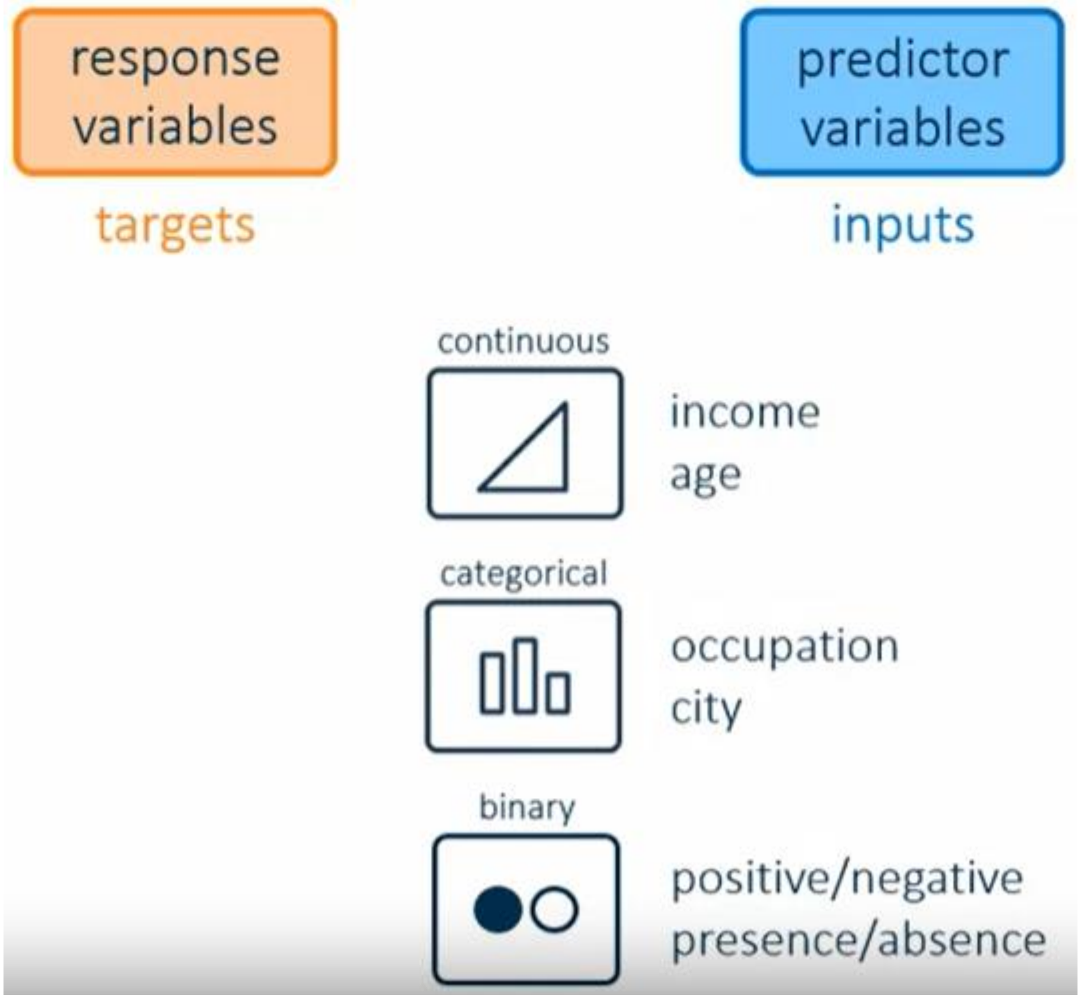
response variables

targets
outcomes
dependent variables

cases
instances
records

predictor variables

inputs
features
explanatory variables
independent variables



response variables

targets

predictor variables

inputs

continuous

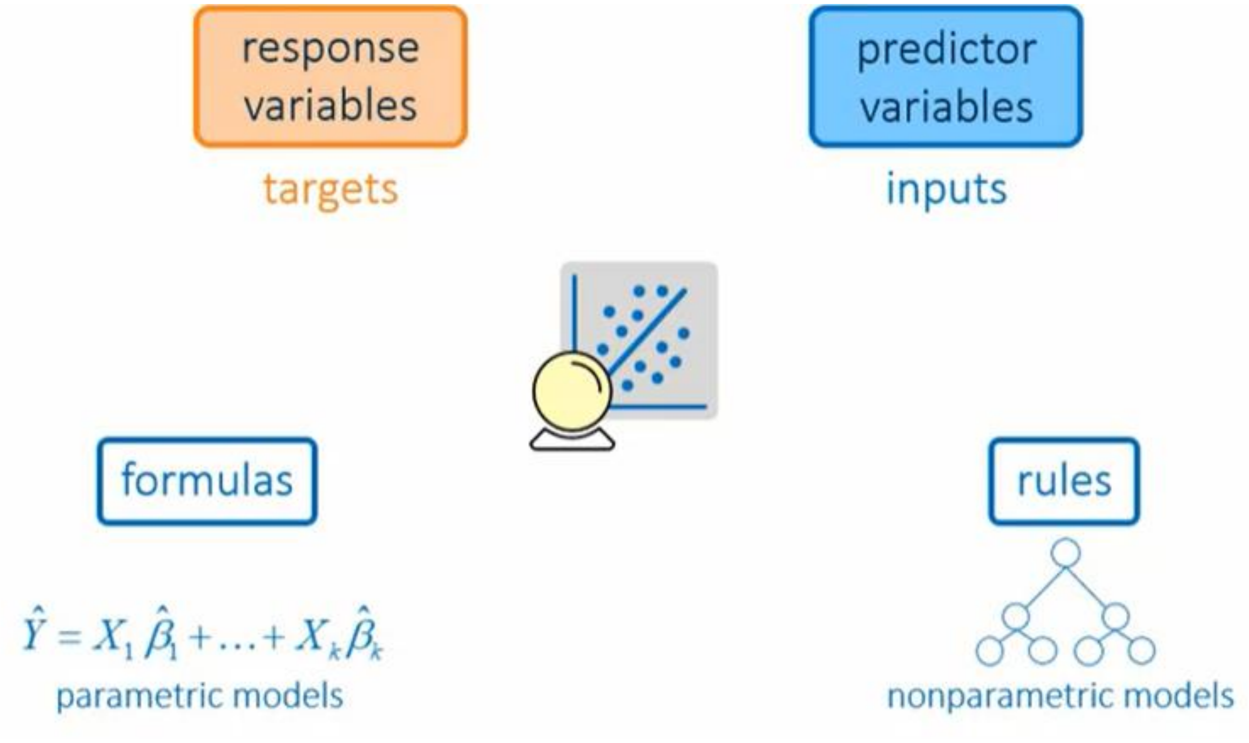
income
age

categorical

occupation
city

binary

positive/negative
presence/absence



response variables

targets

predictor variables

inputs

formulas

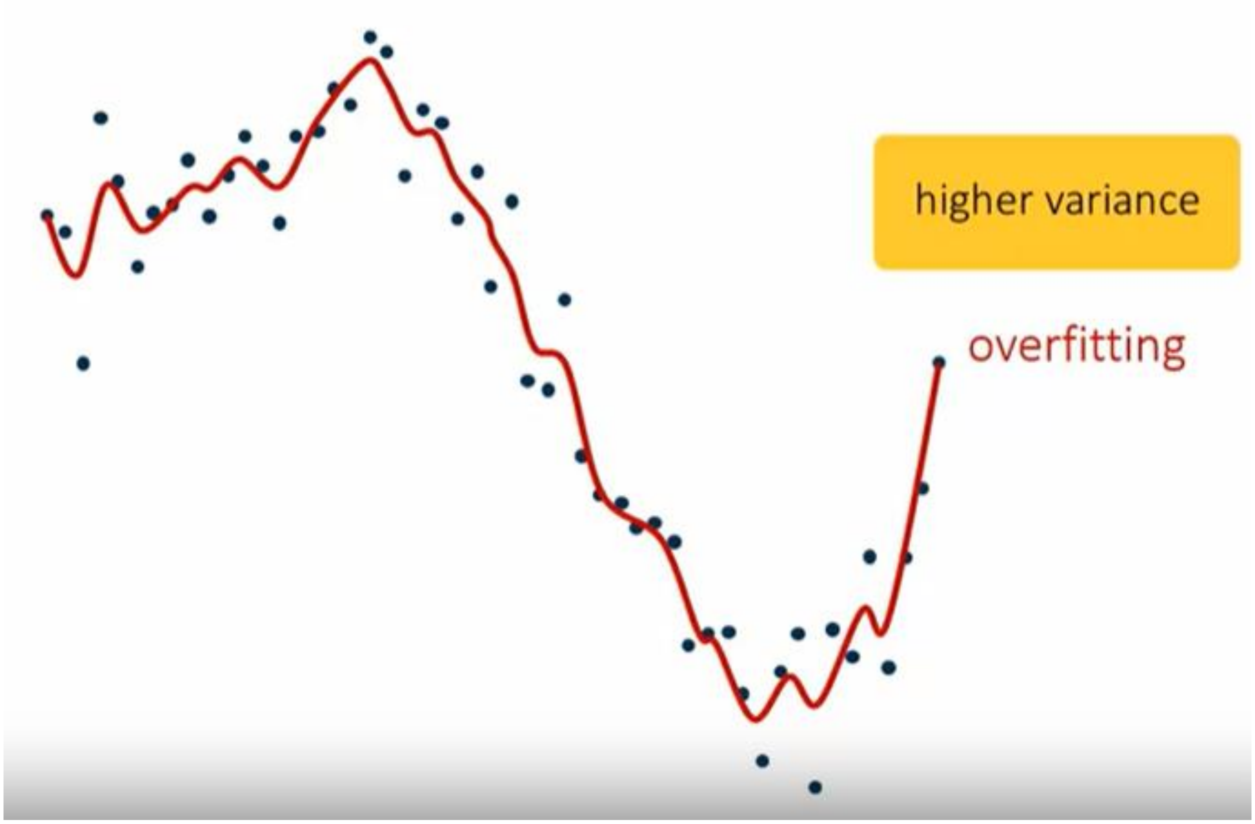
$\hat{Y} = X_1\hat{\beta}_1 + \ldots + X_k\hat{\beta}_k$

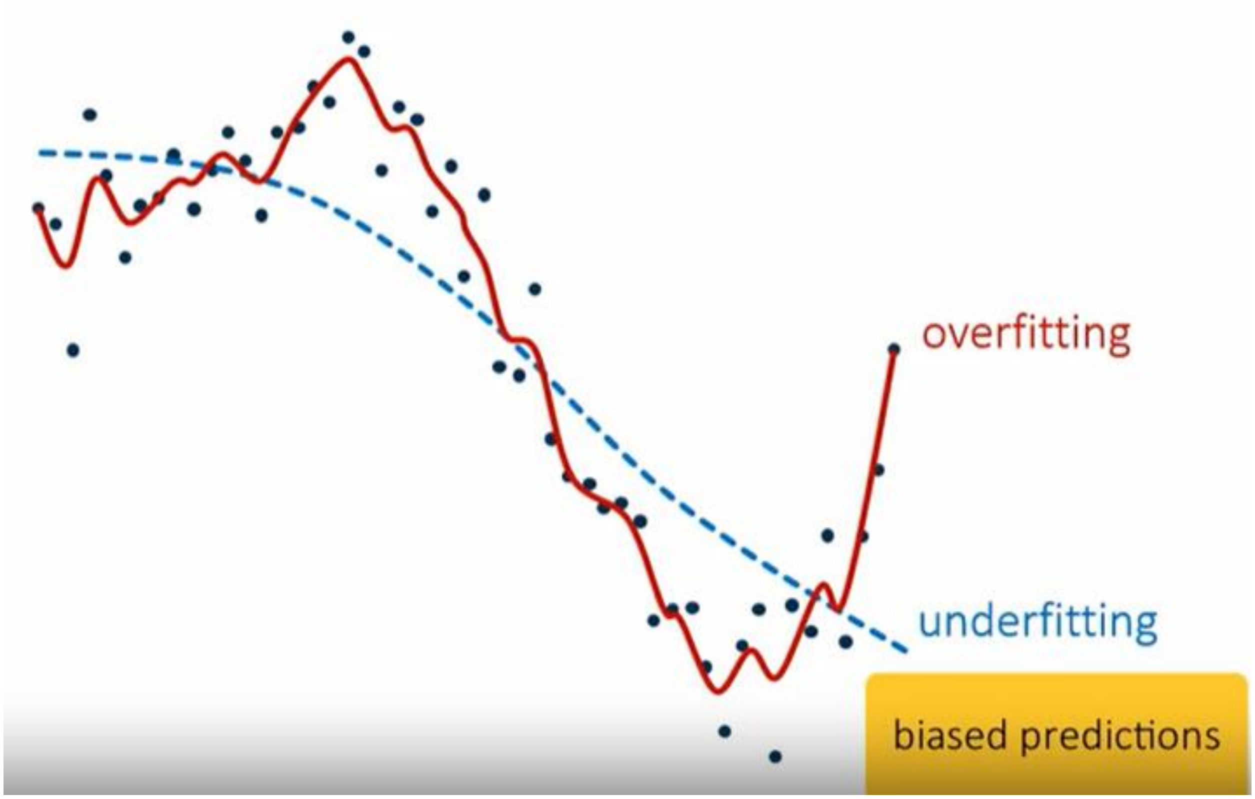
parametric models

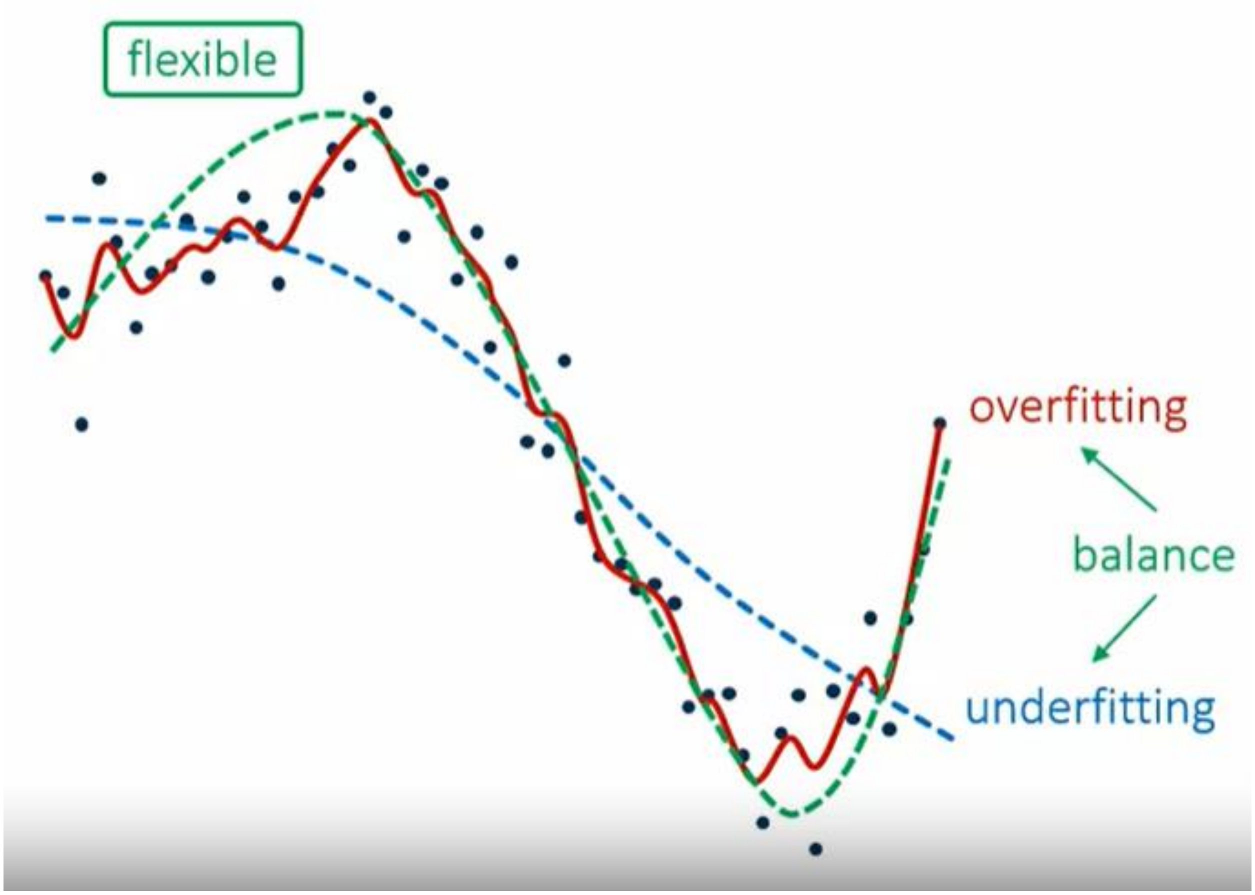
rules

nonparametric models

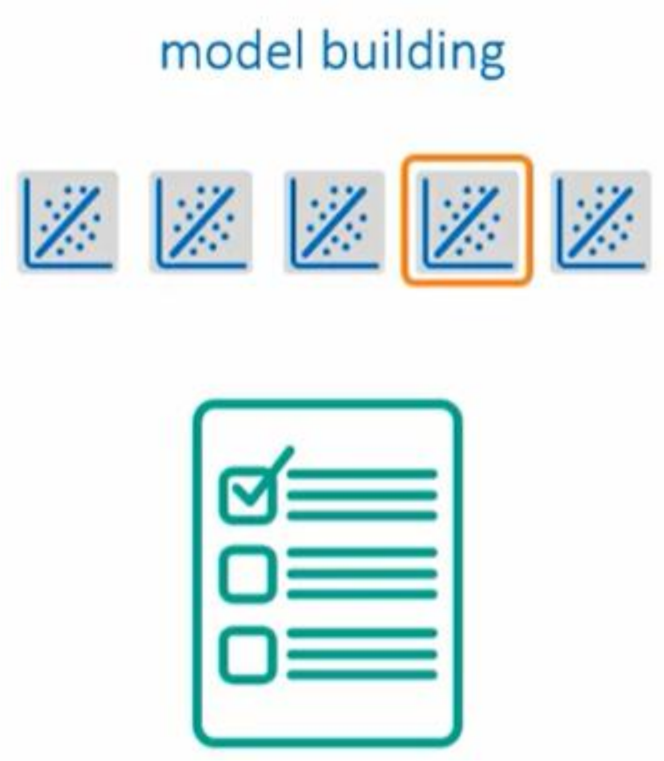
Model Complexity

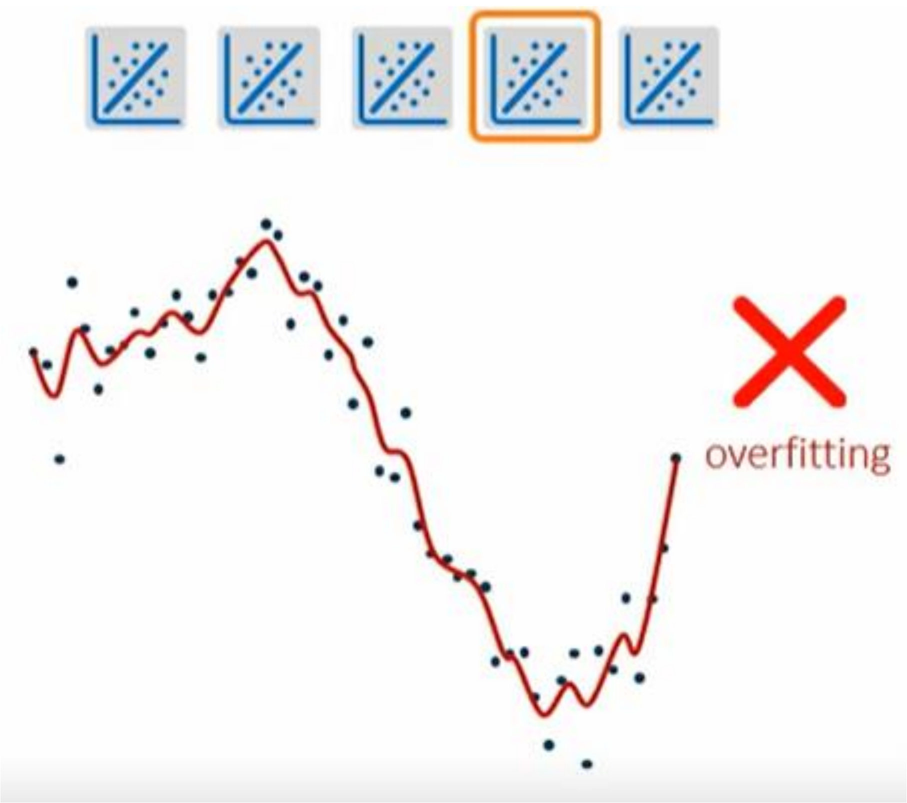

higher variance
overfitting

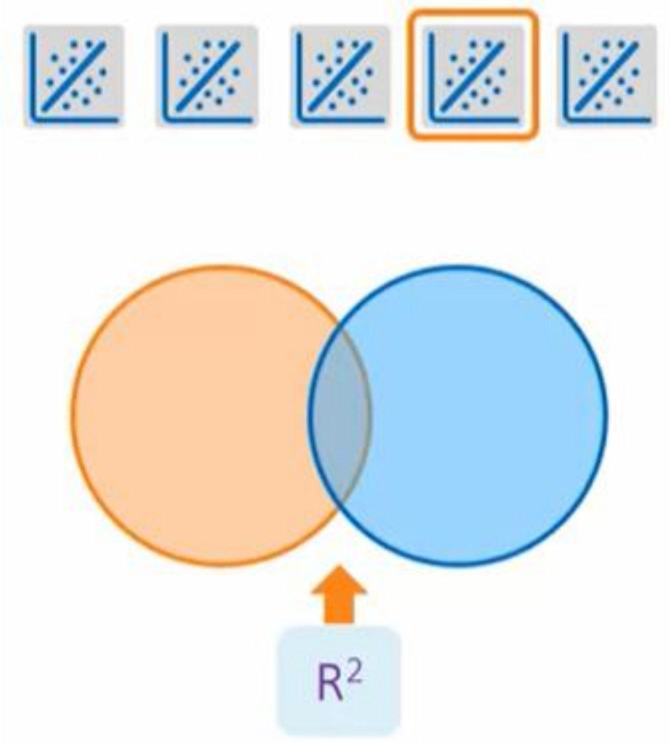

overfitting
underfitting
biased predictions

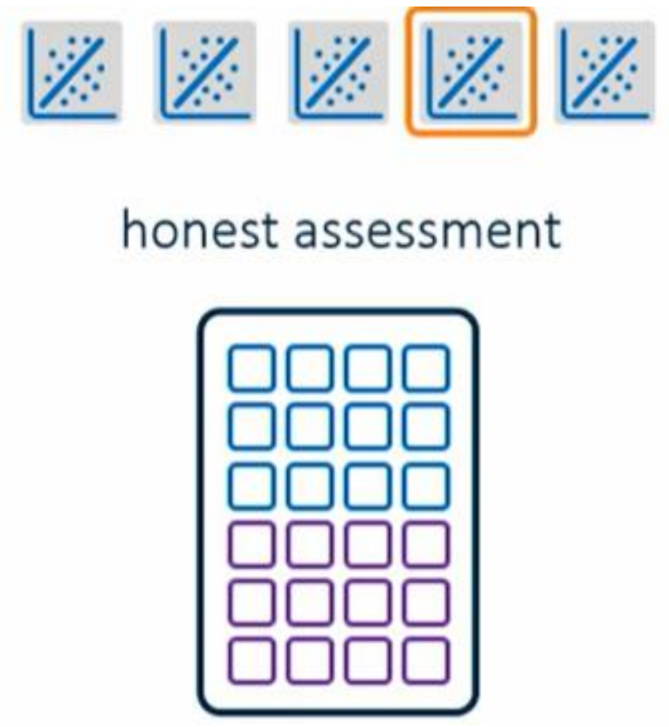

flexible
overfitting
balance
underfitting

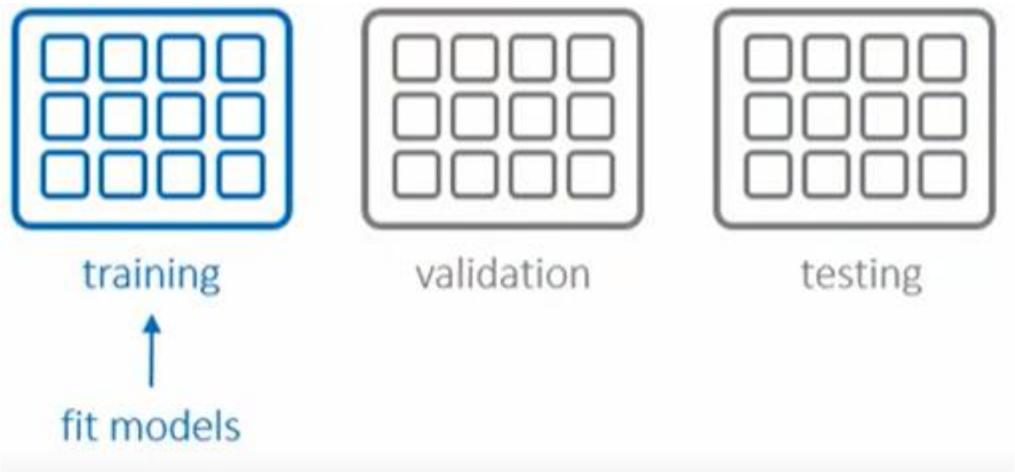
Building a Predictive Model

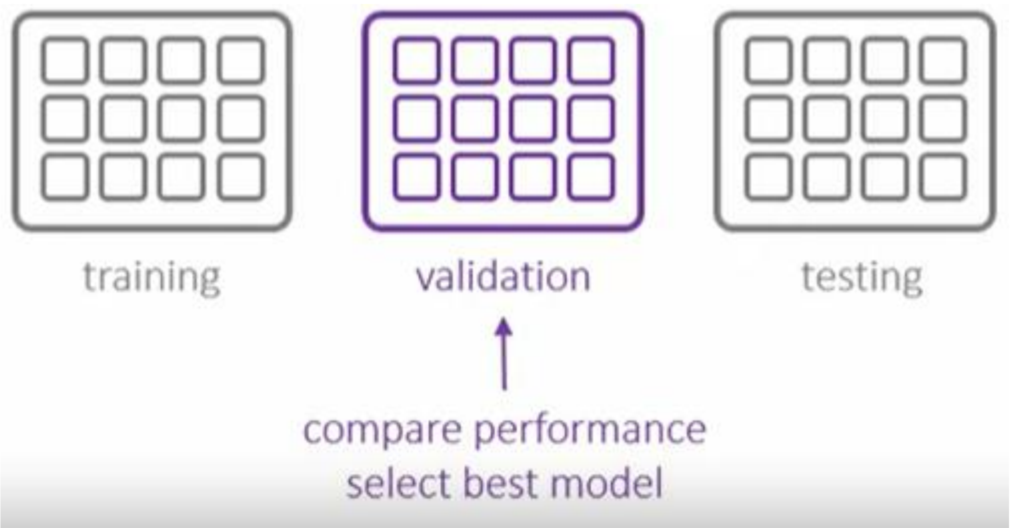


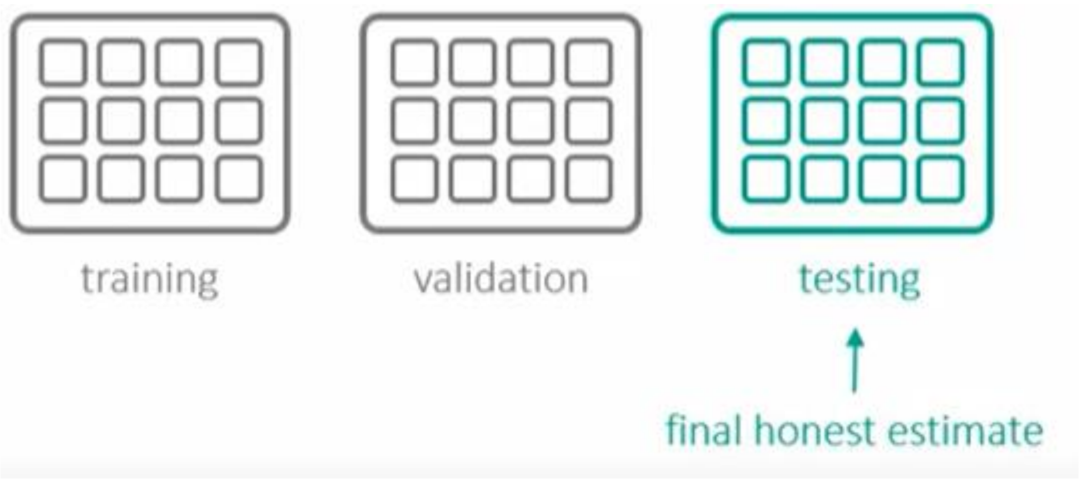


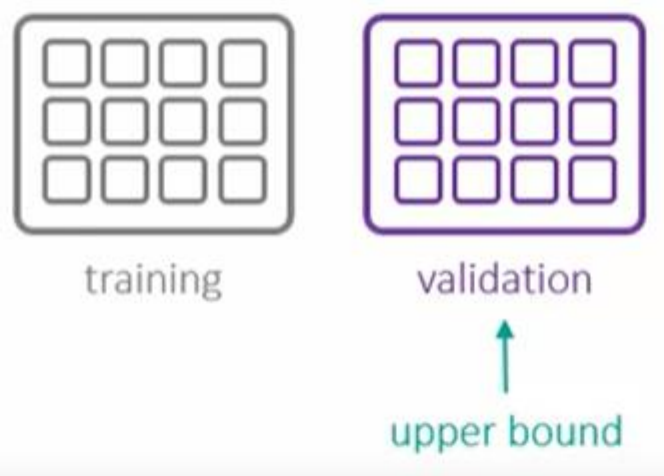

R2


honest assessment

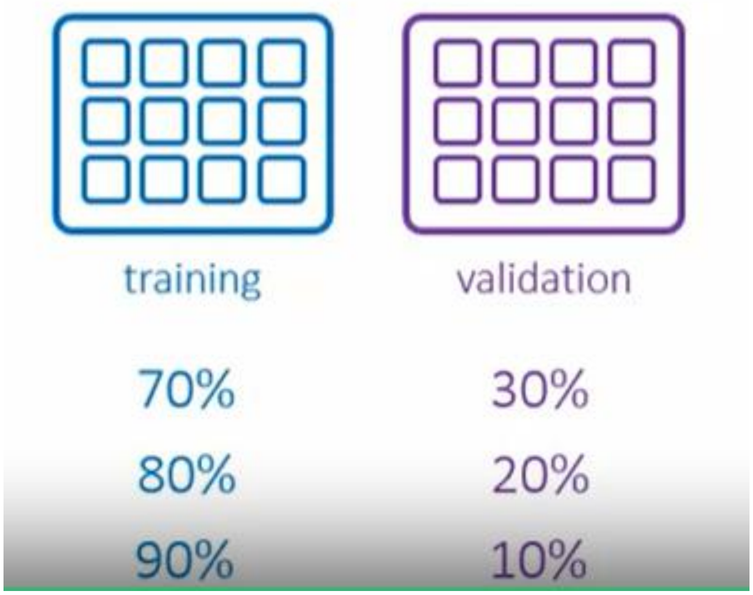

training
validation
testing
fit models


training
validation
testing
compare performance
select best model

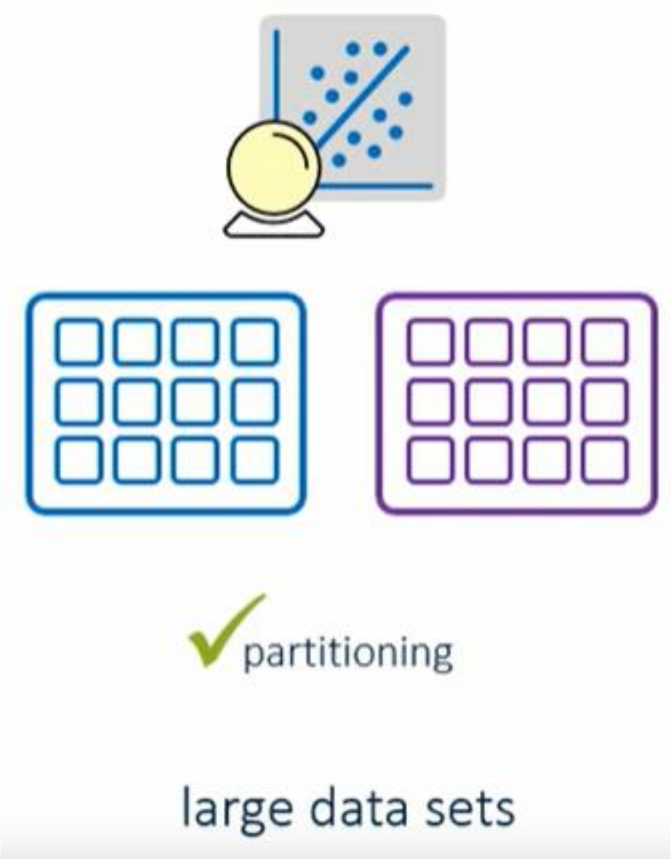

training
validation
testing
final honest estimate



no globally optimal percentage



training	validation
70%	30%
80%	20%
90%	10%



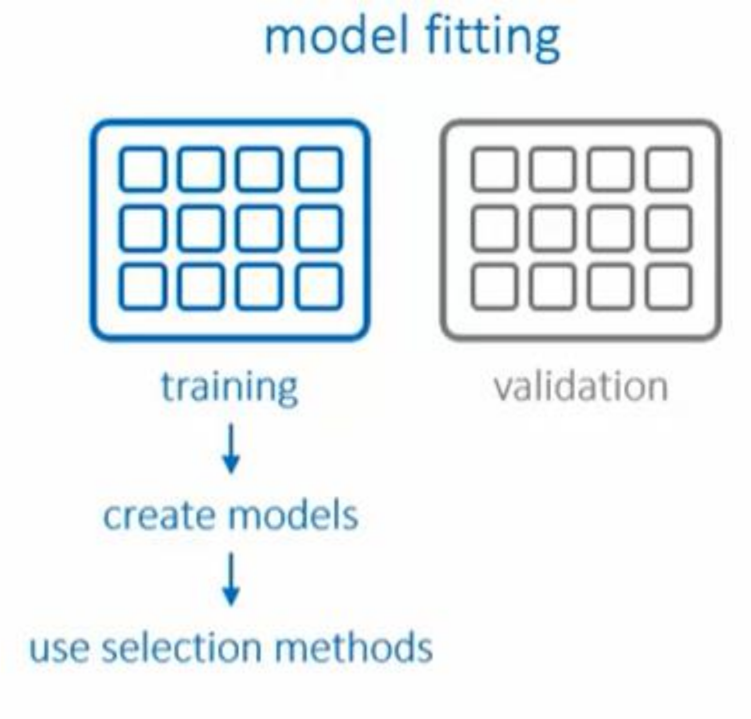
✓ partitioning

large data sets

⊘ partitioning

small or medium data sets

Model Assessment and Selection



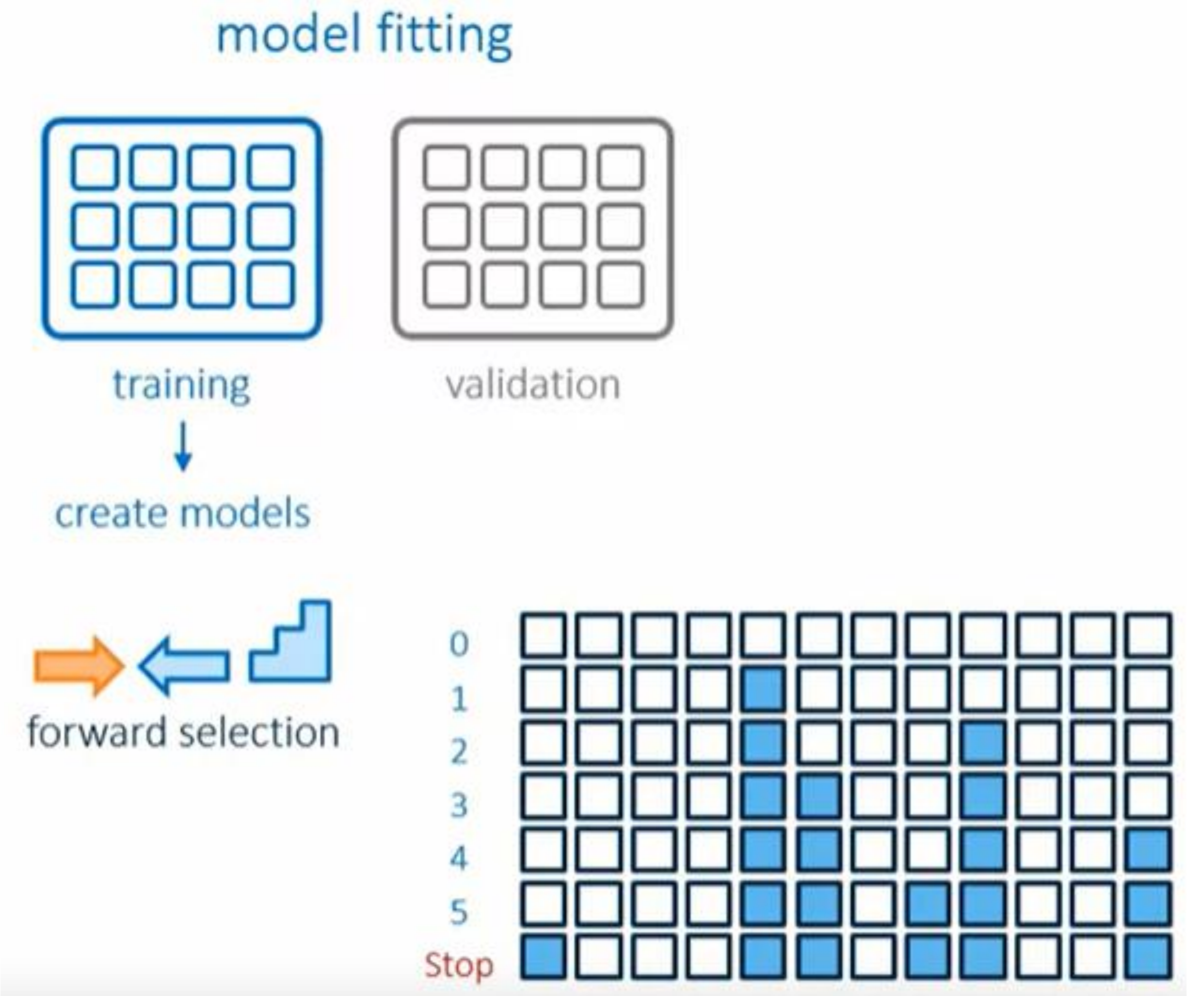
model fitting

training

validation

create models

use selection methods



model fitting

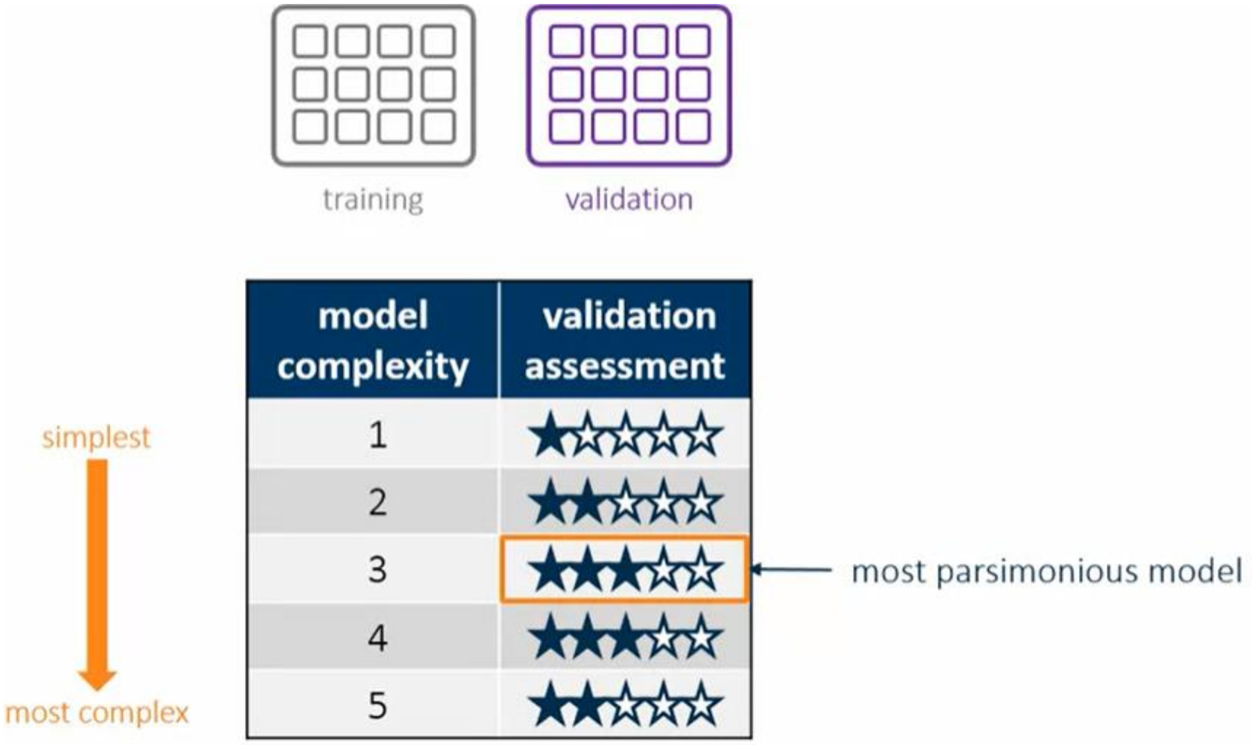
training

validation

create models

forward selection

0	□	□	□	□	□	□	□	□	□	□	□	□
1	□	□	□	□	■	□	□	□	□	□	□	□
2	□	□	□	□	■	□	□	□	■	□	□	□
3	□	□	□	□	■	■	□	□	■	□	□	□
4	□	□	□	□	■	■	□	□	■	□	□	■
5	□	□	□	□	■	■	□	■	■	□	□	■
Stop	■	□	□	□	■	■	□	■	■	□	□	■

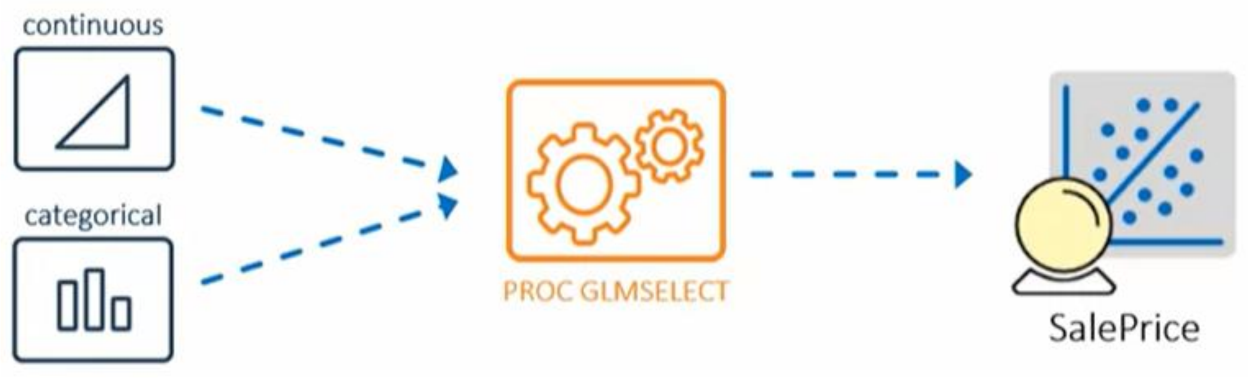


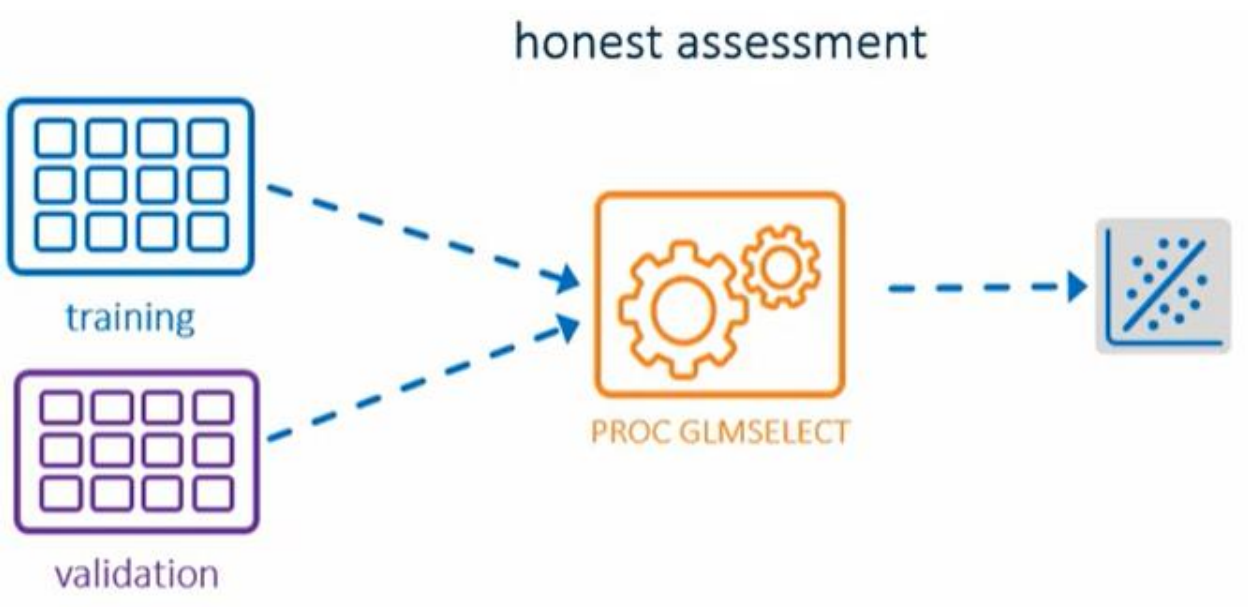
model complexity	validation assessment
1	★☆☆☆☆
2	★★☆☆☆
3	★★★☆☆
4	★★★☆☆
5	★★☆☆☆

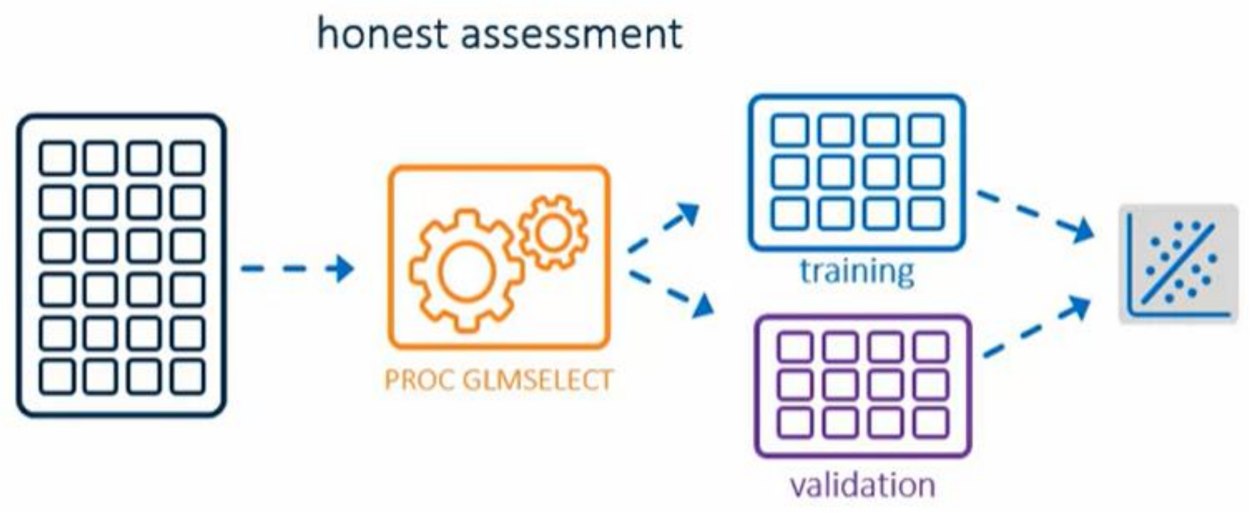
When you use honest assessment, which of the following would be considered the best model?

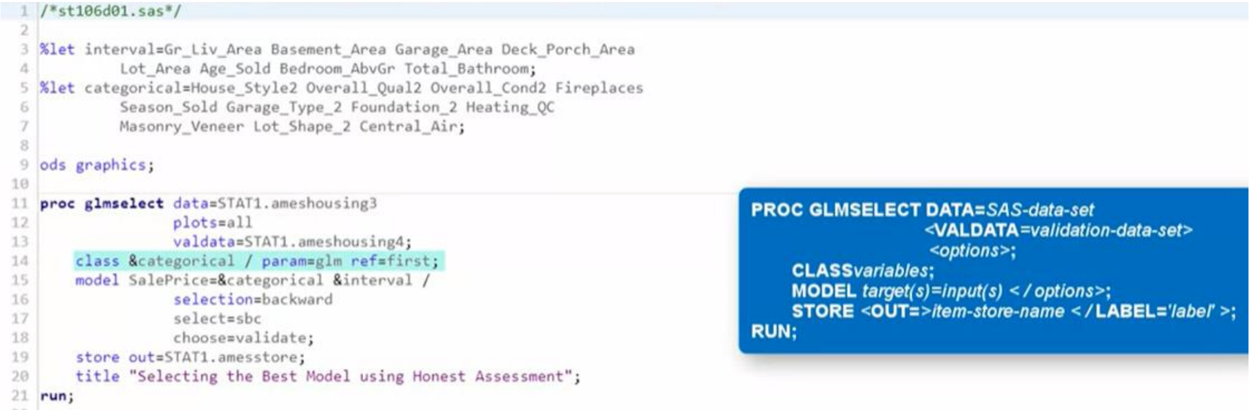
The best model is the simplest (the most parsimonious) model that has the best performance on the validation data. The training data is used to fit the model and generate the possible models to be assessed.

Demo Building a Predictive Model Using PROC GLMSELECT



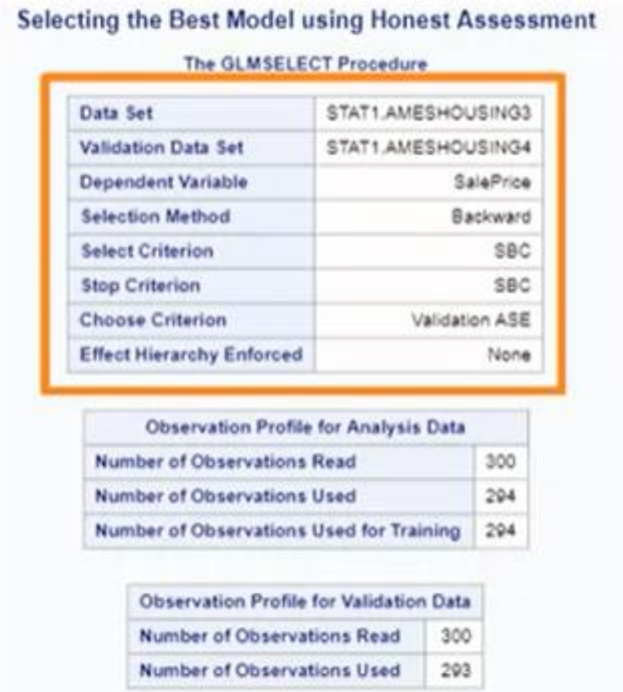






```
/*st106d01.sas*/

%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;
%let categorical=House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces
         Season_Sold Garage_Type_2 Foundation_2 Heating_QC
         Masonry_Veneer Lot_Shape_2 Central_Air;

ods graphics;

proc glmselect data=STAT1.ameshousing3
               plots=all
               valdata=STAT1.ameshousing4;
    class &categorical / param=glm ref=first;
    model SalePrice=&categorical &interval /
               selection=backward
               select=sbc
               choose=validate;
    store out=STAT1.amesstore;
    title "Selecting the Best Model using Honest Assessment";
run;
```

```
PROC GLMSELECT DATA=SAS-data-set
                    <VALDATA=validation-data-set>
                    <options>;
    CLASSvariables;
    MODEL target(s)=input(s) < / options>;
    STORE <OUT=>item-store-name < / LABEL='label' >;
RUN;
```



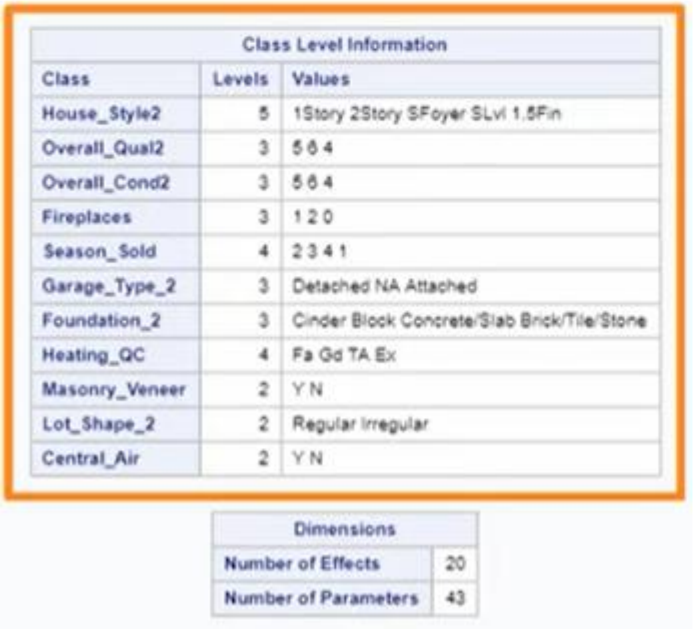
Selecting the Best Model using Honest Assessment

The GLMSELECT Procedure

Data Set	STAT1.AMESHOUSING3
Validation Data Set	STAT1.AMESHOUSING4
Dependent Variable	SalePrice
Selection Method	Backward
Select Criterion	SBC
Stop Criterion	SBC
Choose Criterion	Validation ASE
Effect Hierarchy Enforced	None

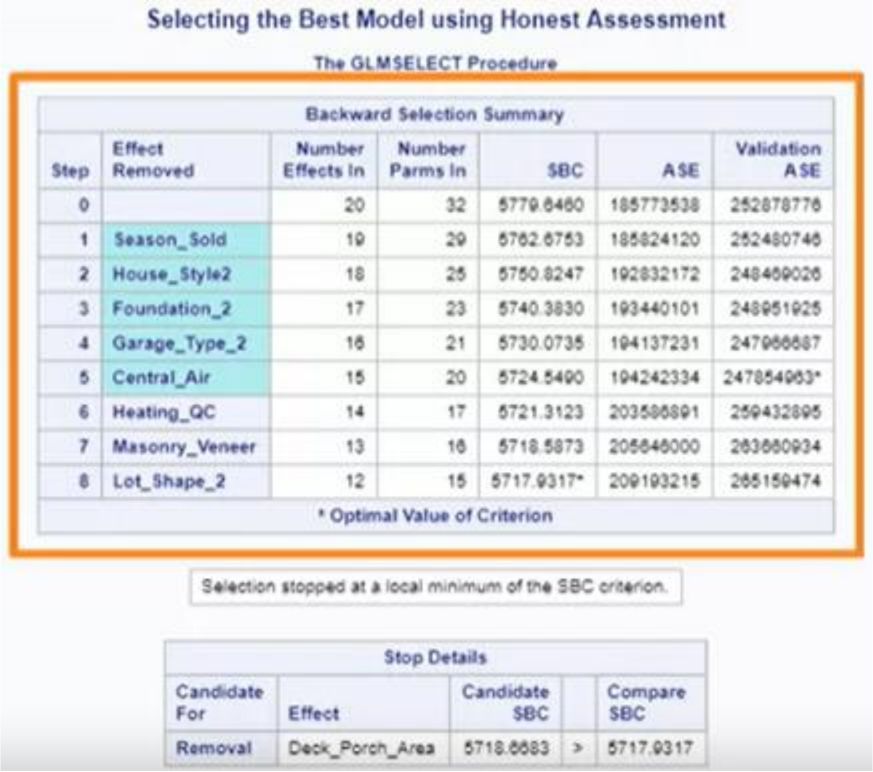
Observation Profile for Analysis Data	
Number of Observations Read	300
Number of Observations Used	294
Number of Observations Used for Training	294

Observation Profile for Validation Data	
Number of Observations Read	300
Number of Observations Used	293



Class Level Information		
Class	Levels	Values
House_Style2	5	1Story 2Story SFoyer SLvl 1.5Fin
Overall_Qual2	3	5 6 4
Overall_Cond2	3	5 6 4
Fireplaces	3	1 2 0
Season_Sold	4	2 3 4 1
Garage_Type_2	3	Detached NA Attached
Foundation_2	3	Cinder Block Concrete/Slab Brick/Tile/Stone
Heating_QC	4	Fa Gd TA Ex
Masonry_Veneer	2	Y N
Lot_Shape_2	2	Regular Irregular
Central_Air	2	Y N

Dimensions	
Number of Effects	20
Number of Parameters	43



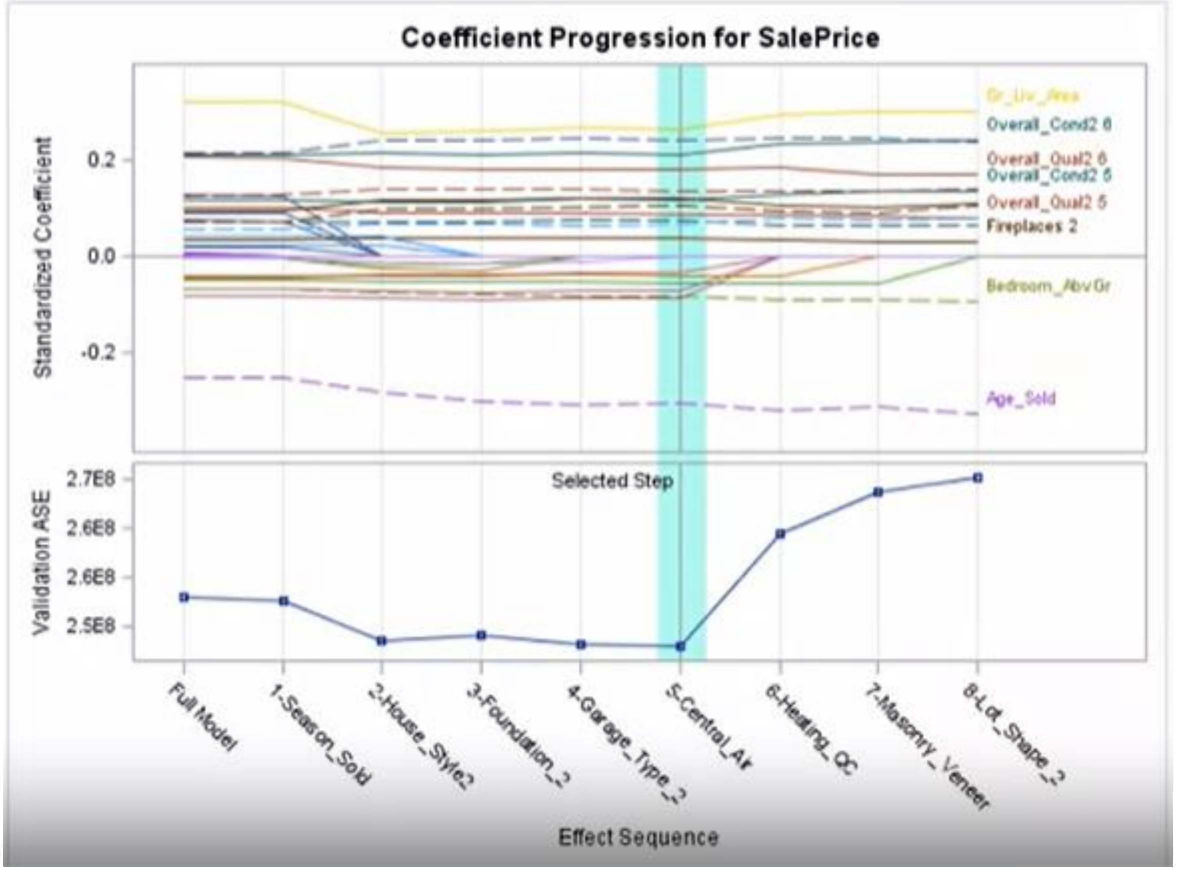
Selecting the Best Model using Honest Assessment

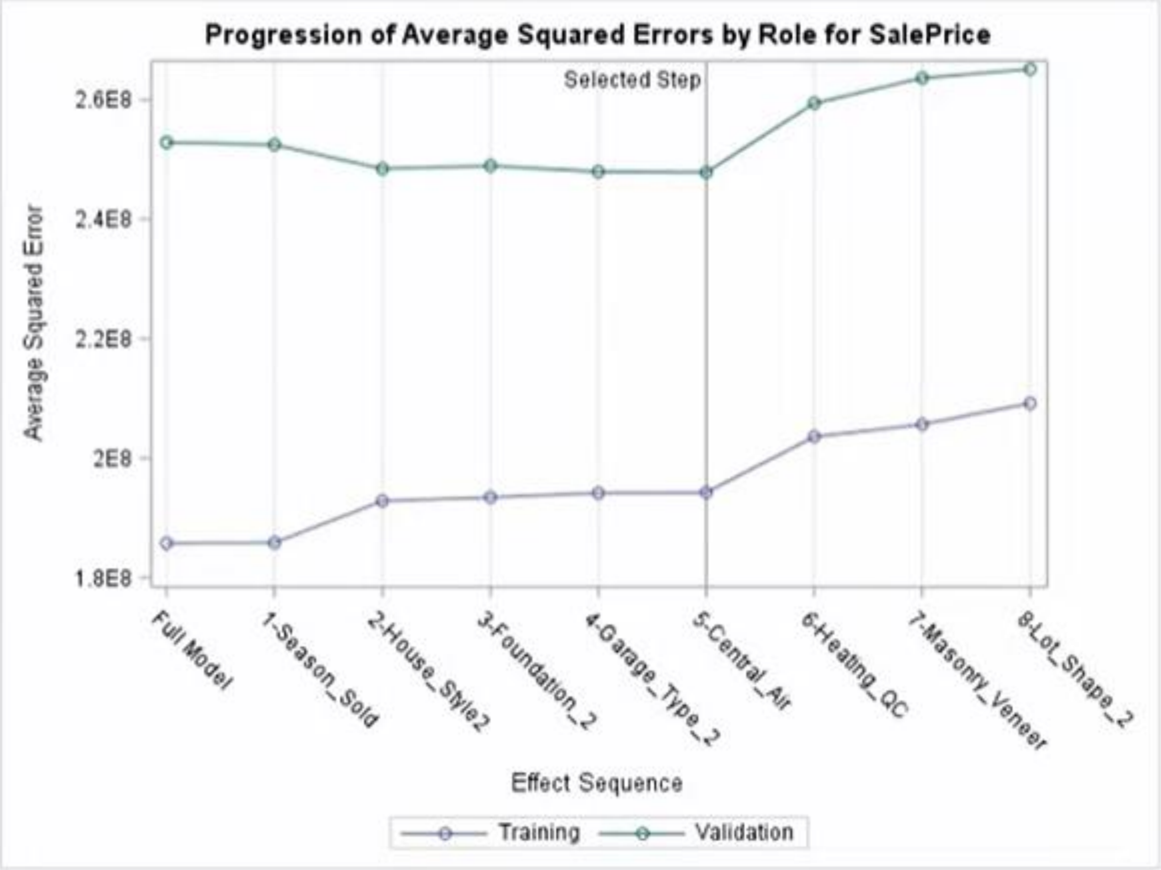
The GLMSELECT Procedure

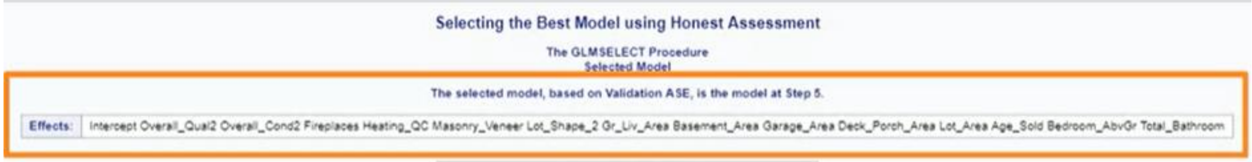
Backward Selection Summary						
Step	Effect Removed	Number Effects In	Number Parms In	SBC	ASE	Validation ASE
0		20	32	5779.6460	185773538	252878776
1	Season_Sold	19	29	5762.6753	185824120	252480746
2	House_Style2	18	25	5750.8247	192832172	248469026
3	Foundation_2	17	23	5740.3830	193440101	248951925
4	Garage_Type_2	16	21	5730.0735	194137231	247966687
5	Central_Air	15	20	5724.5490	194242334	247854963*
6	Heating_QC	14	17	5721.3123	203586891	259432895
7	Masonry_Veneer	13	16	5718.5873	205646000	263660934
8	Lot_Shape_2	12	15	5717.9317*	209193215	265159474
* Optimal Value of Criterion						

Selection stopped at a local minimum of the SBC criterion.

Stop Details				
Candidate For	Effect	Candidate SBC		Compare SBC
Removal	Deck_Porch_Area	5718.6683	>	5717.9317


Coefficient Progression for SalePrice
Standardized Coefficient
0.2
0.0
-0.2
Gr_Liv_Area
Overall_Cond2 6
Overall_Qual2 6
Overall_Cond2 5
Overall_Qual2 5
Fireplaces 2
Bedroom_AbvGr
Age_Sold
Validation ASE
2.7E8
2.6E8
2.6E8
2.5E8
Selected Step
Full Model
1-Season_Sold
2-House_Style2
3-Foundation_2
4-Garage_Type_2
5-Central_Air
6-Heating_QC
7-Masonry_Veneer
8-Lot_Shape_2
Effect Sequence



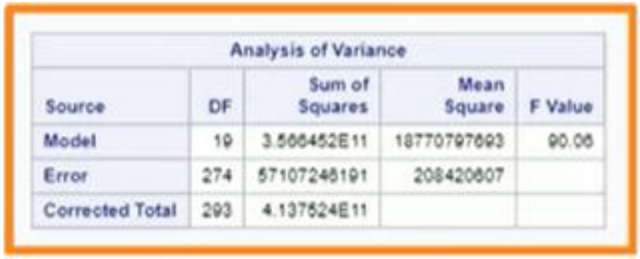


Selecting the Best Model using Honest Assessment

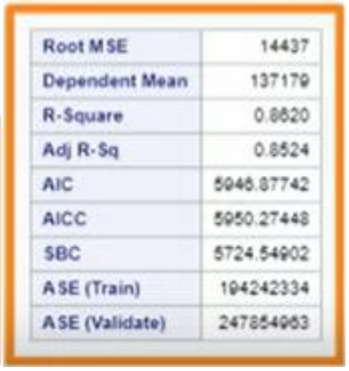
The GLMSELECT Procedure
Selected Model

The selected model, based on Validation ASE, is the model at Step 5.

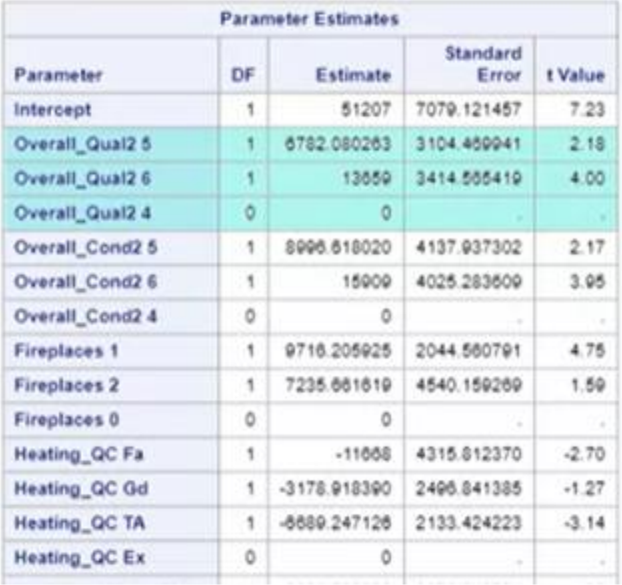
Effects:	Intercept Overall_Qual2 Overall_Cond2 Fireplaces Heating_QC Masonry_Veneer Lot_Shape_2 Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom



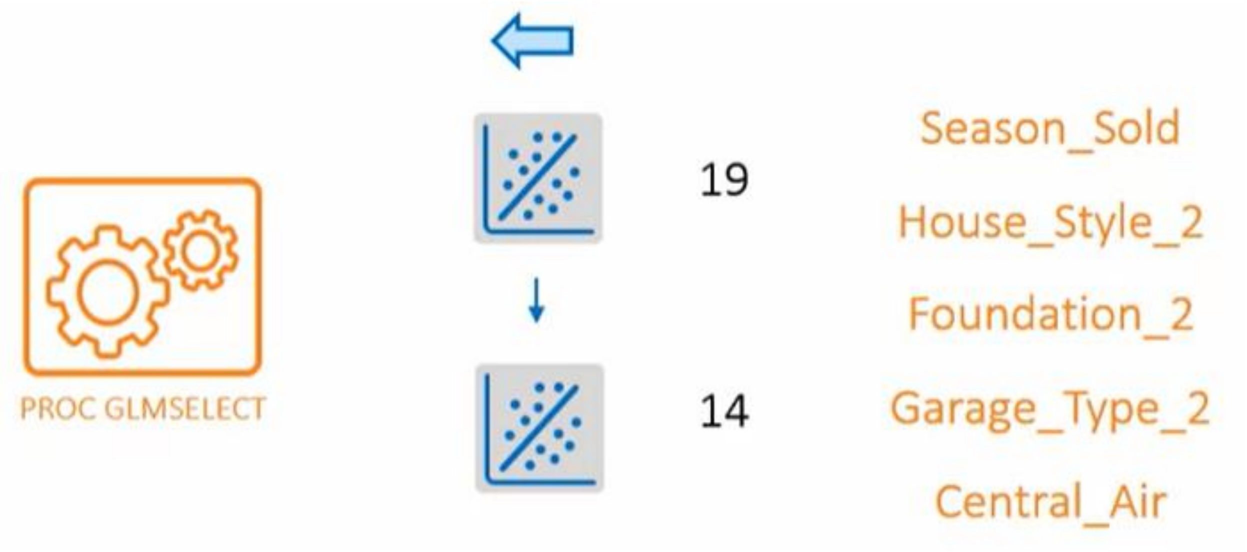
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	19	3.566452E11	18770797693	90.06
Error	274	57107248191	208420807	
Corrected Total	293	4.137524E11		



Root MSE	14437
Dependent Mean	137179
R-Square	0.8620
Adj R-Sq	0.8524
AIC	5946.87742
AICC	5950.27448
SBC	5724.54902
ASE (Train)	194242334
ASE (Validate)	247854963



Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	51207	7079.121457	7.23
Overall_Qual2 5	1	6782.080263	3104.469941	2.18
Overall_Qual2 6	1	13659	3414.565419	4.00
Overall_Qual2 4	0	0	.	.
Overall_Cond2 5	1	8996.618020	4137.937302	2.17
Overall_Cond2 6	1	15909	4025.283609	3.95
Overall_Cond2 4	0	0	.	.
Fireplaces 1	1	9716.205925	2044.560791	4.75
Fireplaces 2	1	7235.661619	4540.159269	1.59
Fireplaces 0	0	0	.	.
Heating_QC Fa	1	-11668	4315.812370	-2.70
Heating_QC Gd	1	-3178.918390	2496.841385	-1.27
Heating_QC TA	1	-6689.247126	2133.424223	-3.14
Heating_QC Ex	0	0	.	.



PROC GLMSELECT

19

14

Season_Sold

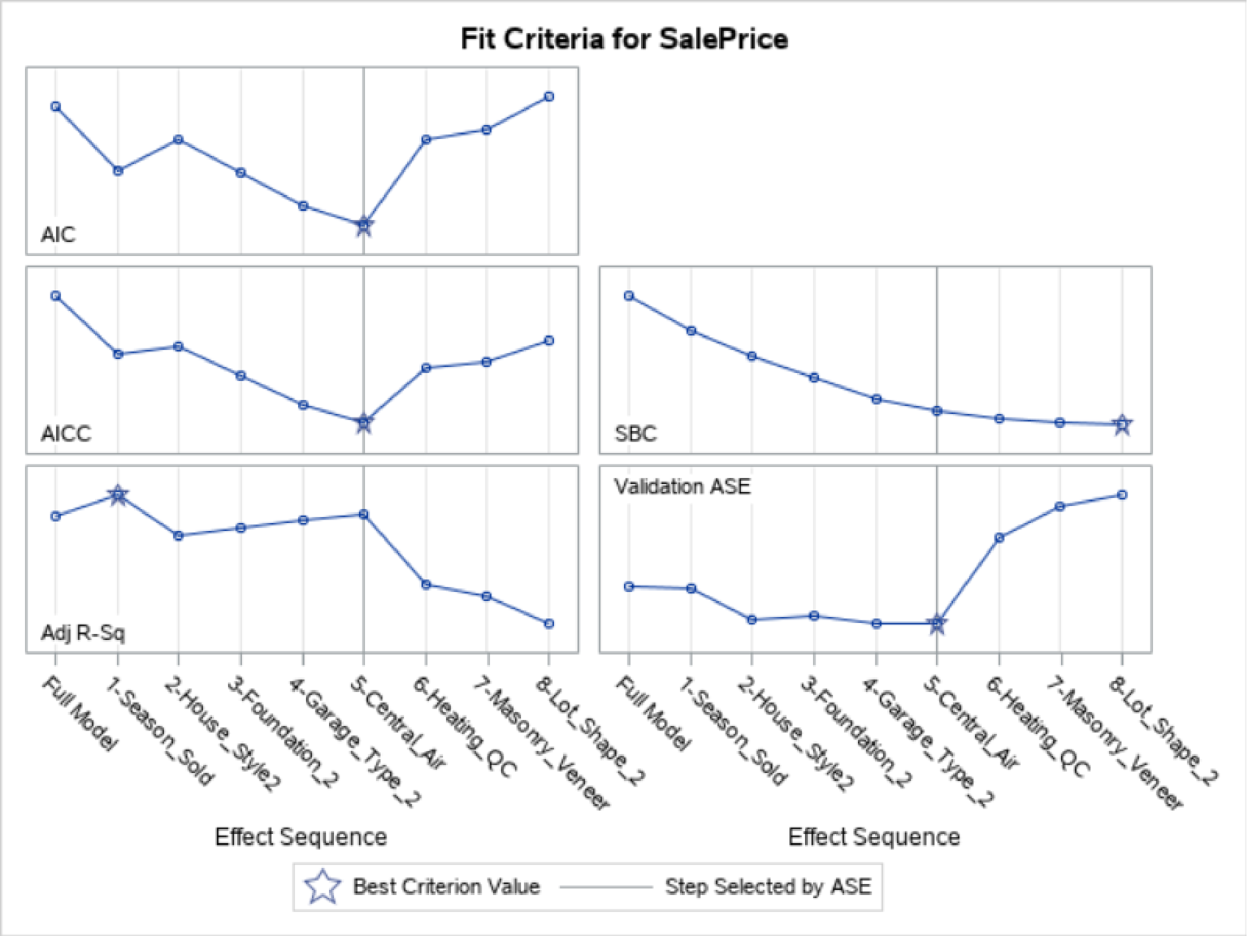
House_Style_2

Foundation_2

Garage_Type_2

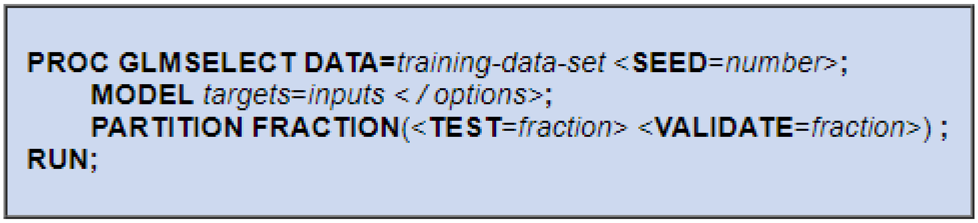
Central_Air

```
/*st106d01.sas*/

%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
        Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;
%let categorical=House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces
        Season_Sold Garage_Type_2 Foundation_2 Heating_QC
        Masonry_Veneer Lot_Shape_2 Central_Air;

ods graphics;

proc glmselect data=STAT1.ameshousing3
            plots=all
            valdata=STAT1.ameshousing4;
   class &categorical / param=glm ref=first;
   model SalePrice=&categorical &interval /
            selection=backward
            select=sbc
            choose=validate;
   store out=STAT1.amesstore;
   title "Selecting the Best Model using Honest Assessment";
run;
```


Fit Criteria for SalePrice
AIC
AICC
SBC
Adj R-Sq
Validation ASE
Full Model
1-Season_Sold
2-House_Style2
3-Foundation_2
4-Garage_Type_2
5-Central_Air
6-Heating_QC
7-Masonry_Veneer
8-Lot_Shape_2
Effect Sequence
Effect Sequence
Best Criterion Value
Step Selected by ASE

Partitioning a Data Set Using PROC GLMSELECT

If you start with a data set that's not yet partitioned, PROC GLMSELECT can partition the data for you. You can request two partitions (training and validation) or three partitions (training, validation and testing). You specify the proportion to use for the validation and test data cases, and you can specify a seed for the partitioning algorithm.



```
PROC GLMSELECT DATA=training-data-set <SEED=number>;
      MODEL targets=inputs < / options>;
      PARTITION FRACTION(<TEST=fraction> <VALIDATE=fraction>) ;
RUN;
```

In the PROC GLMSELECT statement, the DATA= option specifies the input or training data set. You'll use the PARTITION statement to specify how the cases in the input data set are partitioned into holdout samples for model validation, and if desired, testing. The MODEL statement is the same as before.

The PARTITION statement specifies how observations in the input data set are logically partitioned into disjointed subsets for model training, validation, and testing. The FRACTION option specifies the fraction (that is, the proportion) of cases in the input data set that are randomly assigned to a testing role and a validation role. The sum of the specified fractions must be less than 1 and the remaining fraction of the cases in the input data set are assigned to the training role. For example, the statement below requests two partitions (training and validation), and one quarter, or 25%, of the observations are written to the validation data set. The remaining three quarters, or 75%, are written to the training data set.

PARTITION FRACTION(VALIDATE=.25);

The PARTITION statement uses a pseudo-random number generator. To begin the random selection process, it needs a starting "seed," which must be an integer. If you want to reproduce your results in the future, specify an integer greater than zero in the SEED= option. Then, whenever you run the PROC GLMSELECT step and use the same seed value, the selection process is replicated and the same results are generated. If the SEED= value is invalid or omitted, the seed is automatically generated from the computer's clock. In most situations, it's recommended that you use the SEED= option and specify an integer greater than zero.

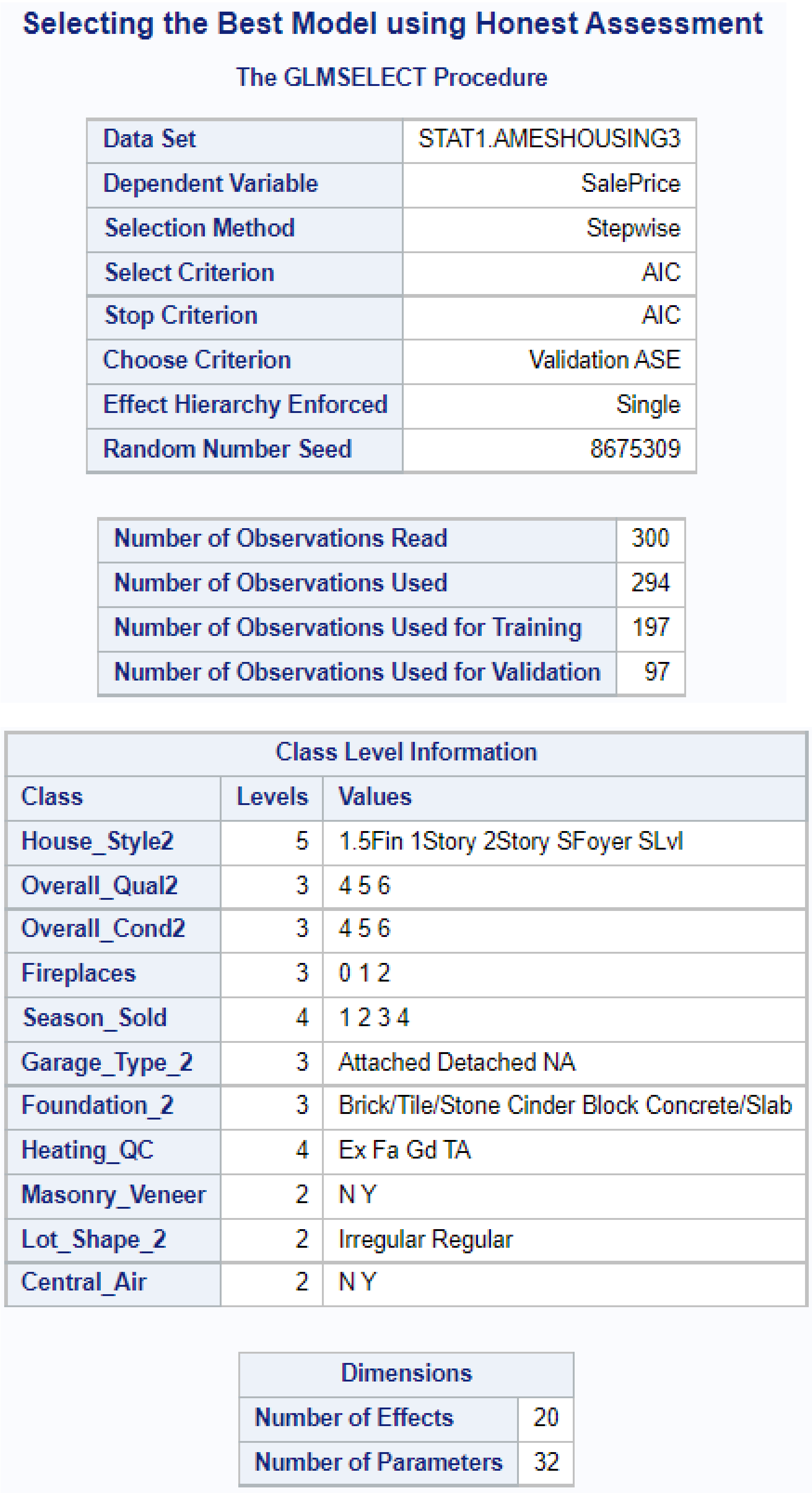
Partitioning a Data Set Using the Predictive Regression Models Task

You can use the Predictive Regression Models task to partition a data set into two or three partitions. If you want two partitions (training and validation), you must specify a sample proportion for the validation cases. This required value, which is a number between 0 and 1, represents the fraction or proportion of observations to be written to the validation partition. The remaining observations are written to the training partition.

If you also want a test partition, then you indicate that in the task, and specify a sample proportion for the testing partition. This value (a number between 0 and 1) represents the fraction or proportion of cases to be written to the testing partition. If you request both validation and testing partitions, then the sum of the specified fractions must be less than one. The remaining observations are written to the training partition.

You can use the random seed option to specify a starting seed for the pseudo-random number generator. If you specify an integer that's greater than zero, you can reproduce the results in the future. If you omit this option, a random seed will be generated, and the results will be different each time you submit the code.

```
/*st106s01.sas*/

%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
        Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;
%let categorical=House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces
        Season_Sold Garage_Type_2 Foundation_2 Heating_QC
        Masonry_Veneer Lot_Shape_2 Central_Air;


/*In this example, the data set ameshousing3 is divided into */
/*training and validation using the PARTITION statement, */
/*along with the SEED= option in the PROC GLMSELECT statement.*/
proc glmselect data=STAT1.ameshousing3
             plots=all
             seed=8675309;
   class &categorical / param=ref ref=first;
   model SalePrice=&categorical &interval /
             selection=stepwise
             (select=aic
             choose=validate) hierarchy=single;
   partition fraction(validate=0.3333);
   title "Selecting the Best Model using Honest Assessment";
run;
```



Selecting the Best Model using Honest Assessment

The GLMSELECT Procedure

Data Set	STAT1.AMESHOUSING3
Dependent Variable	SalePrice
Selection Method	Stepwise
Select Criterion	AIC
Stop Criterion	AIC
Choose Criterion	Validation ASE
Effect Hierarchy Enforced	Single
Random Number Seed	8675309

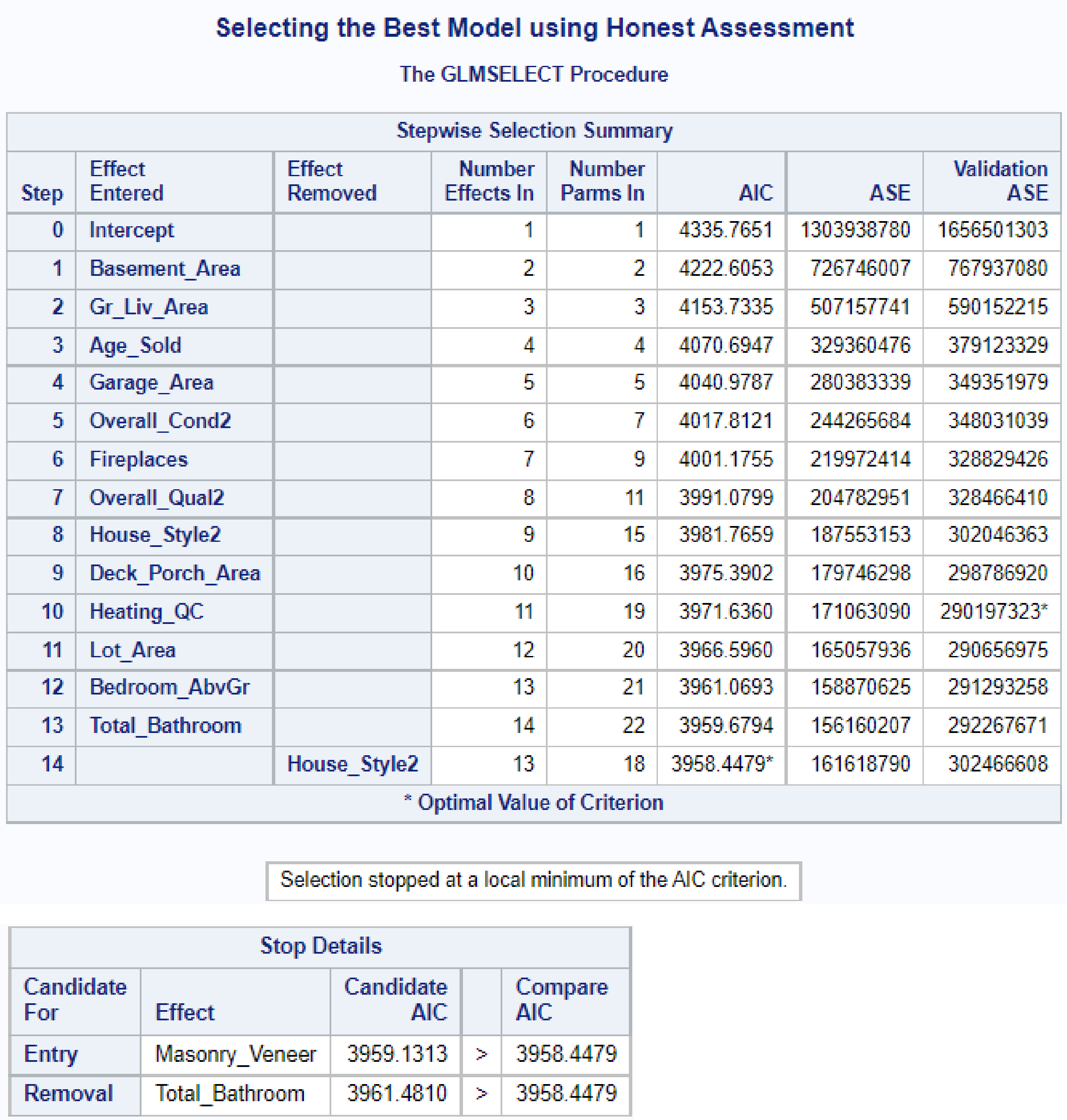
Number of Observations Read	300
Number of Observations Used	294
Number of Observations Used for Training	197
Number of Observations Used for Validation	97

Class Level Information

Class	Levels	Values
House_Style2	5	1.5Fin 1Story 2Story SFoyer SLvl
Overall_Qual2	3	4 5 6
Overall_Cond2	3	4 5 6
Fireplaces	3	0 1 2
Season_Sold	4	1 2 3 4
Garage_Type_2	3	Attached Detached NA
Foundation_2	3	Brick/Tile/Stone Cinder Block Concrete/Slab
Heating_QC	4	Ex Fa Gd TA
Masonry_Veneer	2	N Y
Lot_Shape_2	2	Irregular Regular
Central_Air	2	N Y

Dimensions

Number of Effects	20
Number of Parameters	32



Selecting the Best Model using Honest Assessment

The GLMSELECT Procedure

Stepwise Selection Summary

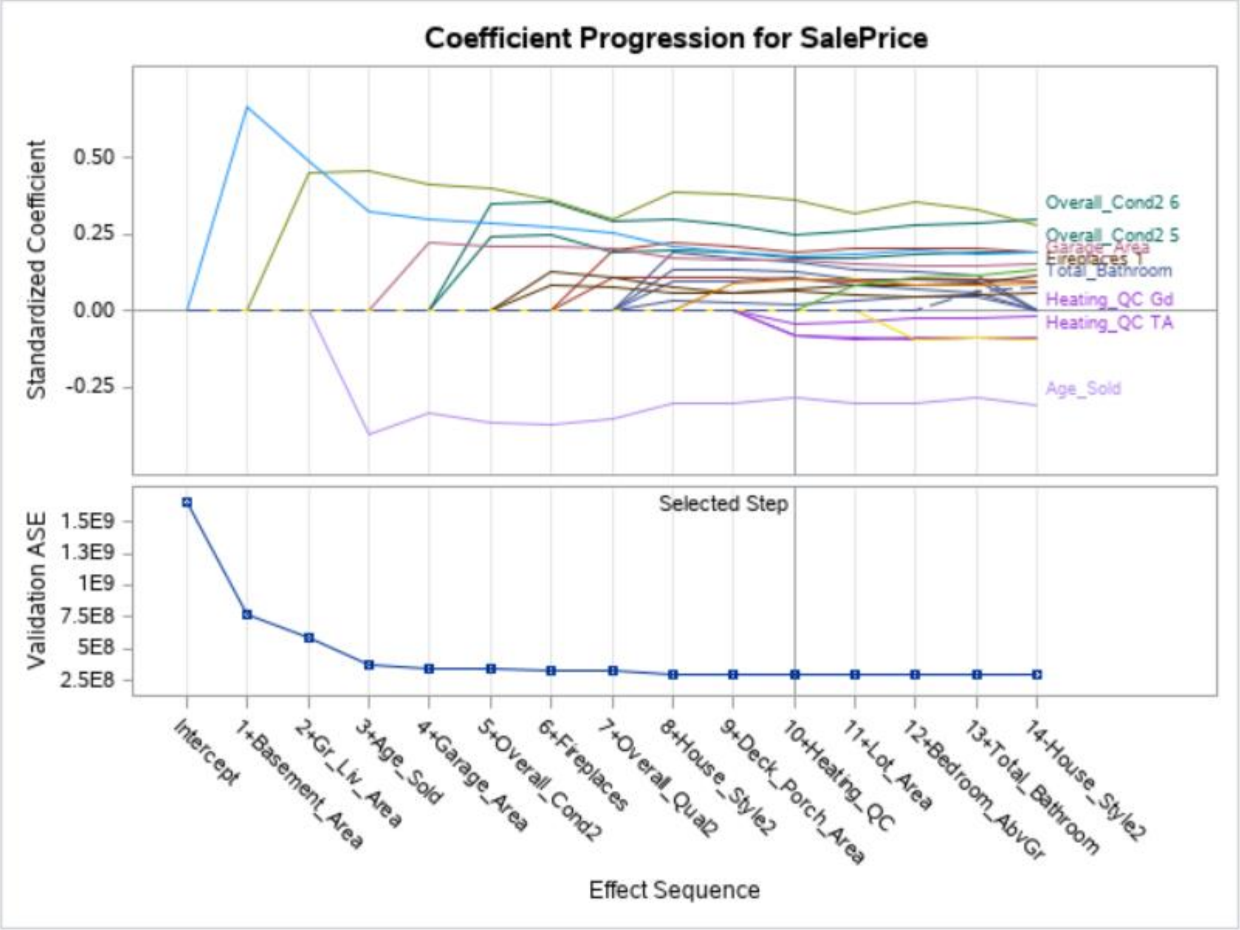
Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	AIC	ASE	Validation ASE
0	Intercept		1	1	4335.7651	1303938780	1656501303
1	Basement_Area		2	2	4222.6053	726746007	767937080
2	Gr_Liv_Area		3	3	4153.7335	507157741	590152215
3	Age_Sold		4	4	4070.6947	329360476	379123329
4	Garage_Area		5	5	4040.9787	280383339	349351979
5	Overall_Cond2		6	7	4017.8121	244265684	348031039
6	Fireplaces		7	9	4001.1755	219972414	328829426
7	Overall_Qual2		8	11	3991.0799	204782951	328466410
8	House_Style2		9	15	3981.7659	187553153	302046363
9	Deck_Porch_Area		10	16	3975.3902	179746298	298786920
10	Heating_QC		11	19	3971.6360	171063090	290197323*
11	Lot_Area		12	20	3966.5960	165057936	290656975
12	Bedroom_AbvGr		13	21	3961.0693	158870625	291293258
13	Total_Bathroom		14	22	3959.6794	156160207	292267671
14		House_Style2	13	18	3958.4479*	161618790	302466608

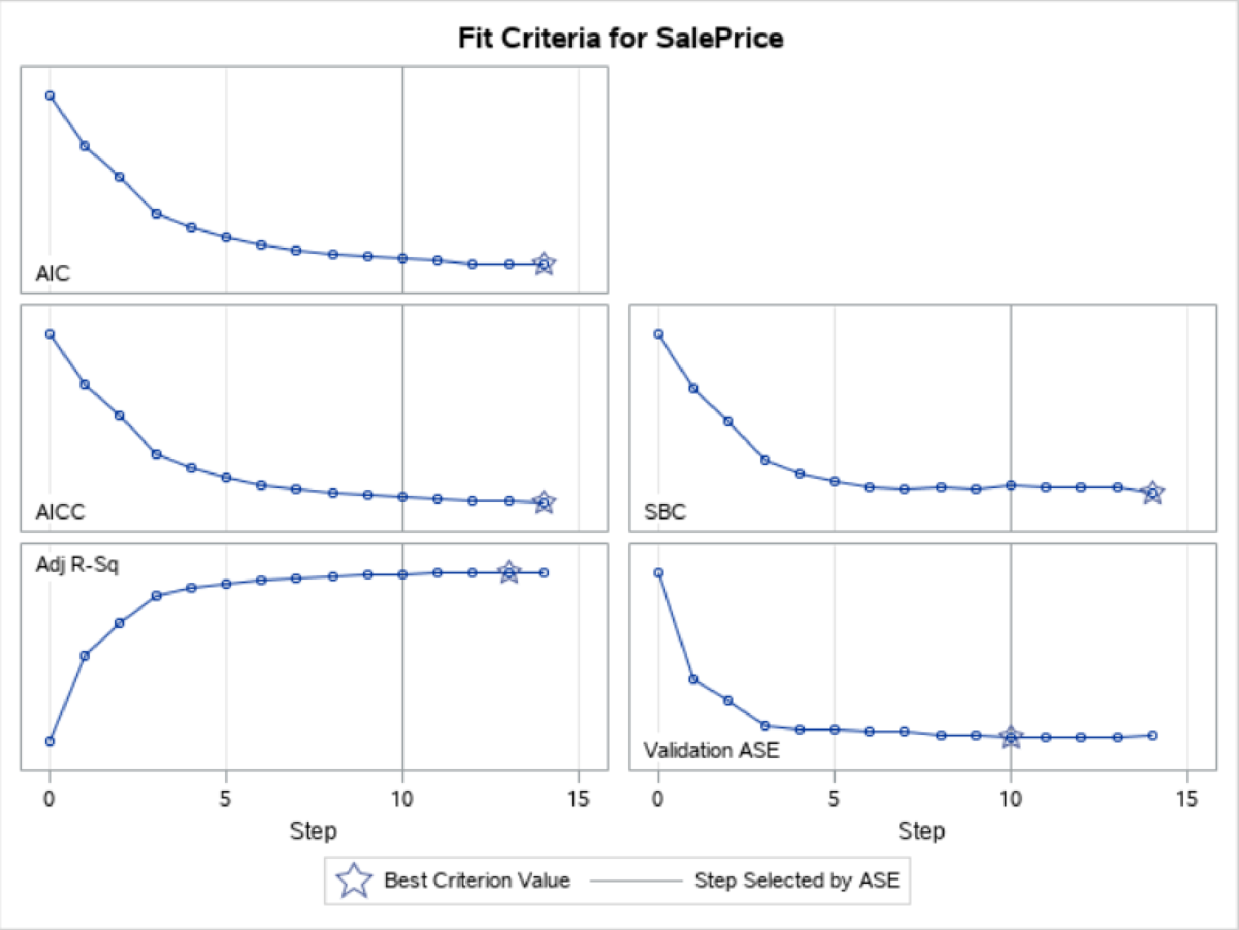
* Optimal Value of Criterion

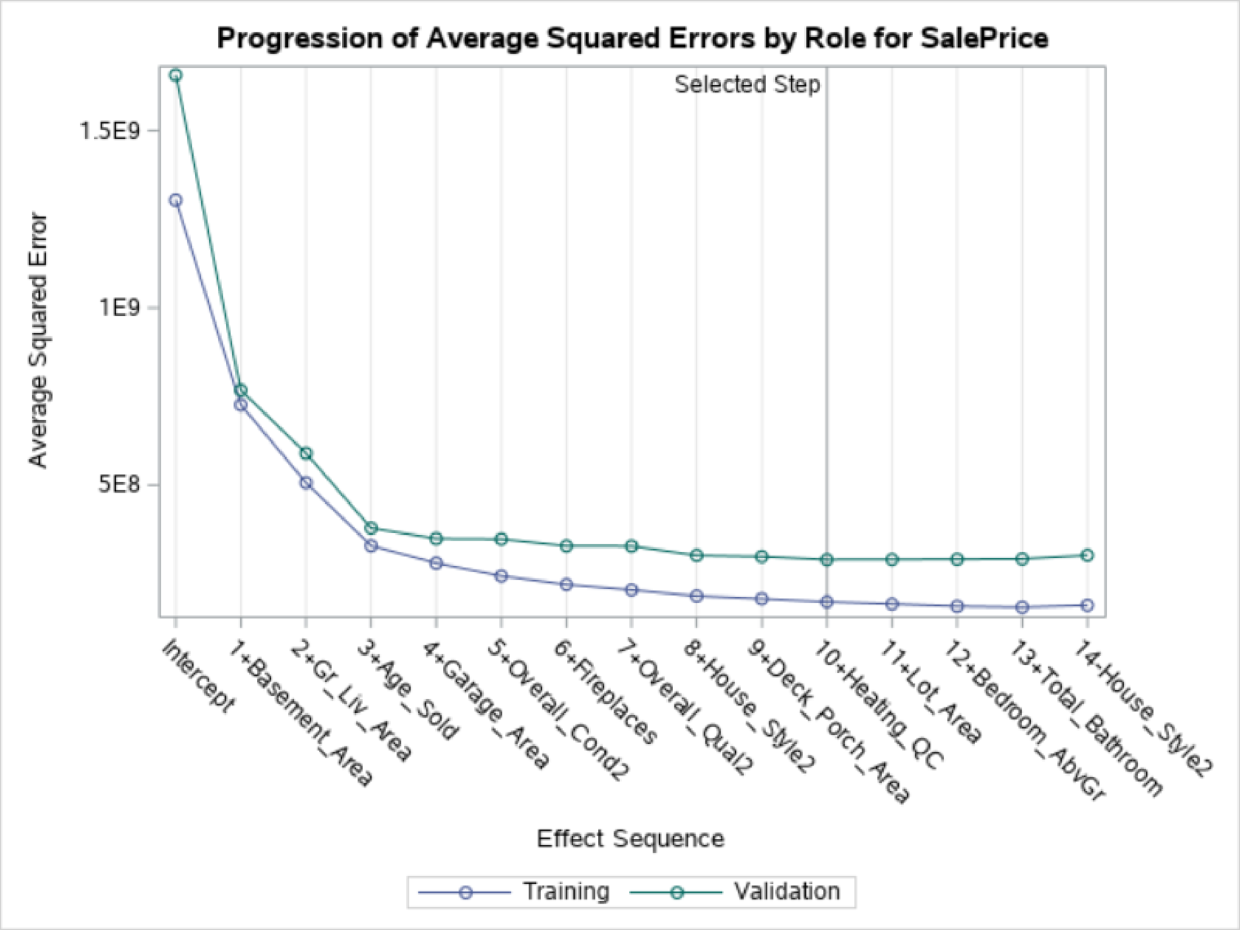
Selection stopped at a local minimum of the AIC criterion.

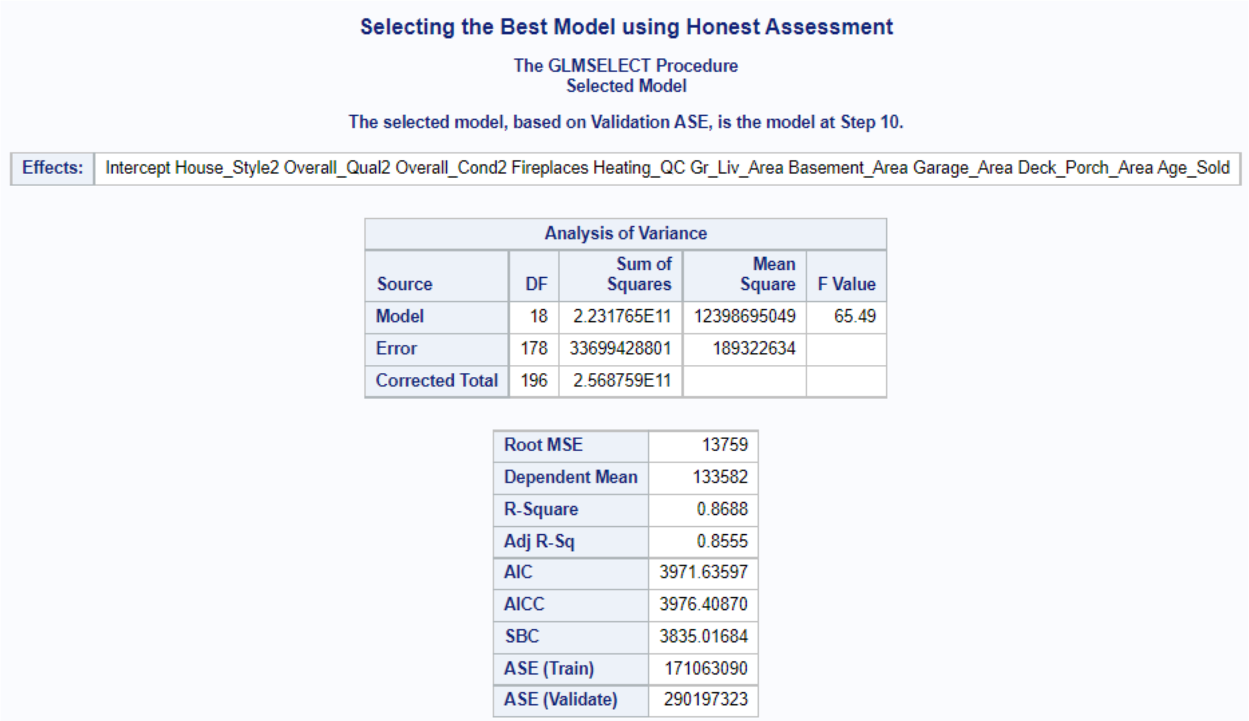
Stop Details

Candidate For	Effect	Candidate AIC		Compare AIC
Entry	Masonry_Veneer	3959.1313	>	3958.4479
Removal	Total_Bathroom	3961.4810	>	3958.4479


Coefficient Progression for SalePrice
Standardized Coefficient
0.50
0.25
0.00
-0.25
Overall_Cond2 6
Overall_Cond2 5
Garage_Area
Fireplaces 1
Total_Bathroom
Heating_QC Gd
Heating_QC TA
Age_Sold
Selected Step
Validation ASE
1.5E9
1.3E9
1E9
7.5E8
5E8
2.5E8
Intercept
1+Basement_Area
2+Gr_Liv_Area
3+Age_Sold
4+Garage_Area
5+Overall_Cond2
6+Fireplaces
7+Overall_Qual2
8+House_Style2
9+Deck_Porch_Area
10+Heating_QC
11+Lot_Area
12+Bedroom_AbvGr
13+Total_Bathroom
14-House_Style2
Effect Sequence


Fit Criteria for SalePrice
AIC
AICC
SBC
Adj R-Sq
Validation ASE
0
5
10
15
Step
Best Criterion Value
Step Selected by ASE





Selecting the Best Model using Honest Assessment

The GLMSELECT Procedure
Selected Model

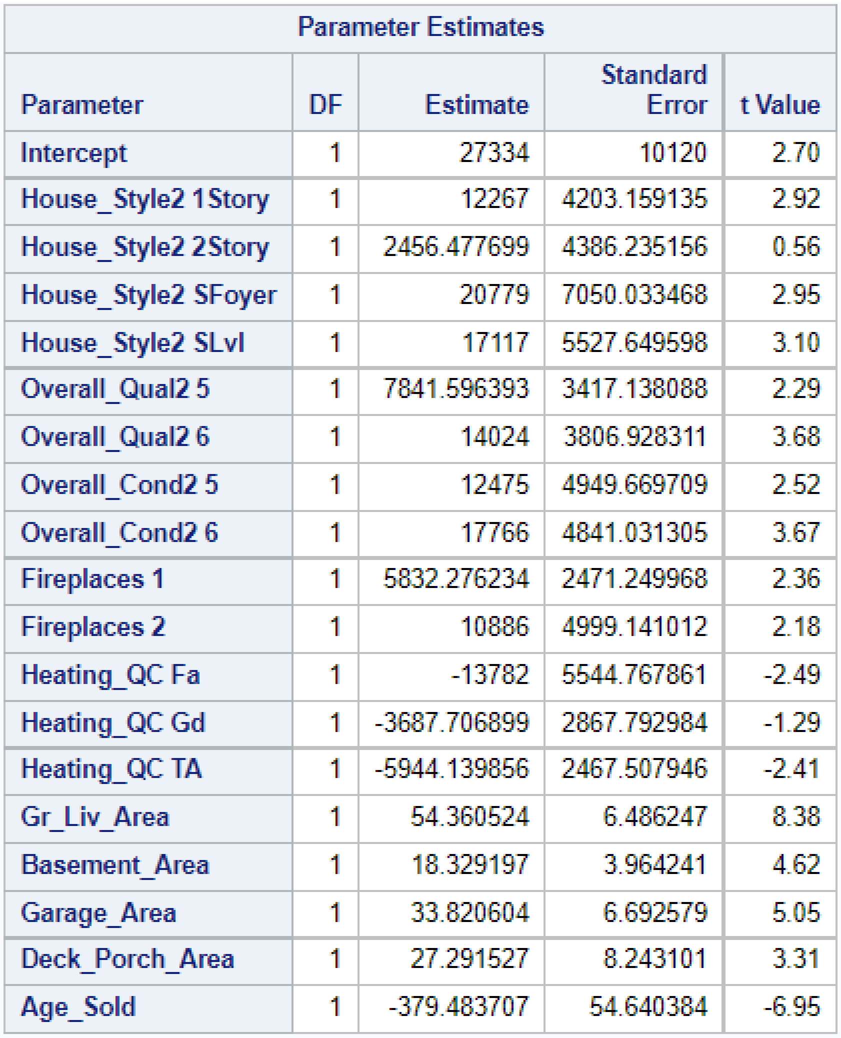
The selected model, based on Validation ASE, is the model at Step 10.

Effects:	Intercept House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces Heating_QC Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area Age_Sold

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	18	2.231765E11	12398695049	65.49
Error	178	33699428801	189322634	
Corrected Total	196	2.568759E11		

Root MSE	13759
Dependent Mean	133582
R-Square	0.8688
Adj R-Sq	0.8555
AIC	3971.63597
AICC	3976.40870
SBC	3835.01684
ASE (Train)	171063090
ASE (Validate)	290197323



Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	27334	10120	2.70
House_Style2 1Story	1	12267	4203.159135	2.92
House_Style2 2Story	1	2456.477699	4386.235156	0.56
House_Style2 SFoyer	1	20779	7050.033468	2.95
House_Style2 SLvl	1	17117	5527.649598	3.10
Overall_Qual2 5	1	7841.596393	3417.138088	2.29
Overall_Qual2 6	1	14024	3806.928311	3.68
Overall_Cond2 5	1	12475	4949.669709	2.52
Overall_Cond2 6	1	17766	4841.031305	3.67
Fireplaces 1	1	5832.276234	2471.249968	2.36
Fireplaces 2	1	10886	4999.141012	2.18
Heating_QC Fa	1	-13782	5544.767861	-2.49
Heating_QC Gd	1	-3687.706899	2867.792984	-1.29
Heating_QC TA	1	-5944.139856	2467.507946	-2.41
Gr_Liv_Area	1	54.360524	6.486247	8.38
Basement_Area	1	18.329197	3.964241	4.62
Garage_Area	1	33.820604	6.692579	5.05
Deck_Porch_Area	1	27.291527	8.243101	3.31
Age_Sold	1	-379.483707	54.640384	-6.95

Practice: Building a Predictive Model Using PROC GLMSELECT

Question 1

Use the **ameshousing3** data set to build a model that predicts the sale prices of homes in Ames, Iowa, that are 1500 square feet or below, based on various home characteristics.

1. Write a PROC GLMSELECT step that predicts the values of **SalePrice**. Partition the **stat1.ameshousing3** data set into a training data set of approximately 2/3 and a validation data set of approximately 1/3. Specify the seed 8675309. Define the **Interval** and **Categorical** macro variables as shown below, and use them to specify the inputs. Use stepwise regression as the selection method, Akaike's information criterion (AIC) to add and or remove effects, and average squared error for the validation data to select the best model. Add the REF=FIRST option in the CLASS statement.

```
%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;
%let categorical=House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces
         Season_Sold Garage_Type_2 Foundation_2 Heating_QC
         Masonry_Veneer Lot_Shape_2 Central_Air;
```

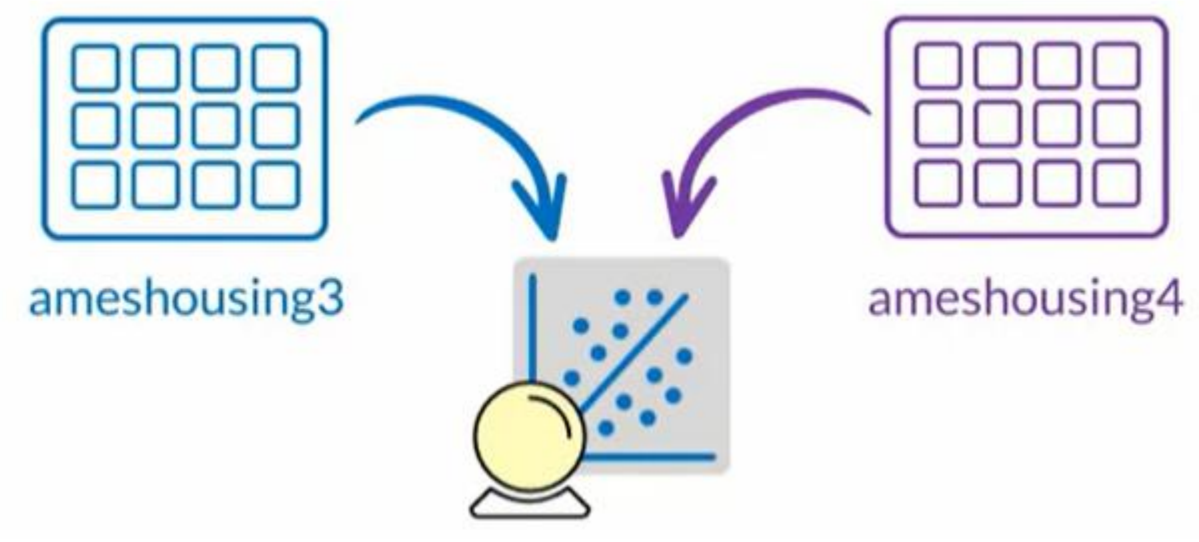
Submit the code and examine the results. Which model did PROC GLMSELECT choose?

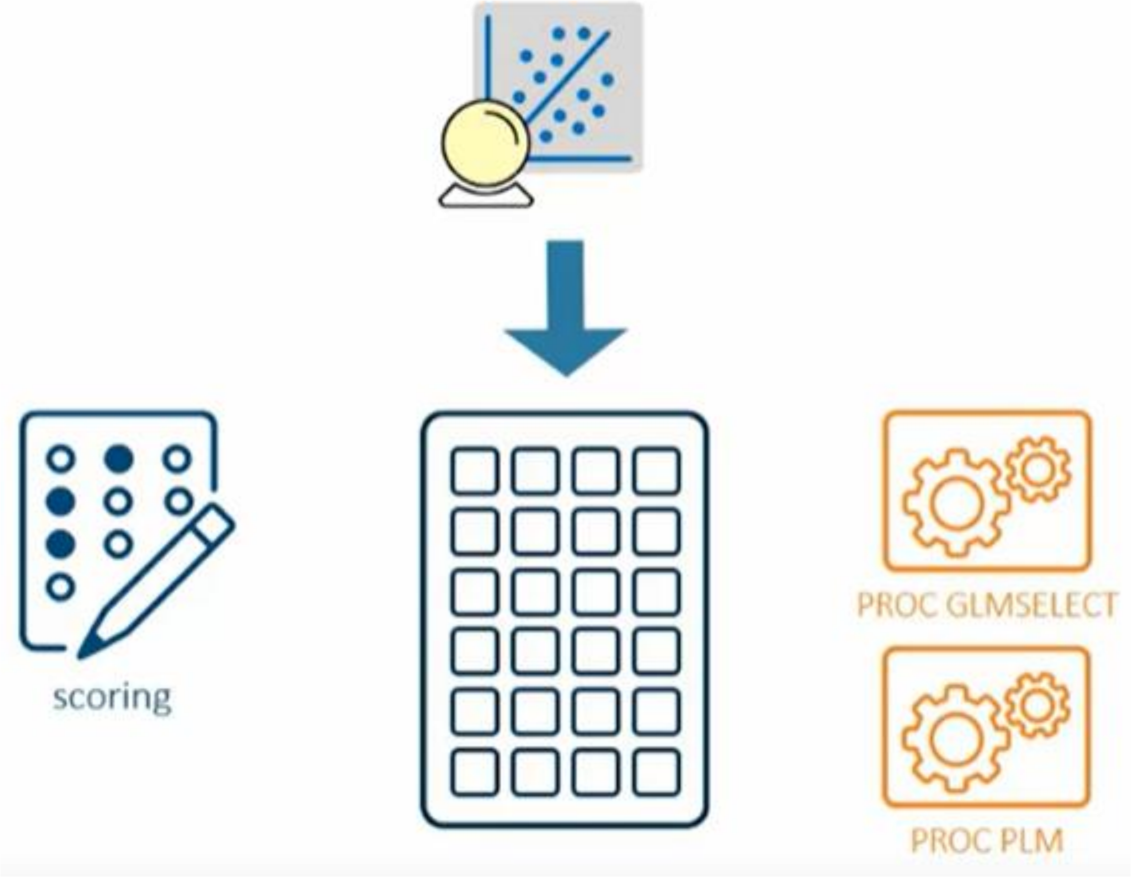
PROC GLMSELECT chose the model at Step 10, which has the following effects:**Intercept**, **Basement_Area**, **Gr_Liv_Area**, **Age_Sold**, **Garage_Area**, **Overall_Cond2**, **Fireplaces**, **Overall_Qual2**, **House_Style2**, **Deck_Porch_Area**, and **Heating_QC**.

```
/*st106s01.sas*/

%let interval=Gr_Liv_Area Basement_Area Garage_Area Deck_Porch_Area 
         Lot_Area Age_Sold Bedroom_AbvGr Total_Bathroom;
%let categorical=House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces 
         Season_Sold Garage_Type_2 Foundation_2 Heating_QC 
         Masonry_Veneer Lot_Shape_2 Central_Air;


/*In this example, the data set ameshousing3 is divided into */
/*training and validation using the PARTITION statement, */
/*along with the SEED= option in the PROC GLMSELECT statement.*/
proc glmselect data=STAT1.ameshousing3
               plots=all 
               seed=8675309;
   class &categorical / param=ref ref=first;
   model SalePrice=&categorical &interval / 
                   selection=stepwise
                   (select=aic 
                   choose=validate) hierarchy=single;
   partition fraction(validate=0.3333);
   title "Selecting the Best Model using Honest Assessment";
run;
```

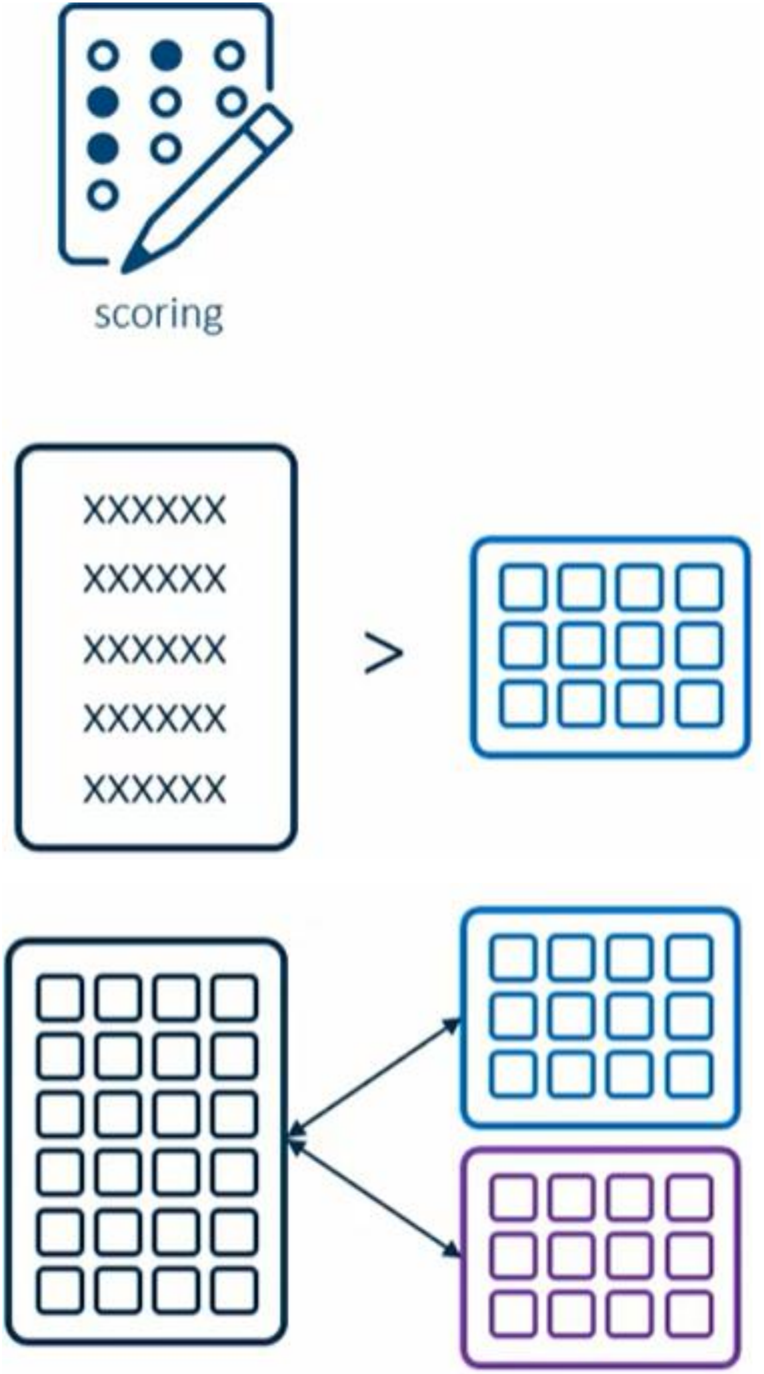
Scoring Predictive Models

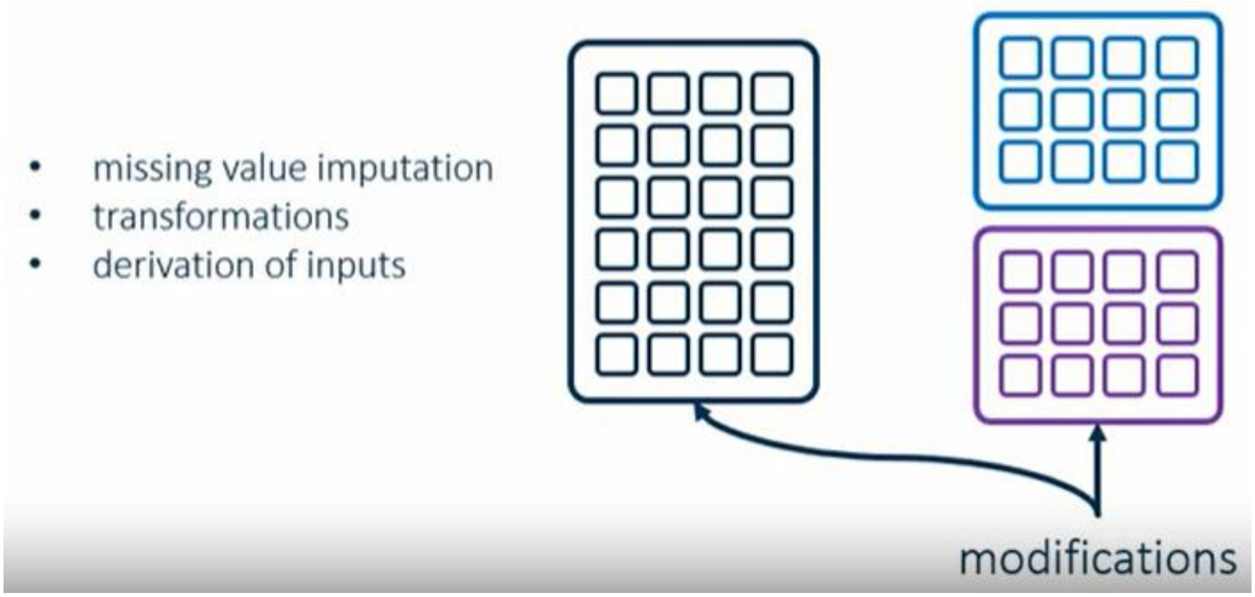
Scenario



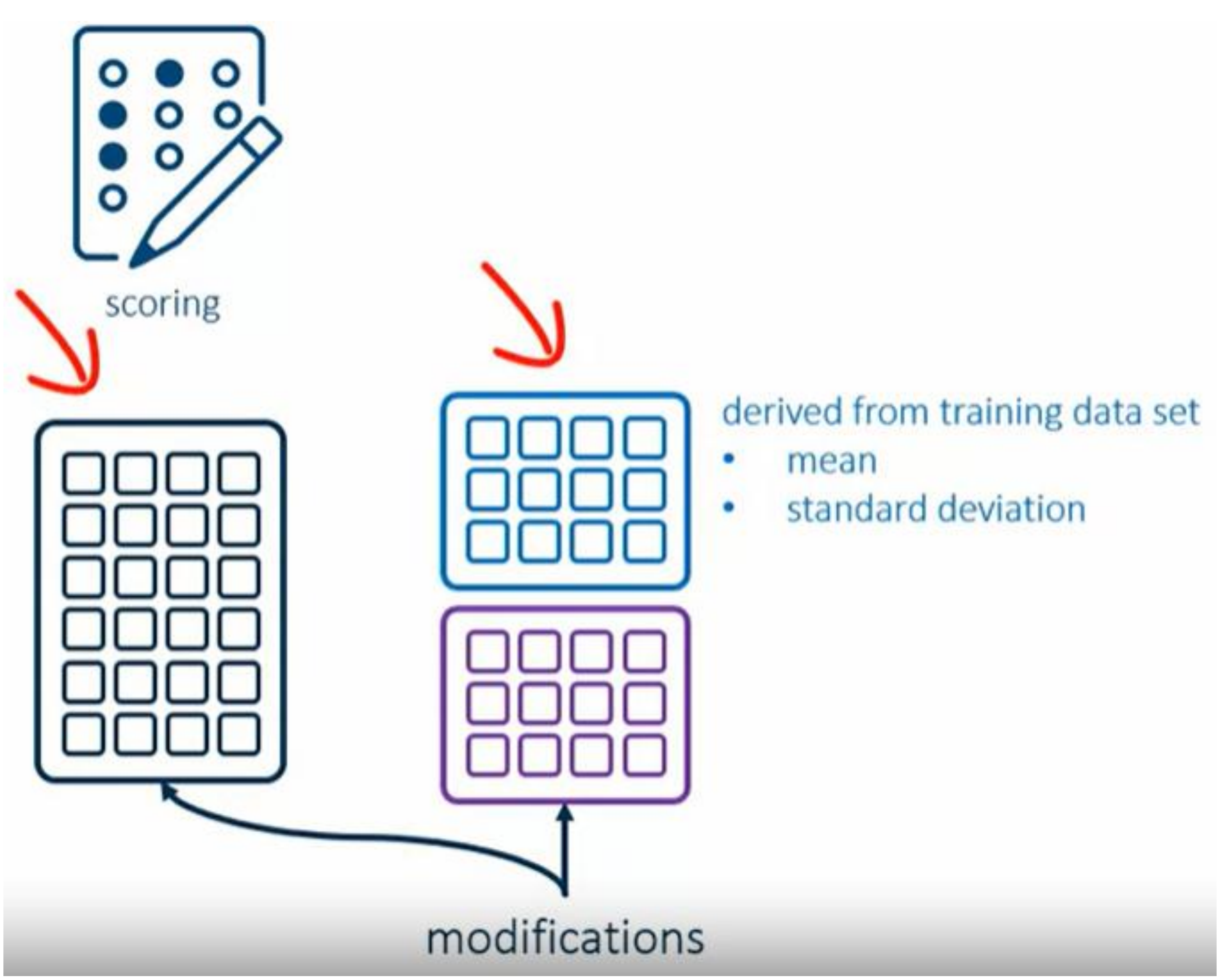

scoring
PROC GLMSELECT
PROC PLM

Preparing for Scoring



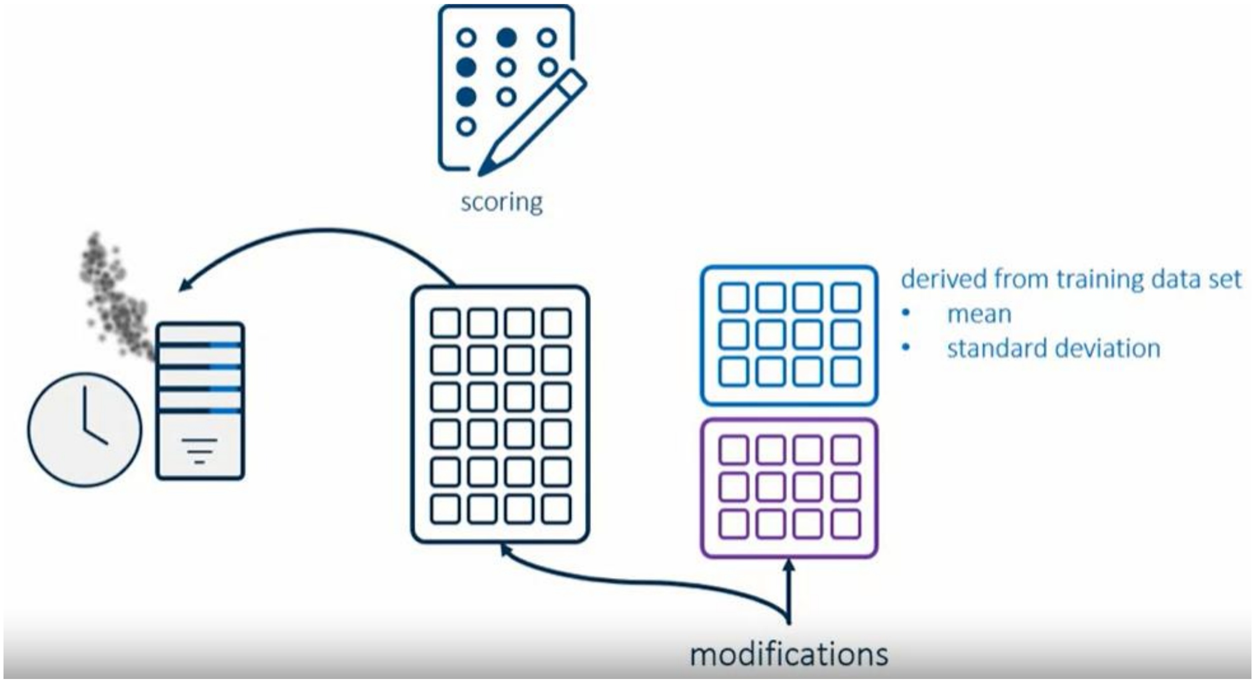


- missing value imputation
- transformations
- derivation of inputs



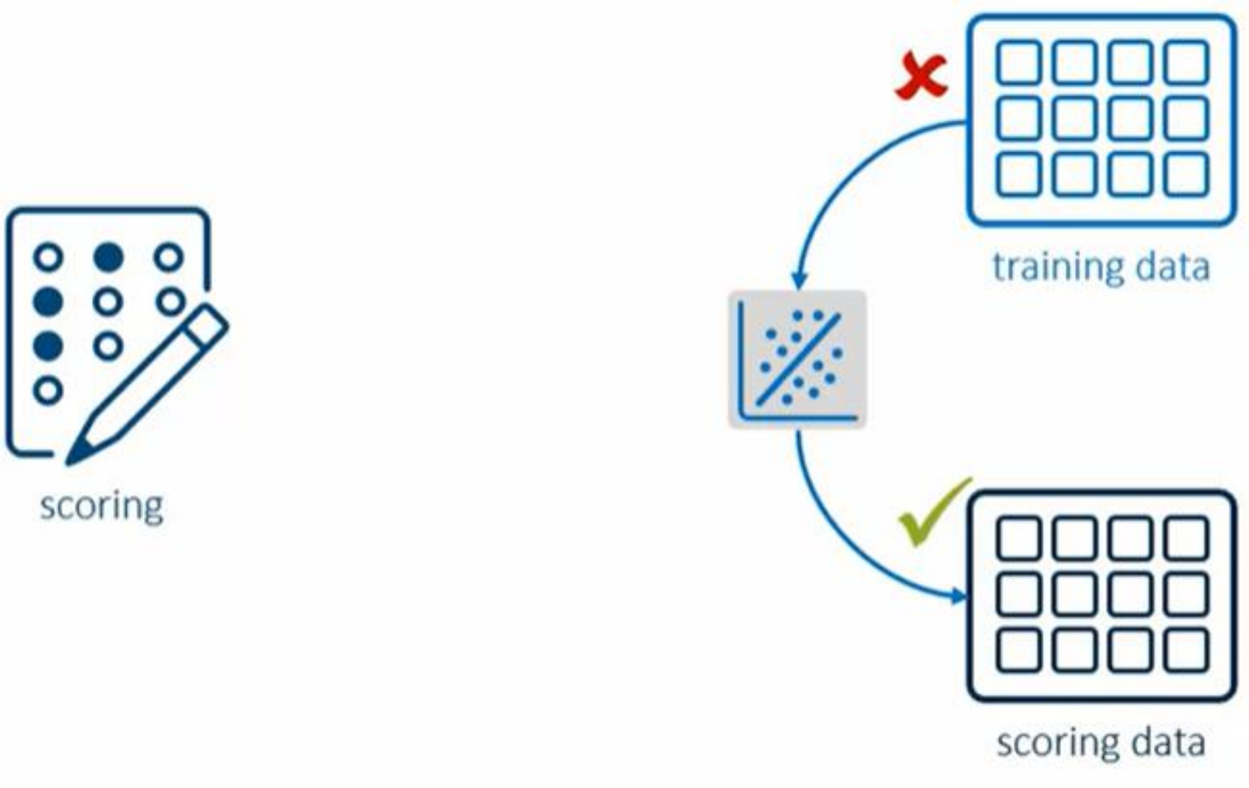
derived from training data set

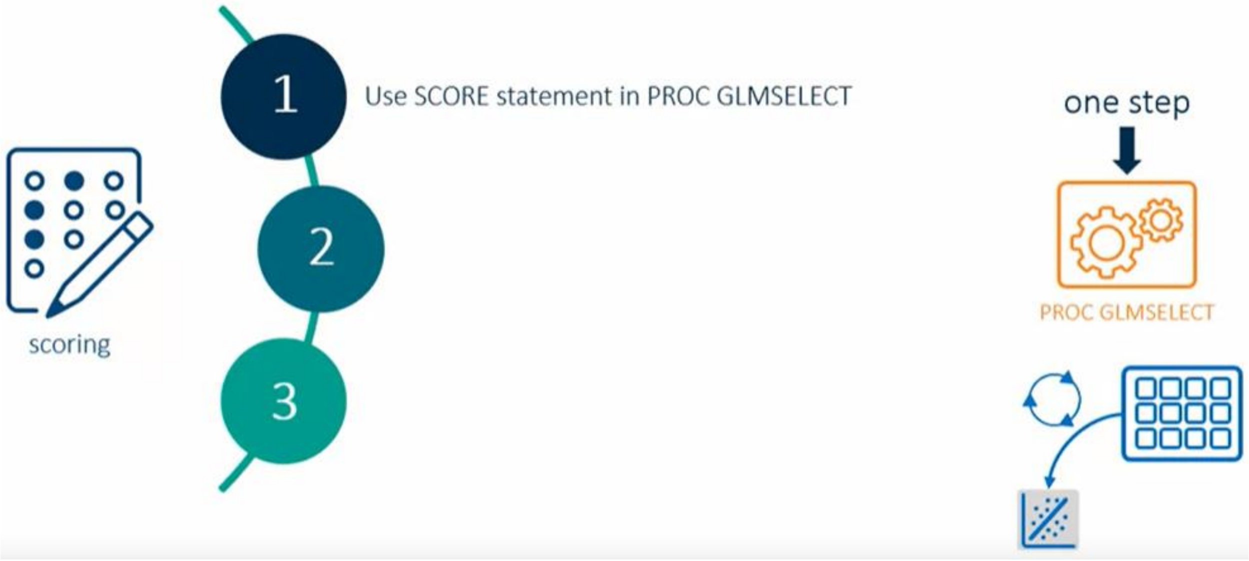
- mean
- standard deviation

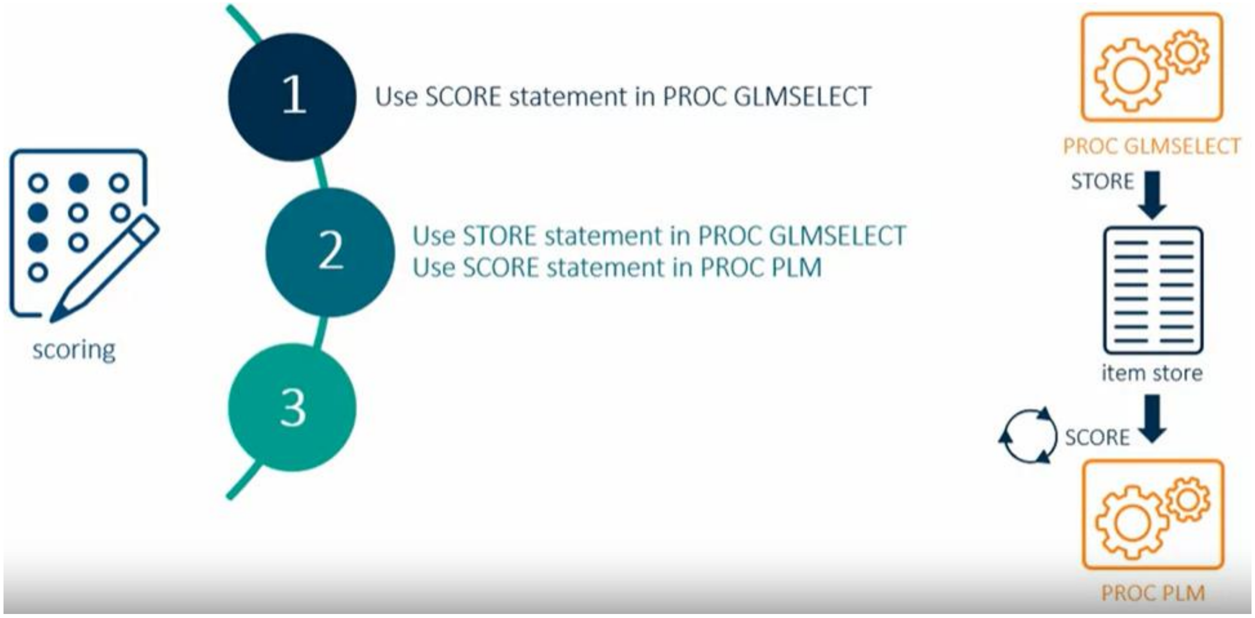


scoring
derived from training data set
• mean
• standard deviation
modifications

Methods of Scoring



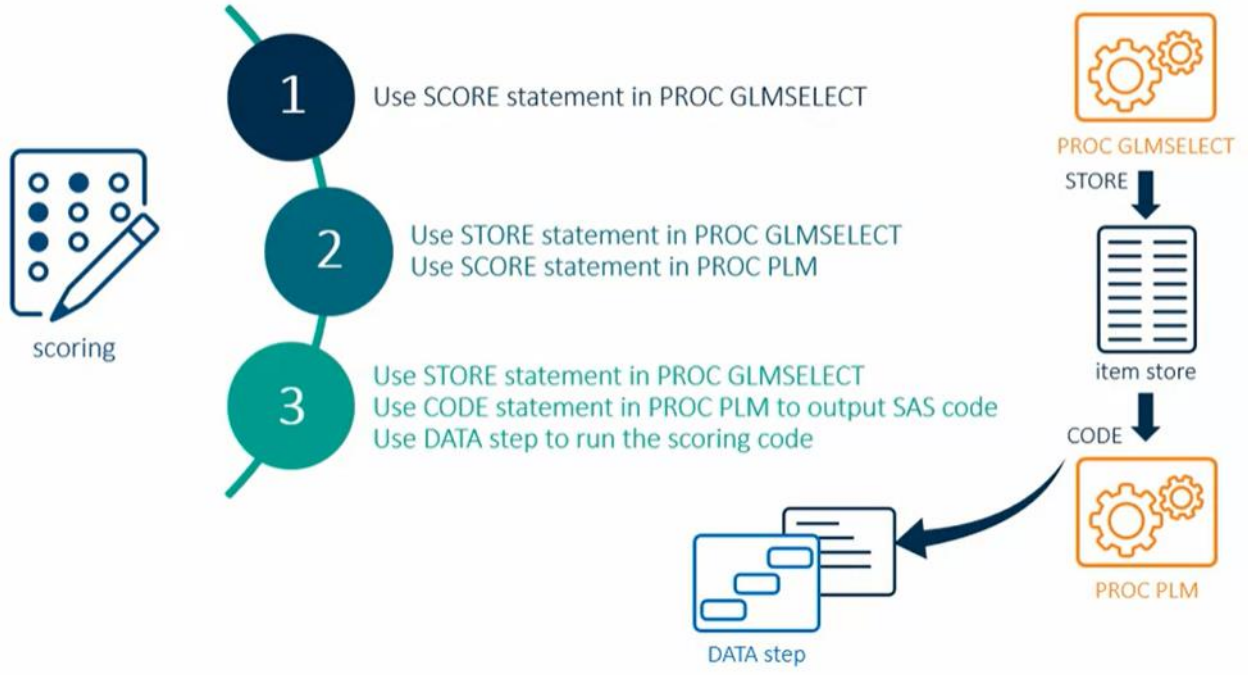

1 Use SCORE statement in PROC GLMSELECT



1 Use SCORE statement in PROC GLMSELECT

2 Use STORE statement in PROC GLMSELECT
Use SCORE statement in PROC PLM

3

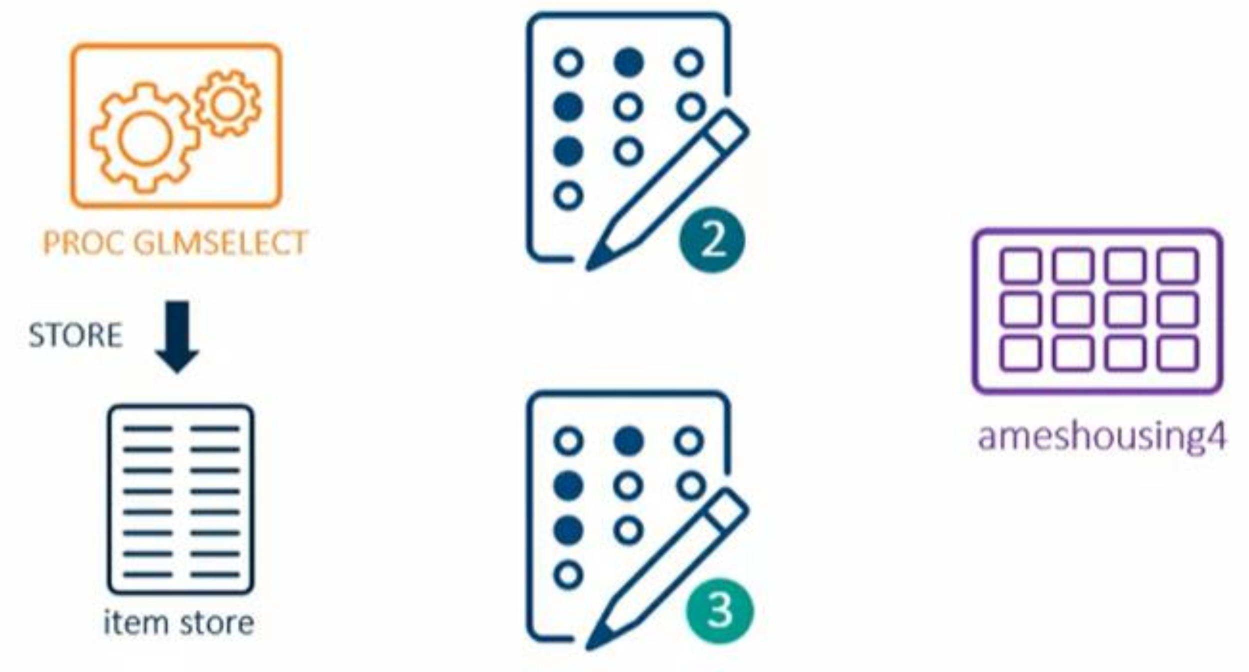


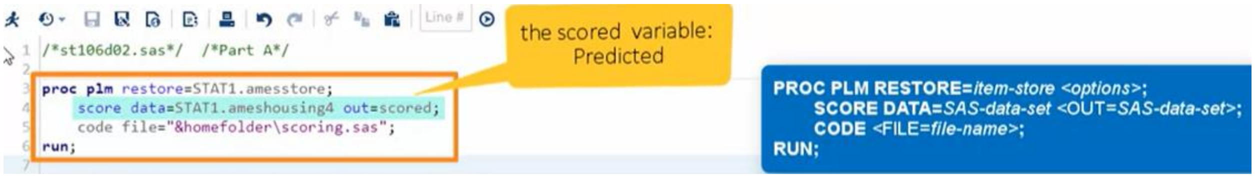
1 Use SCORE statement in PROC GLMSELECT

2 Use STORE statement in PROC GLMSELECT
Use SCORE statement in PROC PLM

3 Use STORE statement in PROC GLMSELECT
Use CODE statement in PROC PLM to output SAS code
Use DATA step to run the scoring code

Demo Scoring Data Using PROC PLM

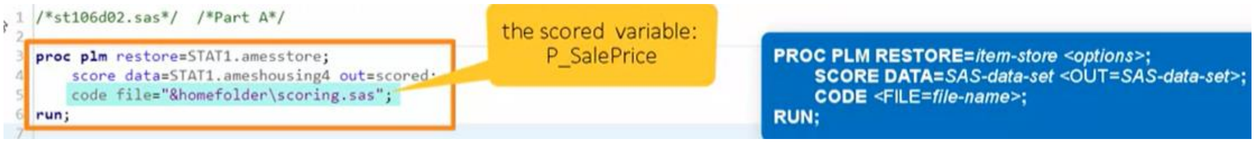




```
/*st106d02.sas*/  /*Part A*/

proc plm restore=STAT1.amesstore;
    score data=STAT1.ameshousing4 out=scored;
    code file="&homefolder\scoring.sas";
run;
```

the scored variable: Predicted

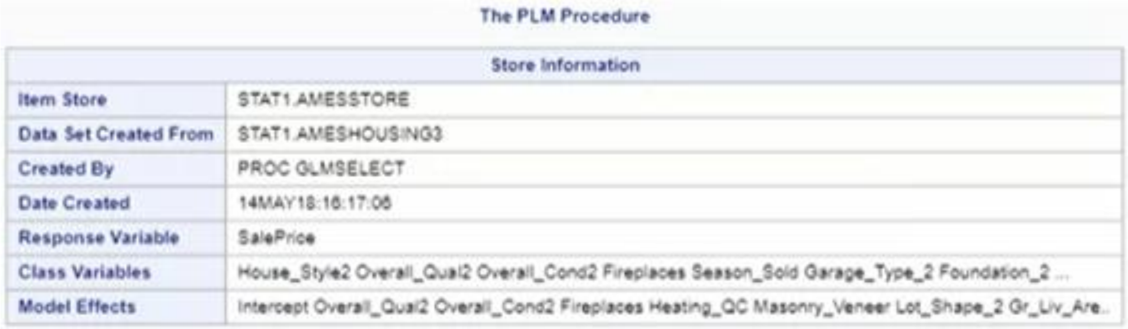
PROC PLM RESTORE=*item-store <options>***;**
    SCORE DATA=*SAS-data-set* <OUT=*SAS-data-set*>;
    CODE <FILE=*file-name*>;
RUN;



```
/*st106d02.sas*/  /*Part A*/

proc plm restore=STAT1.amesstore;
    score data=STAT1.ameshousing4 out=scored;
    code file="&homefolder\scoring.sas";
run;
```

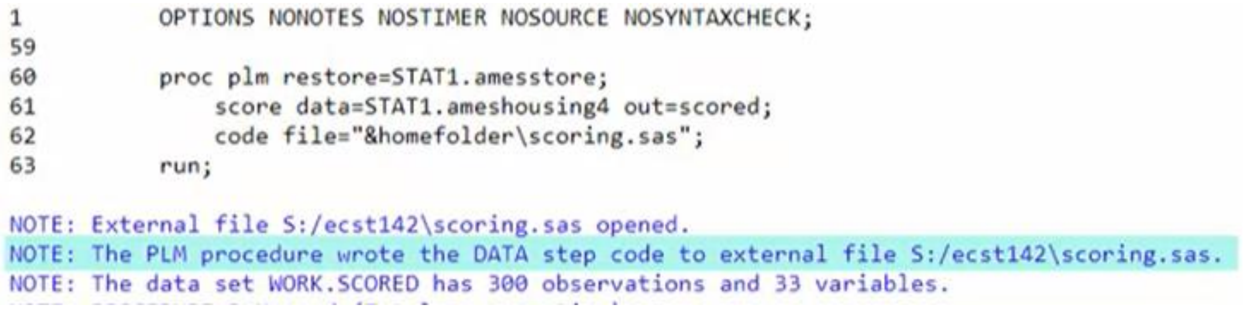
the scored variable: P_SalePrice

PROC PLM RESTORE=*item-store <options>***;**
    SCORE DATA=*SAS-data-set* <OUT=*SAS-data-set*>;
    CODE <FILE=*file-name*>;
RUN;

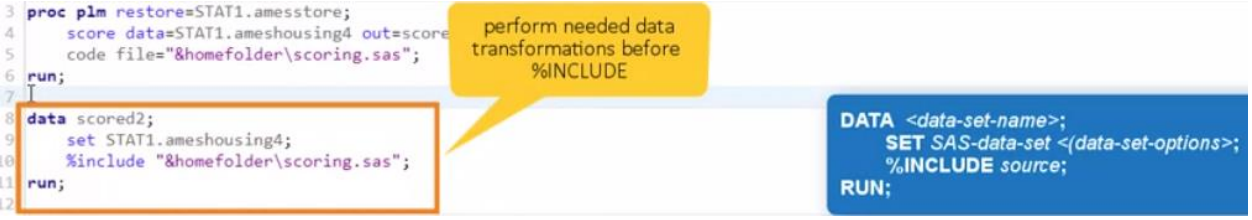


The PLM Procedure

Store Information	
Item Store	STAT1.AMESSTORE
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC GLMSELECT
Date Created	14MAY18:16:17:06
Response Variable	SalePrice
Class Variables	House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces Season_Sold Garage_Type_2 Foundation_2 ...
Model Effects	Intercept Overall_Qual2 Overall_Cond2 Fireplaces Heating_QC Masonry_Veneer Lot_Shape_2 Gr_Liv_Are..



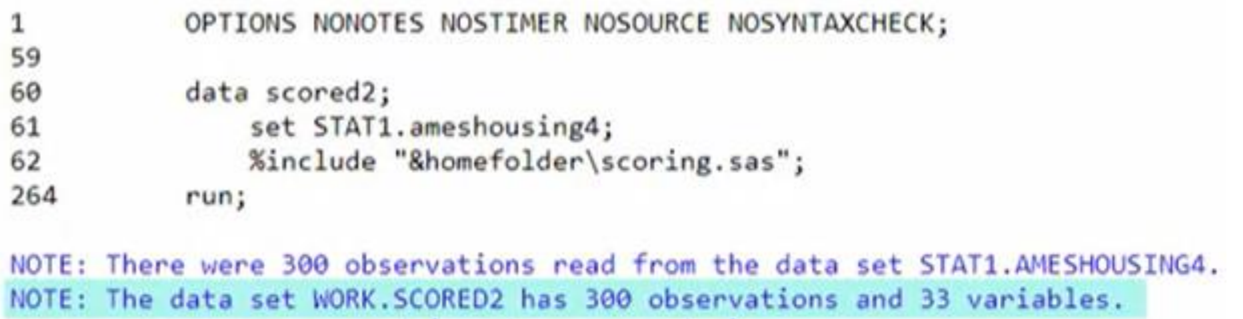
```
OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;

proc plm restore=STAT1.amesstore;
    score data=STAT1.ameshousing4 out=scored;
    code file="&homefolder\scoring.sas";
run;

NOTE: External file S:/ecst142\scoring.sas opened.
NOTE: The PLM procedure wrote the DATA step code to external file S:/ecst142\scoring.sas.
NOTE: The data set WORK.SCORED has 300 observations and 33 variables.
```



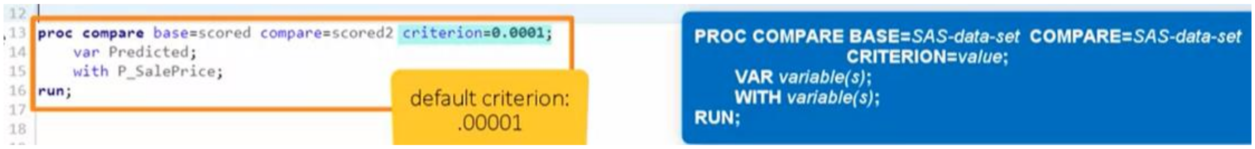
```
proc plm restore=STAT1.amesstore;
    score data=STAT1.ameshousing4 out=score
    code file="&homefolder\scoring.sas";
run;

data scored2;
    set STAT1.ameshousing4;
    %include "&homefolder\scoring.sas";
run;
```

perform needed data transformations before %INCLUDE

DATA *<data-set-name>*;
    SET *SAS-data-set <(data-set-options>*;
    %INCLUDE *source*;
RUN;



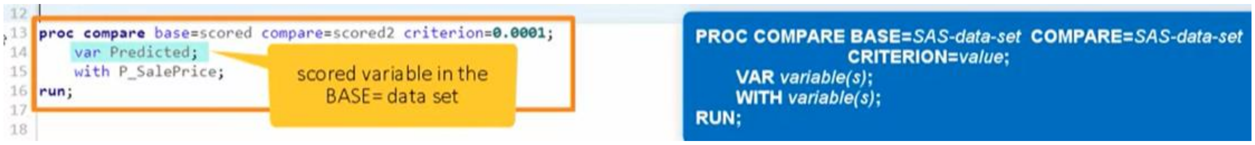
```
OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;

data scored2;
    set STAT1.ameshousing4;
    %include "&homefolder\scoring.sas";
run;

NOTE: There were 300 observations read from the data set STAT1.AMESHOUSING4.
NOTE: The data set WORK.SCORED2 has 300 observations and 33 variables.
```



```
proc compare base=scored compare=scored2 criterion=0.0001;
    var Predicted;
    with P_SalePrice;
run;
```

default criterion: .00001

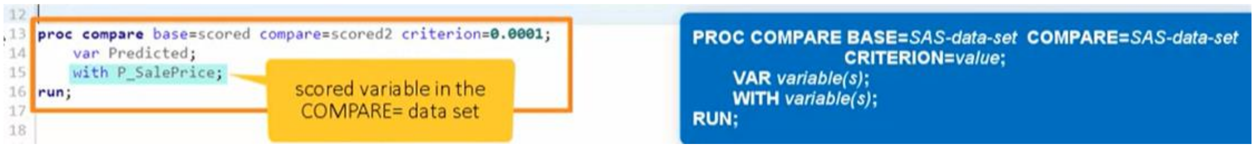
PROC COMPARE BASE=*SAS-data-set* **COMPARE=***SAS-data-set*
    CRITERION=*value*;
  VAR *variable(s)*;
  WITH *variable(s)*;
RUN;



```
proc compare base=scored compare=scored2 criterion=0.0001;
    var Predicted;
    with P_SalePrice;
run;
```

scored variable in the BASE= data set

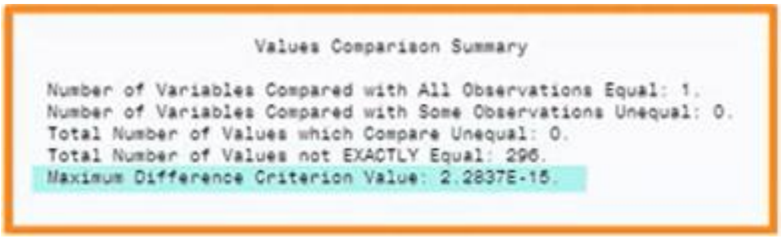
PROC COMPARE BASE=*SAS-data-set* **COMPARE=***SAS-data-set*
    CRITERION=*value*;
  VAR *variable(s)*;
  WITH *variable(s)*;
RUN;



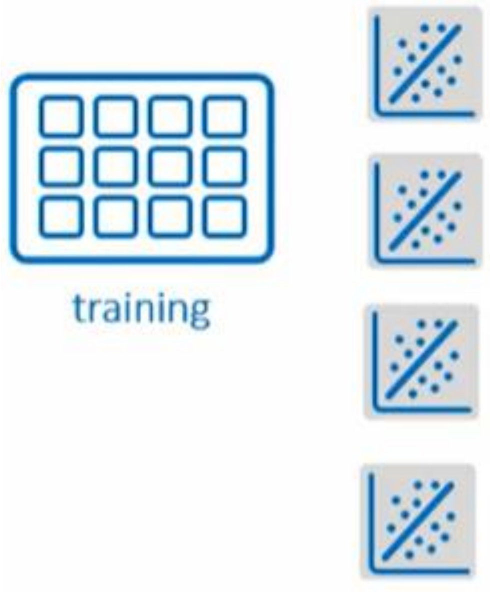
```
proc compare base=scored compare=scored2 criterion=0.0001;
    var Predicted;
    with P_SalePrice;
run;
```

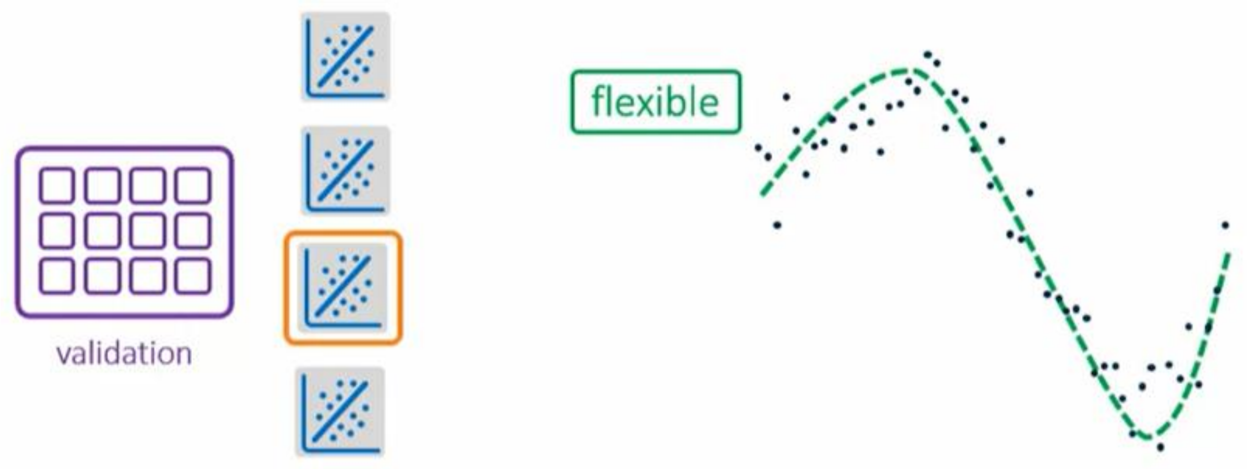
scored variable in the COMPARE= data set

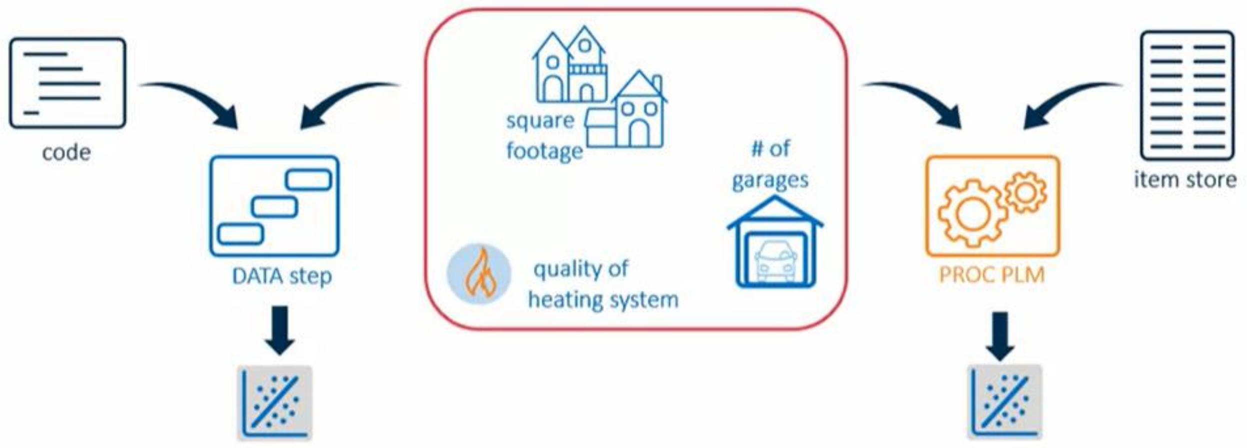
PROC COMPARE BASE=*SAS-data-set* **COMPARE=***SAS-data-set*
    CRITERION=*value*;
  VAR *variable(s)*;
  WITH *variable(s)*;
RUN;



```
                  Values Comparison Summary

Number of Variables Compared with All Observations Equal: 1.
Number of Variables Compared with Some Observations Unequal: 0.
Total Number of Values which Compare Unequal: 0.
Total Number of Values not EXACTLY Equal: 296.
Maximum Difference Criterion Value: 2.2837E-15.
```

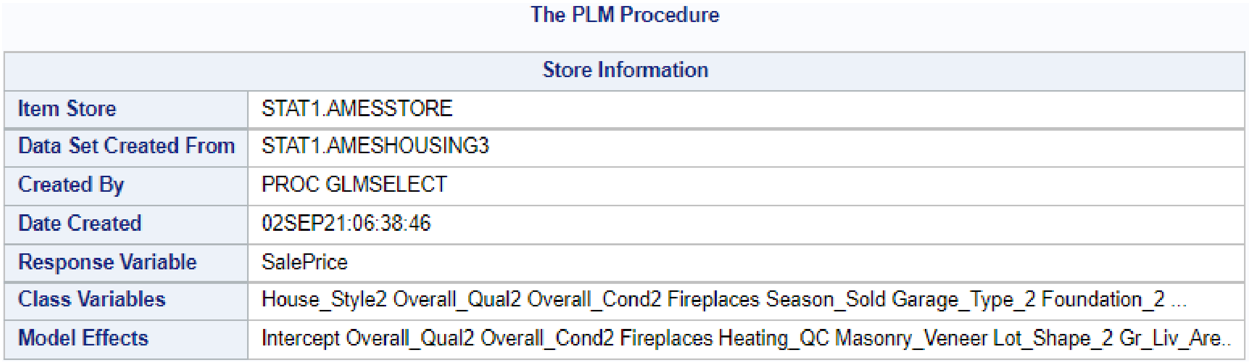

training


validation
flexible


code
DATA step
square footage
# of garages
quality of heating system
item store
PROC PLM

/*st106d02.sas*/ /*Part A*/

```
proc plm restore=STAT1.amesstore;
   score data=STAT1.ameshousing4 out=scored;
   code file="&homefolder\scoring.sas";
run;
```



The PLM Procedure

Store Information	
Item Store	STAT1.AMESSTORE
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC GLMSELECT
Date Created	02SEP21:06:38:46
Response Variable	SalePrice
Class Variables	House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces Season_Sold Garage_Type_2 Foundation_2 ...
Model Effects	Intercept Overall_Qual2 Overall_Cond2 Fireplaces Heating_QC Masonry_Veneer Lot_Shape_2 Gr_Liv_Are..

```
data scored2;
   set STAT1.ameshousing4;
   %include "&homefolder\scoring.sas";
run;
```

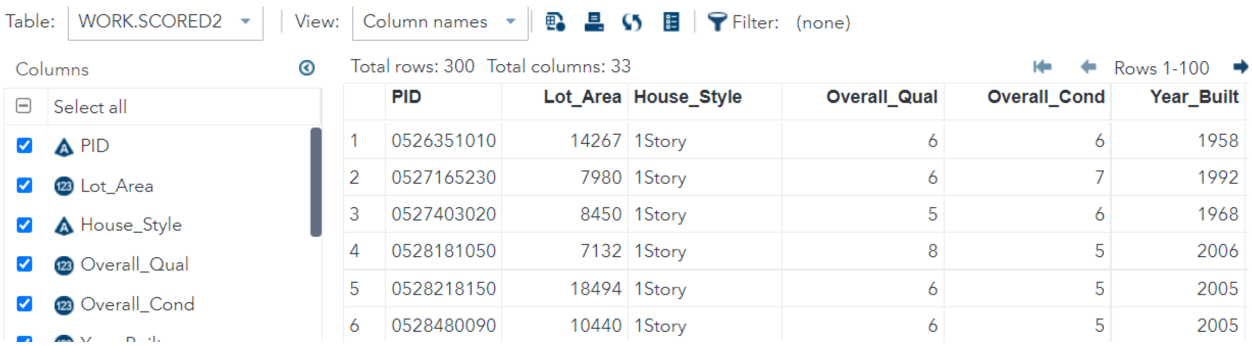


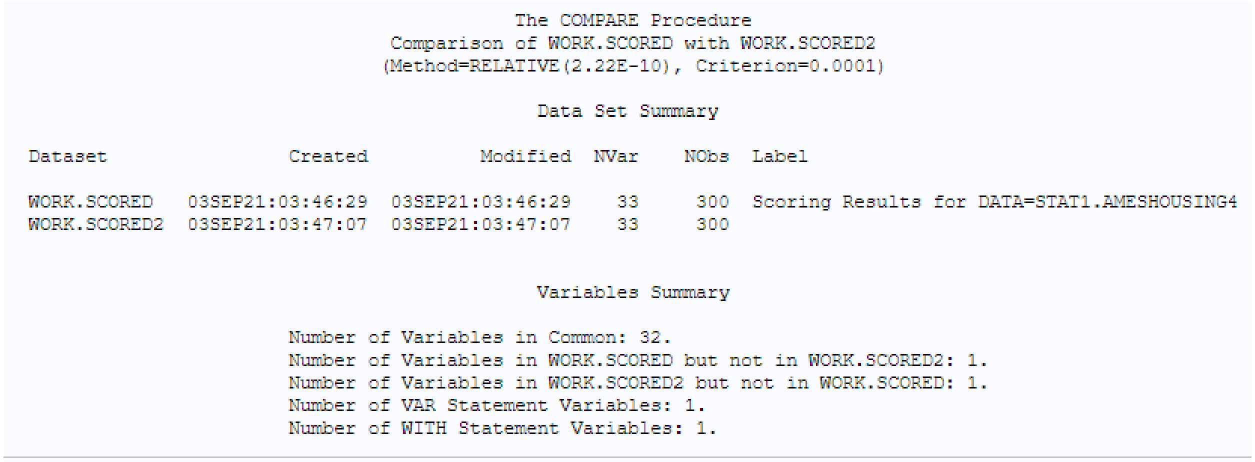
Table: WORK.SCORED2 | View: Column names | Filter: (none)

Columns
- Select all
- ☑ PID
- ☑ Lot_Area
- ☑ House_Style
- ☑ Overall_Qual
- ☑ Overall_Cond

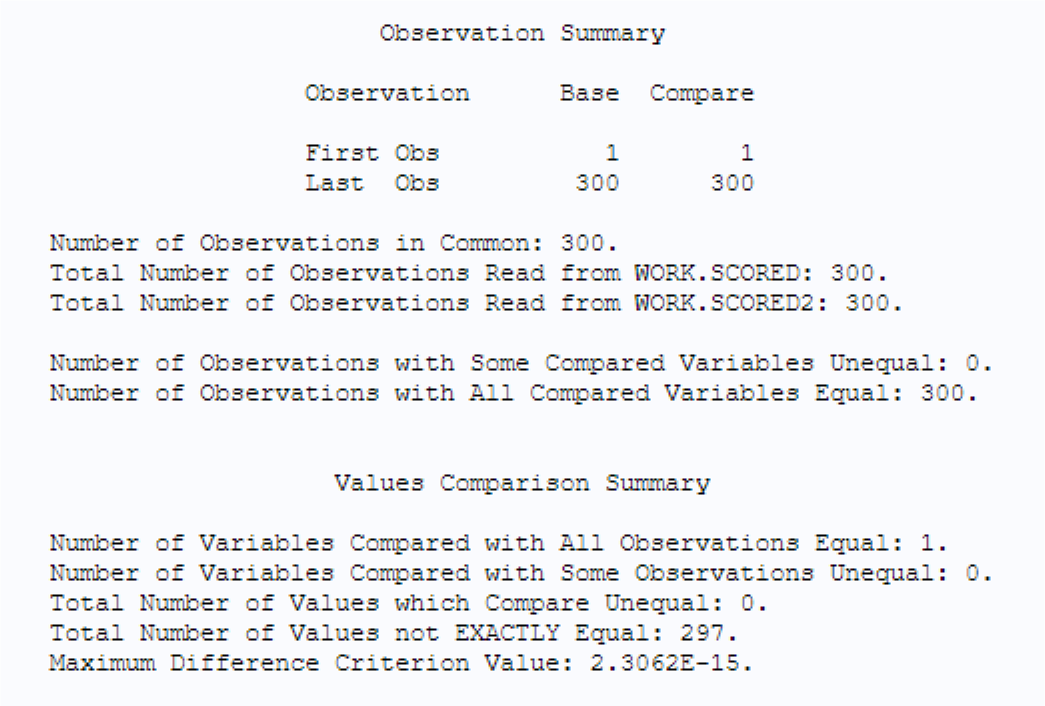
Total rows: 300 Total columns: 33 | Rows 1-100

	PID	Lot_Area	House_Style	Overall_Qual	Overall_Cond	Year_Built
1	0526351010	14267	1Story	6	6	1958
2	0527165230	7980	1Story	6	7	1992
3	0527403020	8450	1Story	5	6	1968
4	0528181050	7132	1Story	8	5	2006
5	0528218150	18494	1Story	6	5	2005
6	0528480090	10440	1Story	6	5	2005

```
proc compare base=scored compare=scored2 criterion=0.0001;
   var Predicted;
   with P_SalePrice;
run;
```



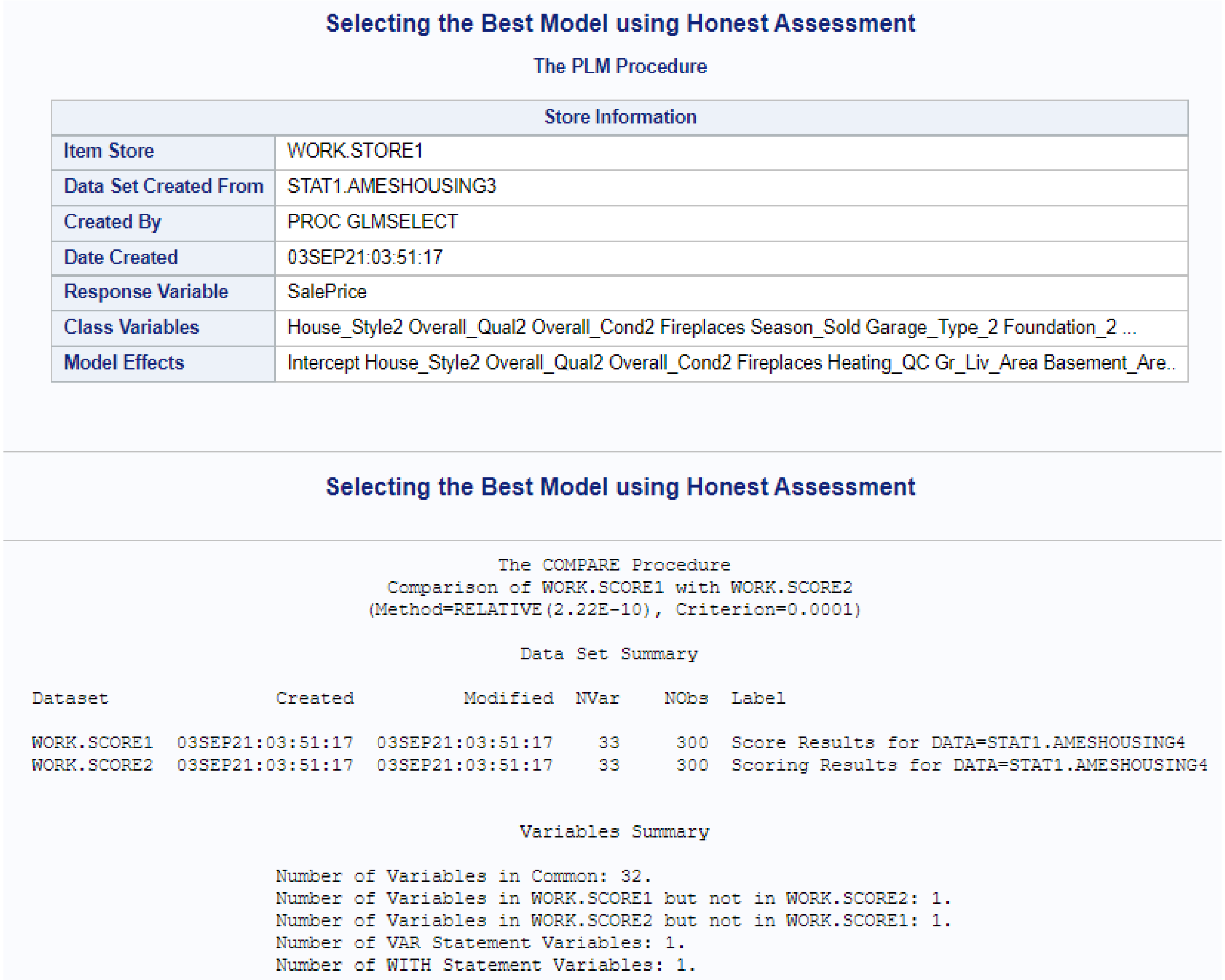
```
                                  The COMPARE Procedure
                        Comparison of WORK.SCORED with WORK.SCORED2
                       (Method=RELATIVE(2.22E-10), Criterion=0.0001)

                                    Data Set Summary

Dataset                 Created          Modified  NVar    NObs  Label

WORK.SCORED    03SEP21:03:46:29  03SEP21:03:46:29    33     300  Scoring Results for DATA=STAT1.AMESHOUSING4
WORK.SCORED2   03SEP21:03:47:07  03SEP21:03:47:07    33     300


                                    Variables Summary

                       Number of Variables in Common: 32.
                       Number of Variables in WORK.SCORED but not in WORK.SCORED2: 1.
                       Number of Variables in WORK.SCORED2 but not in WORK.SCORED: 1.
                       Number of VAR Statement Variables: 1.
                       Number of WITH Statement Variables: 1.
```



```
                         Observation Summary

                    Observation      Base  Compare

                    First Obs           1        1
                    Last  Obs         300      300

   Number of Observations in Common: 300.
   Total Number of Observations Read from WORK.SCORED: 300.
   Total Number of Observations Read from WORK.SCORED2: 300.

   Number of Observations with Some Compared Variables Unequal: 0.
   Number of Observations with All Compared Variables Equal: 300.


                      Values Comparison Summary

   Number of Variables Compared with All Observations Equal: 1.
   Number of Variables Compared with Some Observations Unequal: 0.
   Total Number of Values which Compare Unequal: 0.
   Total Number of Values not EXACTLY Equal: 297.
   Maximum Difference Criterion Value: 2.3062E-15.
```

```
/*st106s02.sas*/


proc glmselect data=STAT1.ameshousing3

          seed=8675309

          noprint;

  class &categorical / param=ref ref=first;

  model SalePrice=&categorical &interval /

          selection=stepwise

          (select=aic

          choose=validate) hierarchy=single;

  partition fraction(validate=0.3333);

  score data=STAT1.ameshousing4 out=score1;

  store out=store1;

  title "Selecting the Best Model using Honest Assessment";

run;


proc plm restore=store1;
```

```
   score data=STAT1.ameshousing4 out=score2;
run;


proc compare base=score1 compare=score2 criterion=0.0001;
   var P_SalePrice;
   with Predicted;
run;
```



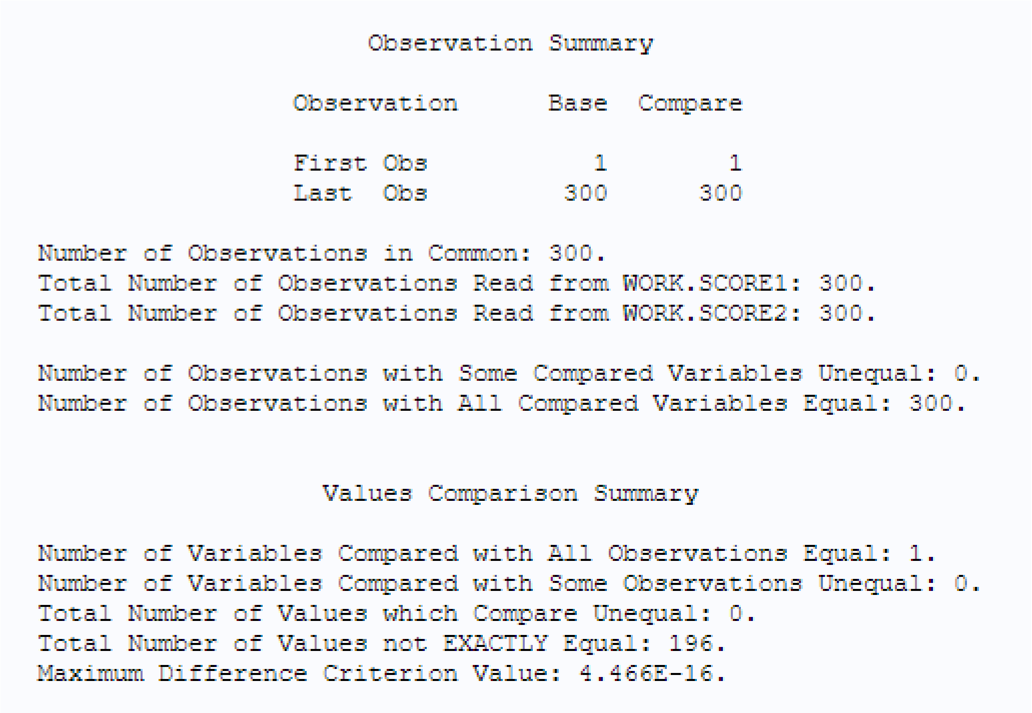
Selecting the Best Model using Honest Assessment

The PLM Procedure

Store Information	
Item Store	WORK.STORE1
Data Set Created From	STAT1.AMESHOUSING3
Created By	PROC GLMSELECT
Date Created	03SEP21:03:51:17
Response Variable	SalePrice
Class Variables	House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces Season_Sold Garage_Type_2 Foundation_2 ...
Model Effects	Intercept House_Style2 Overall_Qual2 Overall_Cond2 Fireplaces Heating_QC Gr_Liv_Area Basement_Are..

Selecting the Best Model using Honest Assessment

```
                                     The COMPARE Procedure
                          Comparison of WORK.SCORE1 with WORK.SCORE2
                         (Method=RELATIVE(2.22E-10), Criterion=0.0001)

                                       Data Set Summary

Dataset                Created          Modified  NVar    NObs  Label

WORK.SCORE1  03SEP21:03:51:17  03SEP21:03:51:17    33     300  Score Results for DATA=STAT1.AMESHOUSING4
WORK.SCORE2  03SEP21:03:51:17  03SEP21:03:51:17    33     300  Scoring Results for DATA=STAT1.AMESHOUSING4


                                       Variables Summary

                      Number of Variables in Common: 32.
                      Number of Variables in WORK.SCORE1 but not in WORK.SCORE2: 1.
                      Number of Variables in WORK.SCORE2 but not in WORK.SCORE1: 1.
                      Number of VAR Statement Variables: 1.
                      Number of WITH Statement Variables: 1.
```



```
                        Observation Summary

                   Observation      Base  Compare

                   First Obs           1        1
                   Last  Obs         300      300

Number of Observations in Common: 300.
Total Number of Observations Read from WORK.SCORE1: 300.
Total Number of Observations Read from WORK.SCORE2: 300.

Number of Observations with Some Compared Variables Unequal: 0.
Number of Observations with All Compared Variables Equal: 300.


                     Values Comparison Summary

Number of Variables Compared with All Observations Equal: 1.
Number of Variables Compared with Some Observations Unequal: 0.
Total Number of Values which Compare Unequal: 0.
Total Number of Values not EXACTLY Equal: 196.
Maximum Difference Criterion Value: 4.466E-16.
```

Practice: Scoring Using the SCORE Statement in PROC GLMSELECT

Question 1

You want to re-create the model that was built in the previous practice (based on **stat1.ameshousing3**), create an item store, and then use the item store to score the new cases in **stat1.ameshousing4**. You'll score the data in two ways (using PROC GLMSELECT and PROC PLM) and compare the results.

Open the solution program from the previous practice, **st106s01.sas**. There is no need to examine the results, so make the following changes to the code:

- Remove the PLOTS= option.
- Add the NOPRINT option to the PROC GLMSELECT statement.
- Remove the TITLE statement

Here's the modified code:

```
proc glmselect data=STAT1.ameshousing3
               seed=8675309
               noprint;
   class &categorical / param=ref ref=first;
   model SalePrice=&categorical &interval /
               selection=stepwise
               (select=aic
               choose=validate) hierarchy=single;
   partition fraction(validate=0.3333);
run;
```

In the PROC GLMSELECT step, add a STORE statement to create an item store named **store1**, and a SCORE statement to score the data in **stat1.ameshousing4**. Add a PROC PLM step that uses the item store, store1, to score the data in **stat1.ameshousing4**. **Note**: Be sure to use different names for the two scored data sets. Add a PROC COMPARE step to compare the scoring results from PROC GLMSELECT and PROC PLM. Submit the code and examine the results.

Does the PROC COMPARE output indicate any differences between the predictions produced by the two scoring methods?

The two scoring methods produce the same predictions. **Note:** Depending on the version of SAS and SAS/STAT that you are using, your results might look somewhat different from the output shown here. However, the results should indicate that these data sets do not differ.

```
/*st106s02.sas*/

proc glmselect data=STAT1.ameshousing3
               seed=8675309
               noprint;
   class &categorical / param=ref ref=first;
   model SalePrice=&categorical &interval /
               selection=stepwise
               (select=aic
               choose=validate) hierarchy=single;
   partition fraction(validate=0.3333);
   score data=STAT1.ameshousing4 out=score1;
   store out=store1;
   title "Selecting the Best Model using Honest Assessment";
run;

proc plm restore=store1;
   score data=STAT1.ameshousing4 out=score2;
run;

proc compare base=score1 compare=score2 criterion=0.0001;
   var P_SalePrice;
   with Predicted;
run;
```