Getting Started with Building Regression Models for Prediction

Robert N. Rodriguez
SAS Institute (retired)

Michigan SAS Users Group
January 18, 2024

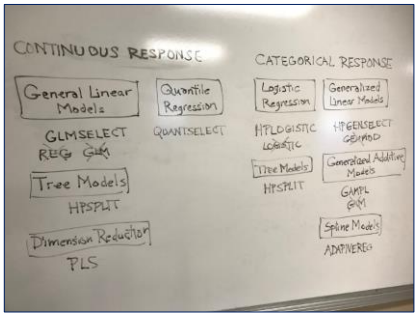
1

An unforgettable conversation:
"I'm starting to use regression models for prediction …"

Which variables should I include in my model?

Which procedure do I use?

What's the difference between all these procedures?

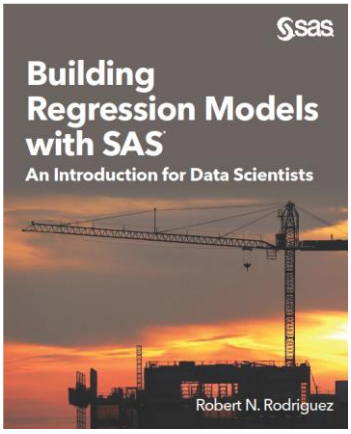


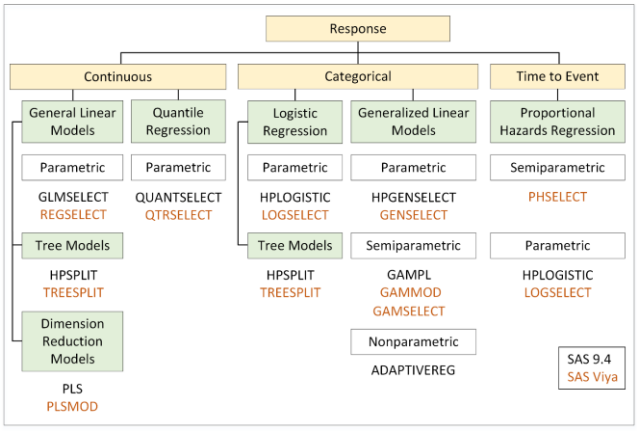
My whiteboard explanation

2

2

Someone ought to write a book about this …





The diagram that began on the whiteboard

3

3

This presentation will be a balloon ride over a vast landscape

SAS users on a balloon safari near Magaliesburg, South Africa

4

4

Our itinerary

- Concepts
- Example 1: Building a general linear model
- Example 2: Building a generalized linear model
- More about the book

5

5

Preliminary Concepts

6

6

Regression models are viewed differently in the fields of statistics and machine learning

Model form

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \ , \qquad i = 1, \ldots, n$$

In statistics, the predictors x_j are planned;
the focus is on the parameters β_j

In machine learning, algorithms choose the variables from a database;
the focus is prediction of the response *y*

7

7

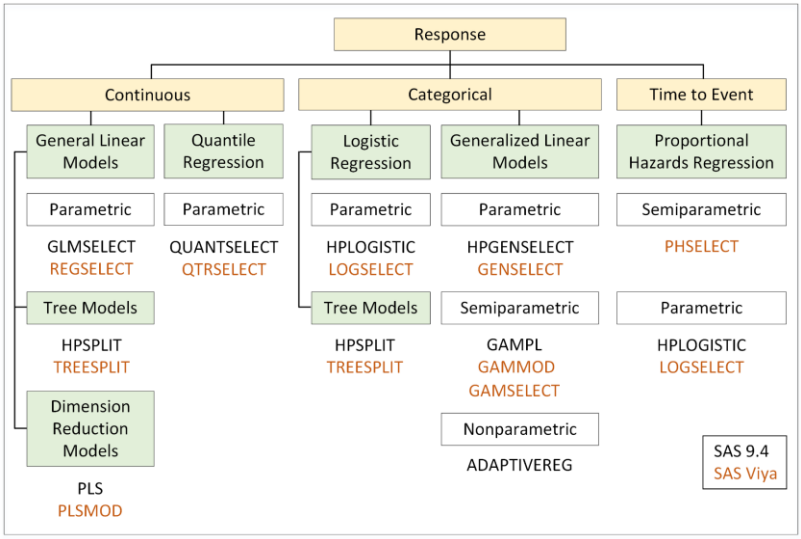
Data science spans a major divide between statistics and machine learning

Models	Statistics	Machine Learning
Goal	Inference, understanding	Predictive accuracy
Data	Planned, observational	Large databases
Meaning	Simplification of reality	Representation of reality
Creation	Specified from knowledge	Learned from data
Form	Interpretable	Black box
Validity	Theory and assumptions	Performance on new data
Concern	Biased inference	Biased data

8

8

Regression models might not achieve the predictive accuracy of supervised learning, but they are versatile and interpretable



9

9

"Model building" includes more than traditional model selection; it is the use of algorithms to learn the form of the model from the data

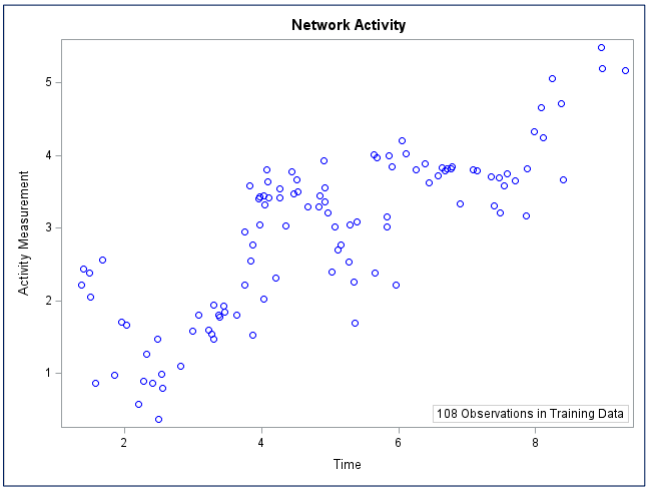
Methods for Building General Linear Models

Best Subset	Sequential Selection	Shrinkage	Dimension Reduction	Model Averaging
	Forward selection Forward swap selection Backward elimination Stepwise selection	Ridge regression Lasso Least angle regression Group lasso Elastic net	Variable clustering Principal components regression Partial least squares regression	Bootstrap Akaike weights

10

10

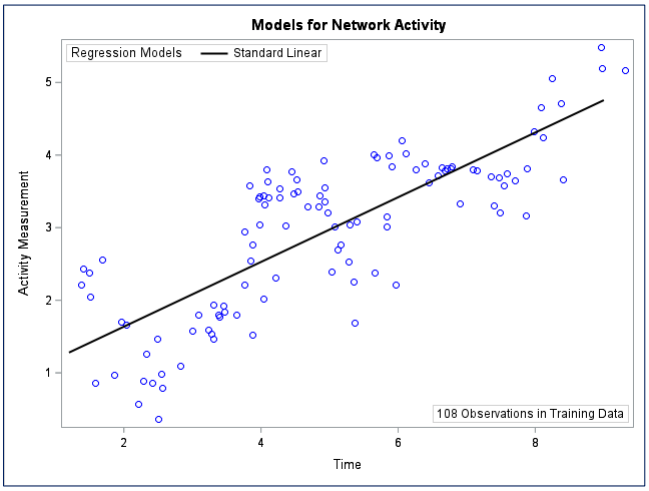
A good predictive model must make a trade-off between bias (poor fit to data structure) and variance (too much wiggliness)



11

11

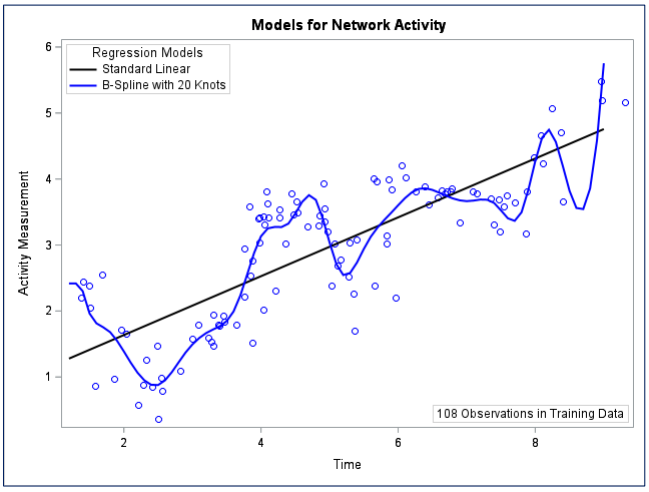
A linear regression model is too smooth to capture the oscillations (low variance but high bias), so it generalizes poorly to new data



12

12

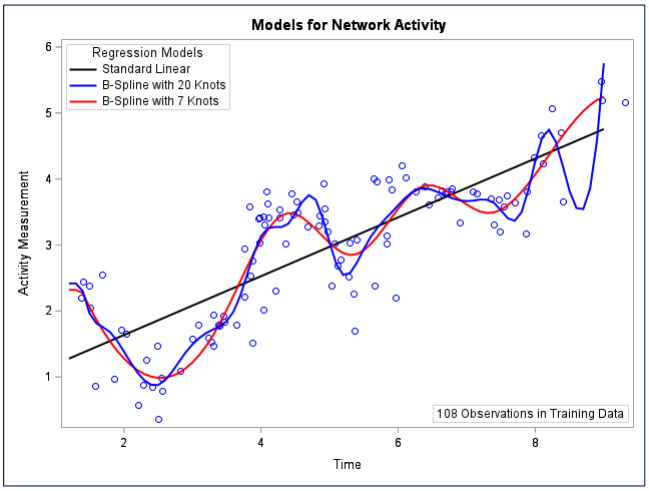
A spline model with 20 knots also predicts poorly because its wiggliness accommodates noise (high variance but low bias)



13

13

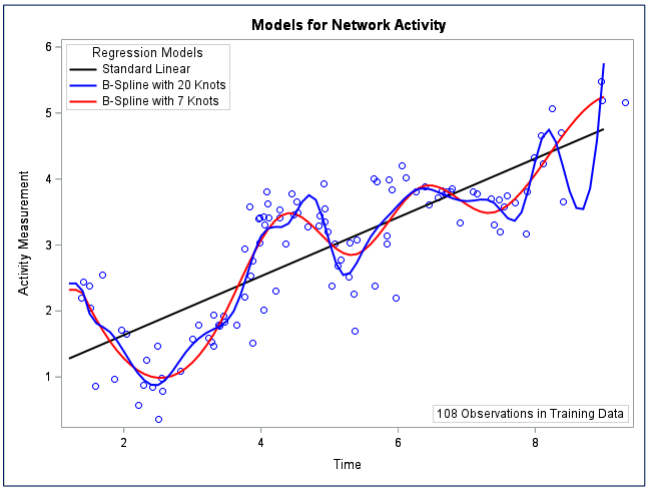
A spline model with seven knots (in red) predicts well because it makes a good trade-off between bias and variance



14

14

To achieve a good trade-off, algorithms are used to find the right number of knots or effects for the model (the tuning parameter)



15

15

Example 1: Building a General Linear Model to Predict the Close Rate of a Store

16

16

The close rate for a retail store is the percent of people who enter and make a purchase

Predicting the close rate of new stores helps to assess profitability

Understanding which variables are associated with close rate is critical for growing the business

17

17

Data for 35 variables and 500 stores are available for building a predictive model

Variable	Description	Levels of Categorical Predictors
CloseRate	Response	
Region	Geographic region	East, West, South, Midwest
Training	Training status	None, In Progress, Complete
X1 ... X20	Store characteristics	
P1 ... P6	Promotional activities	
L1 ... L6	Special layouts	

18

18

A general linear model is appropriate for predicting close rate because it can include both classification effects and continuous variables

GLM Procedure	GLMSELECT Procedure
Fits and analyzes models	Fits and builds models
Handles moderate-to-large data	Handles large-to-massive data
Designed for inference	Designed for prediction

The REG procedure does not handle classification effects, and its features for model selection are outmoded

19

19

In the GLMSELECT procedure, you specify the candidate variables with the MODEL and CLASS statements

```
proc glmselect plots=coefficients data=Stores;
   class Region(ref='Midwest' split) Training(ref='None');
   model CloseRate = Region Training X1-X20 L1-L6 P1-P6 /
      selection=forward(choose=sbc select=sbc stop=sbc);
run;
```

Without the SELECTION= option, the procedure would fit the model using all the variables

20

20

Forward selection with the Schwarz Bayesian criterion (SBC) is one of many methods available in the GLMSELECT procedure

```
proc glmselect plots=coefficients data=Stores;
   class Region(ref='Midwest' split) Training(ref='None');
   model CloseRate = Region Training X1-X20 L1-L6 P1-P6 /
      selection=forward(choose=sbc select=sbc stop=sbc);
run;
```

Forward selection starts with the intercept and adds a predictor at each step

21

21

CHOOSE=SBC specifies that the model chosen is the one that minimizes SBC

```
proc glmselect plots=coefficients data=Stores;
   class Region(ref='Midwest' split) Training(ref='None');
   model CloseRate = Region Training X1-X20 L1-L6 P1-P6 /
      selection=forward(choose=sbc select=sbc stop=sbc);
run;
```

Minimizing SBC tries to avoid overfitting by penalizing the complexity of the model (the number of predictors p)

$$SBC = n \log(\text{training average squared error}) + p \log(n)$$

22

22

SELECT=SBC specifies that the predictor added at each step is the one that most decreases SBC

```
proc glmselect plots=coefficients data=Stores;
   class Region(ref='Midwest' split) Training(ref='None');
   model CloseRate = Region Training X1-X20 L1-L6 P1-P6 /
      selection=forward(choose=sbc select=sbc stop=sbc);
run;
```

23

23

STOP=SBC specifies that selection stops when the predictor that would be added next yields a larger SBC

```
proc glmselect plots=coefficients data=Stores;
   class Region(ref='Midwest' split) Training(ref='None');
   model CloseRate = Region Training X1-X20 L1-L6 P1-P6 /
      selection=forward(choose=sbc select=sbc stop=sbc);
run;
```

24

24

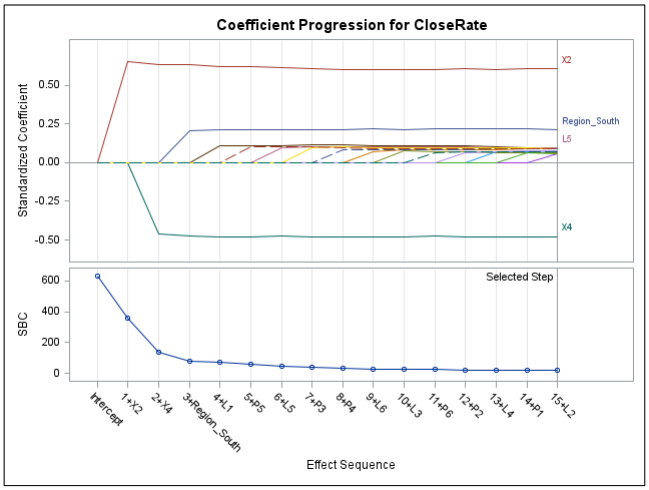
The forward method selects X2, X4, Region_South, and all the layout and promotional variables

Forward Selection Summary				
Step	Effect Entered	Number Effects In	Number Parms In	SBC
0	Intercept	1	1	631.5807
1	X2	2	2	360.0538
2	X4	3	3	137.2902
3	Region_South	4	4	81.3798
4	L1	5	5	69.1589
5	P5	6	6	58.8027
6	L5	7	7	48.5608
7	P3	8	8	39.3333
8	P4	9	9	32.8399
9	L6	10	10	29.2646
10	L3	11	11	24.8623
11	P6	12	12	23.4529
12	P2	13	13	21.2727
13	L4	14	14	18.2626
14	P1	15	15	17.1062
15	L2	16	16	16.8729*
* Optimal Value of Criterion				

25

25

SBC decreases slowly after the first three effects are added, revealing that simpler (sparser) models compete with the chosen model



26

26

By default, the table of parameter estimates does not show *p*-values because they are not reliable

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	59.563030	0.181698	327.81
Region_South	1	0.852461	0.091167	9.35
X2	1	3.959738	0.150641	26.29
X4	1	-3.015353	0.142871	-21.11
L1	1	0.629862	0.149481	4.21
L2	1	0.366923	0.146395	2.51
L3	1	0.412324	0.146743	2.81
L4	1	0.469072	0.144829	3.24
L5	1	0.602895	0.141213	4.27
L6	1	0.530262	0.139327	3.81
P1	1	0.392566	0.149940	2.62
P2	1	0.485213	0.144584	3.36
P3	1	0.592262	0.146886	4.03
P4	1	0.602274	0.148382	4.06
P5	1	0.574614	0.146799	3.91
P6	1	0.462066	0.147162	3.14

27

27

The SCORE statement uses the parameter estimates to compute predicted values of CloseRate for new data

```
proc glmselect plots=coefficients data=Stores;
   class Region(ref='Midwest' split) Training(ref='None');
   model CloseRate = Region Training X1-X20 L1-L6 P1-P6 /
      selection=forward(choose=sbc select=sbc stop=sbc);
   score data=NewStores out=CloseRateScores;
run;
```

StoreID	X1	X2	X3	L1	L2	P1	P2	Region	Training	p_CloseRate
601	-0.340	-0.268	0.376	0.451	0.633	0.094	-0.122	East	None	58.3912
602	0.466	0.225	-0.069	0.021	0.093	0.496	0.107	East	InProgress	60.5080
603	0.436	0.469	-0.250	0.193	0.557	-0.225	-0.316	East	None	62.1660

Partial listing of CloseRateScores

28

28

With enough data, you can stop the selection based on average squared error (ASE) computed from observations set aside for validation

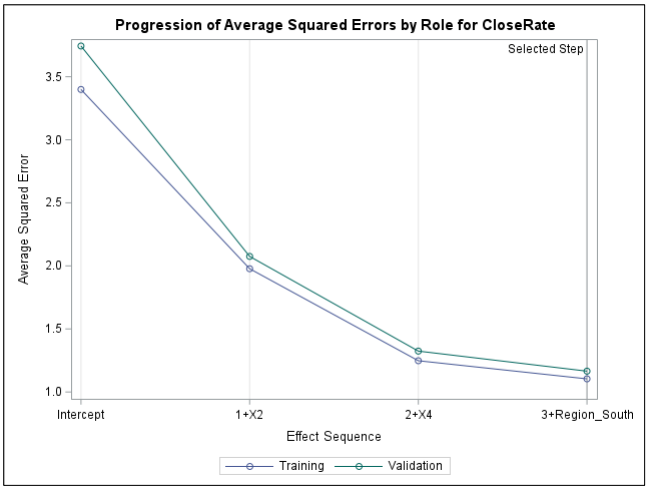
```
proc glmselect plot=aseplot data=Stores seed=13551;
   class Region(ref='Midwest' split) Training(ref='None');
   model CloseRate = Region Training X1-X20 L1-L6 P1-P6 /
      selection=forward(choose=sbc select=sbc stop=validate);
   partition fraction(validate=0.3);
run;
```

Validation ASE is an honest measure of prediction error because it is computed from data not used to determine the model

29

29

Validation ASE is minimized by the first three effects (sparse model), whereas training ASE continues to decrease with additional effects



30

30

The lasso method overcomes various limitations of sequential selection by penalizing the least squares estimates of the parameters

$$\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq t$

- Selects models by shrinking some coefficients, setting others to zero
- Produces sparser, more interpretable models
- Applies to wide data ($p > n$)
- Involves choosing t (validation, cross validation, bootstrap)

31

31

The lasso method with five-fold cross validation selects the sparse model with effects X2, X4, and Region_South

```
proc glmselect data=Stores seed=59915;
   class Region(ref='Midwest') Training(ref='None');
   model CloseRate = Region Training X1-X20 L1-L6 P1-P6 /
      cvmethod=random(5)
      selection=lasso(choose=cvex maxsteps=15);
run;
```

32

32

Example 2: Building a Generalized Linear Model to Predict Claim Frequency

33

33

Poisson regression models—one class of generalized linear models—are used to predict the frequency of insurance claims

22 variables for 20,361 auto insurance policyholders

Variable	Description	Levels of Categorical Variables
NumberClaims	Response	
Exposure	Length of time	
AgePolicyHolder	Age	
AreaType	Type of area	`City, Rural, Town, Village`
<more variables>		
WorkStatus	Employment	`Full Time, Not Working, ...`

34

34

The HPGENSELECT is the appropriate procedure for building a generalized linear model

GENMOD Procedure	HPGENSELECT Procedure
Fits models	Fits and builds models
Handles moderate-to-large data	Handles large-to-massive data
Designed for inference	Designed for prediction

35

35

You specify the Poisson distribution, log link function, and candidate effects with the MODEL and CLASS statements

```
proc hpgenselect data=ClaimFrequency fmtlibxml=ClmsFmts;
   class &ClassVars / param=ref;
   model NumberClaims = &AllCandidates /
      distribution=poisson link=log offset=logExposure;
   selection method=forward(choose=validate stop=sbc);
   partition rolevar=Role(train='1' validate='2' test='3');
run;
```

36

36

The SELECTION statement provides various methods and criteria, including ASE based on validation data

```
proc hpgenselect data=ClaimFrequency fmtlibxml=ClmsFmts;
   class &ClassVars / param=ref;
   model NumberClaims = &AllCandidates /
      distribution=poisson link=log offset=logExposure;
   selection method=forward(choose=validate stop=sbc);
   partition rolevar=Role(train='1' validate='2' test='3');
run;
```

37

37

The ROLEVAR= variable designates observations for training, validation, and testing

```
proc hpgenselect data=ClaimFrequency fmtlibxml=ClmsFmts;
   class &ClassVars / param=ref;
   model NumberClaims = &AllCandidates /
      distribution=poisson link=log offset=logExposure;
   selection method=forward(choose=validate stop=sbc);
   partition rolevar=Role(train='1' validate='2' test='3');
run;
```

The values of ROLE were pre-assigned with the SURVEYSELECT procedure

38

38

Selection stopped at Step 4 when SBC was minimized, but validation ASE was minimized at Step 6 when AreaType entered

Selection Summary				
Step	Effect Entered	Number Effects In	SBC	Validation ASE
0	Intercept	1	27938.7784	1.5709
1	NYoungDrvrs	2	27910.1597	1.5633
2	VehicleOwned	3	27892.5942	1.5610
3	Deductible	4	27891.2697	1.5579
4	StdHouseIncome	5	27884.8648*	1.5545
5	NumberNoFault	6	27885.5394	1.5529
6	AreaType	7	27903.3595	1.5502*

Minimum SBC (stop selection)

Minimum ASE (choose model)

The procedure used a stop horizon of three steps

39

39

The CODE statement creates and saves DATA step code for scoring new data

```
proc hpgenselect data=ClaimFrequency fmtlibxml=ClmsFmts;
   class &ClassVars / param=ref;
   model NumberClaims = &AllCandidates /
      distribution=poisson link=log offset=logExposure;
   selection method=forward(choose=validate stop=sbc);
   partition rolevar=Role(train='1' validate='2' test='3');
   code file='ClaimFreqScore.sas';
run;
```

40

40

You can include the scoring code in a DATA step to predict the number of claims for new policyholders

```
data ClaimPrediction;
   set NewPolicyHolderData;
   %inc 'ClaimFreqScore.sas';
run;
```

AgePolicyHolder	AreaType	Citations	CreditScore	Deductible	Education	WorkStatus	P_NumberClaims
66	City	None	800 to 850	High	Associate	Full Time	0.529
49	City	None	750 to 799	None	Master	Full Time	0.164
48	City	None	800 to 850	None	Some college	Not Working	1.227

Partial listing of ClaimPrediction

41

41

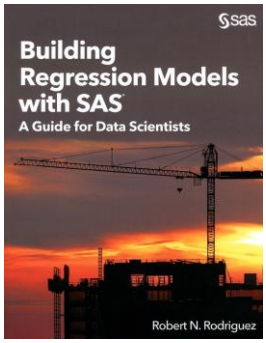
What You Will Find in the Book

42

42

The book is designed for readers who are unfamiliar with the models and the concepts of predictive modeling

- ✓ Introduction to theory and methods
- ✓ Examples (data and code on book website)
- ✓ Macros that simplify tasks
- ✓ Tables comparing methods and features
- ✓ Road signs for common misconceptions
- ✓ Problems with model building (Chapter 3)

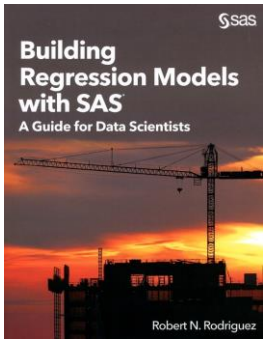


43

43

The book also includes specialized topics for readers who are familiar with the basics and want to learn more

- ✓ Selection bias
- ✓ Spline effects
- ✓ Multicollinearity
- ✓ Scoring methods
- ✓ Computational algorithms
- ✓ Information criteria
- ✓ Notes about leading personalities



44

44

Colin Mallows (1930-2023) introduced C_p, a measure of prediction error, during the early 1970s

Intended C_p for finding a unique set of predictors or multiple sets that do well

Criticized algorithms that automate model selection by minimizing C_p, SBC, and AIC, describing them as "blind … they don't look at the data." Suggested model averaging (Chapter 7)

Advocated considering *what data are needed for the problem* and examining the data *before* making assumptions

Colin and his wife Jean
IMS Bulletin, January 2024

45

45

Time to Land and Wrap Up

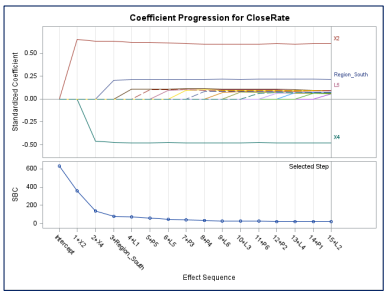
46

46

With so many methods for building regression models, which one should you use?

- No one method stands above the rest, so learn how they work and try different methods
- Explore sensitivity to tuning parameters
- Identify competing models with diagnostic plots

No method will succeed if you do not understand what your variables are measuring!



Newer SAS/STAT procedures that build regression models provide three major benefits

- Versatility for handling different types of responses and effects
- Modern algorithms for good predictive performance
- Interpretable models that you can understand and explain