Data modeling: CSCI E-106

Applied Linear Statistical Models

Chapter 9 – Building the Regression Model I: Model Selection and Validation

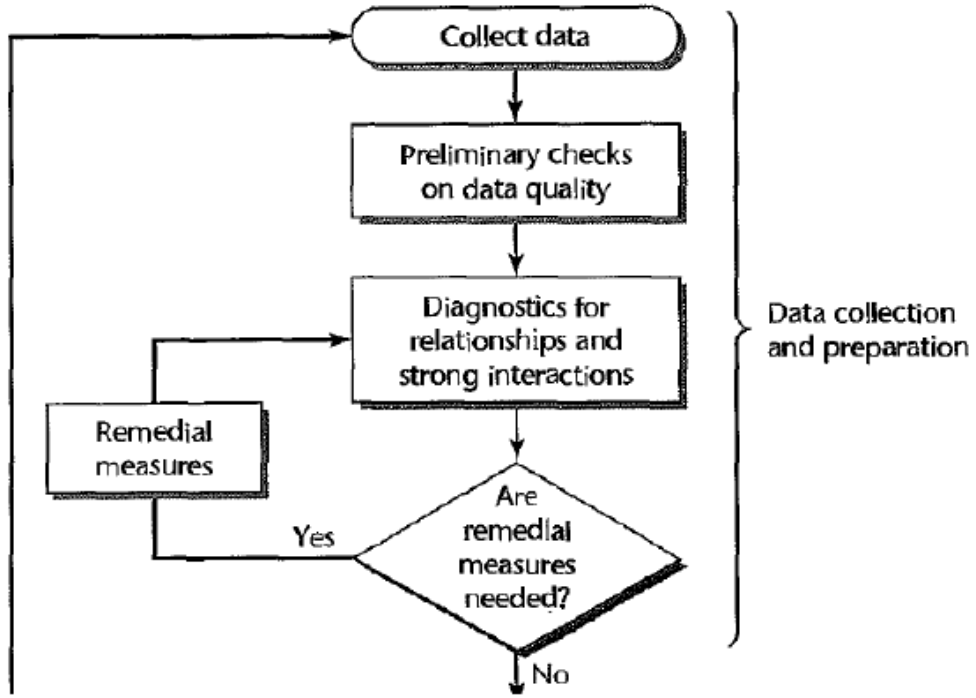
Overview of Model Building Process

- This strategy involves three or, sometimes, four phases:
  - Data collection and preparation
  - Reduction of explanatory or predictor variables (for exploratory observational studies)
  - Model refinement and selection
  - Model validation

Overview of Model Building Process, cont'd

Strategy for building a regression model

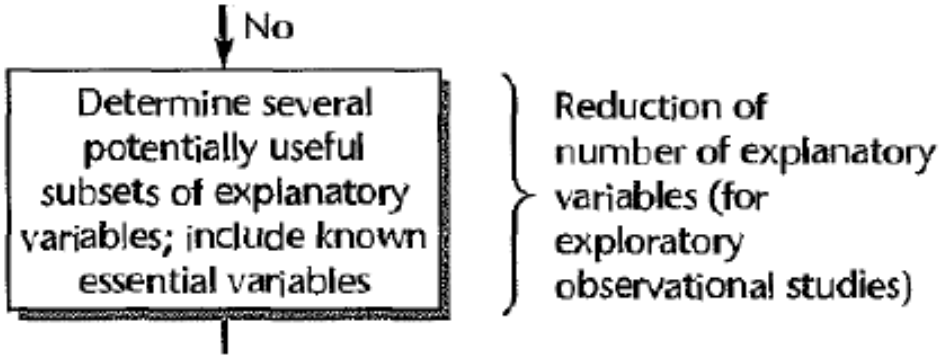
1. Data collection and preparation



Overview of Model Building Process, cont'd

Strategy for building a regression model

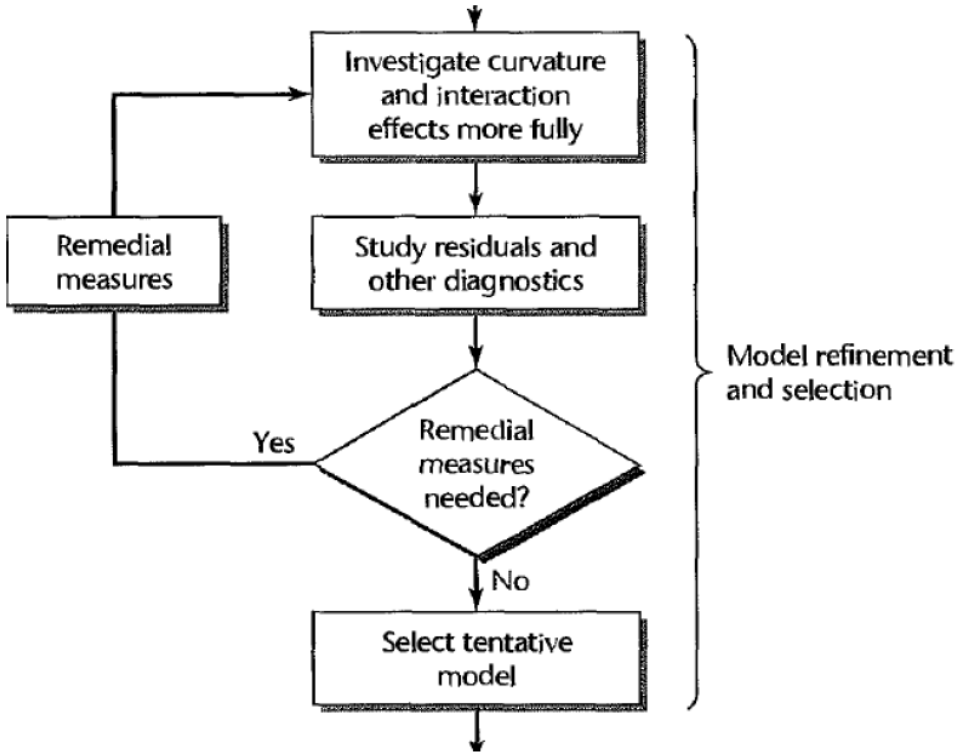
1. Data collection and preparation
2. Reduction of number of explanatory variables



Overview of Model Building Process, cont'd

Strategy for building a regression model

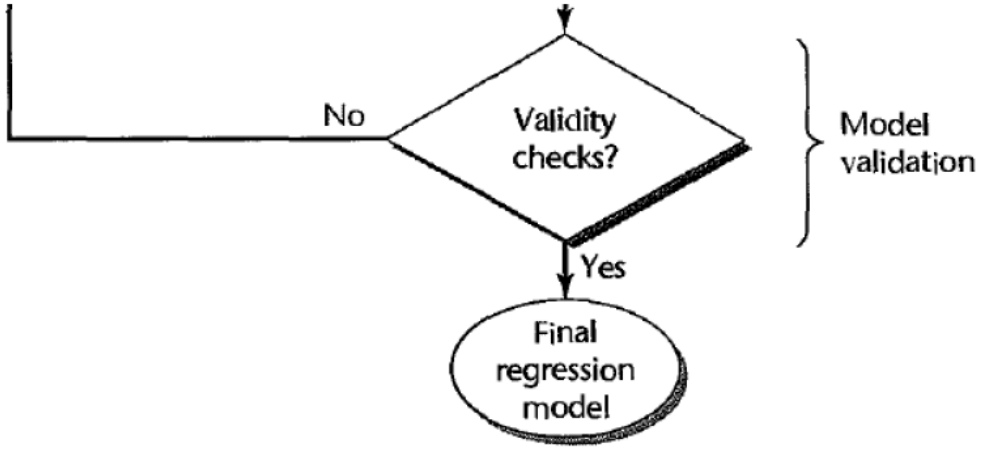
1. Data collection and preparation
2. Reduction of number of explanatory variables
3. Model refinement and selection



Overview of Model Building Process, cont'd

Strategy for building a regression model

1. Data collection and preparation
2. Reduction of number of explanatory variables
3. Model refinement and selection
4. Model Validation



Surgical Unit Example

A hospital surgical unit was interested in predicting survival in patients undergoing a particular type of liver operation. 108 patients was available for analysis. Variables:

- X_1: blood clotting score
- X_2: prognostic index
- X_3: enzyme function test score
- X_4: liver function test score
- X_5: age, in years
- X_6: indicator variable for gender (0 = male; 1 = female)
- X_7 and X_8: indicator variables for history of alcohol use:

Alcohol Use	X_7	X_8
None	0	0
Moderate	1	0
Severe	0	1

Surgical Unit Example, cont'd

- Y: survival time
- Original data: 108 patients
- Preliminary study: the first 54 patients with the first four Variables

TABLE 9.1 Potential Predictor Variables and Response Variable—Surgical Unit Example.

Case Number i	Blood-Clotting Score X_{i1}	Prognostic Index X_{i2}	Enzyme Test X_{i3}	Liver Test X_{i4}	Age X_{i5}	Gender X_{i6}	Alc. Use: Mod. X_{i7}	Alc. Use: Heavy X_{i8}	Survival Time Y_i	$Y_i' = \ln Y_i$
1	6.7	62	81	2.59	50	0	1	0	695	6.544
2	5.1	59	66	1.70	39	0	0	0	403	5.999
3	7.4	57	83	2.16	55	0	0	0	710	6.565
...	...	...	...	...	...	...	...	...	...	...
52	6.4	85	40	1.21	58	0	0	1	579	6.361
53	6.4	59	85	2.33	63	0	1	0	550	6.310
54	8.8	78	72	3.20	56	0	0	0	651	6.478

Surgical Unit Example, cont'd

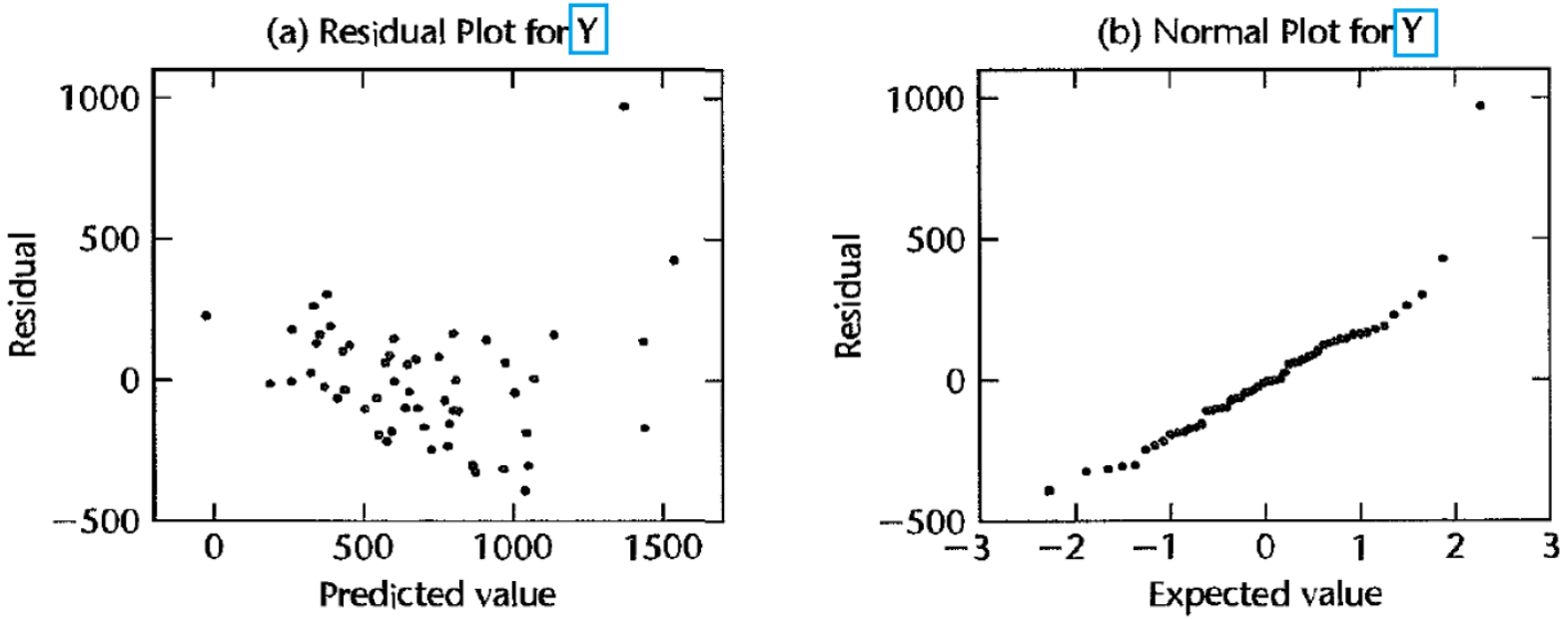
A first-order regression model based on all predictor variables was fitted to serve as a starting point:

- basic plots/correlation diagnosis
- several cases as outlying
- examine the influence of these cases

Surgical Unit Example, cont'd

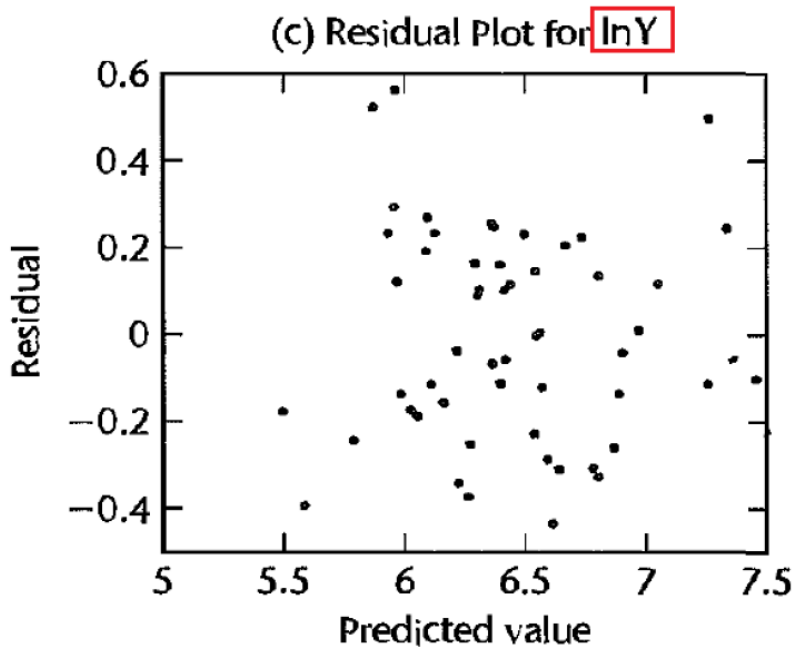
Residual plots:

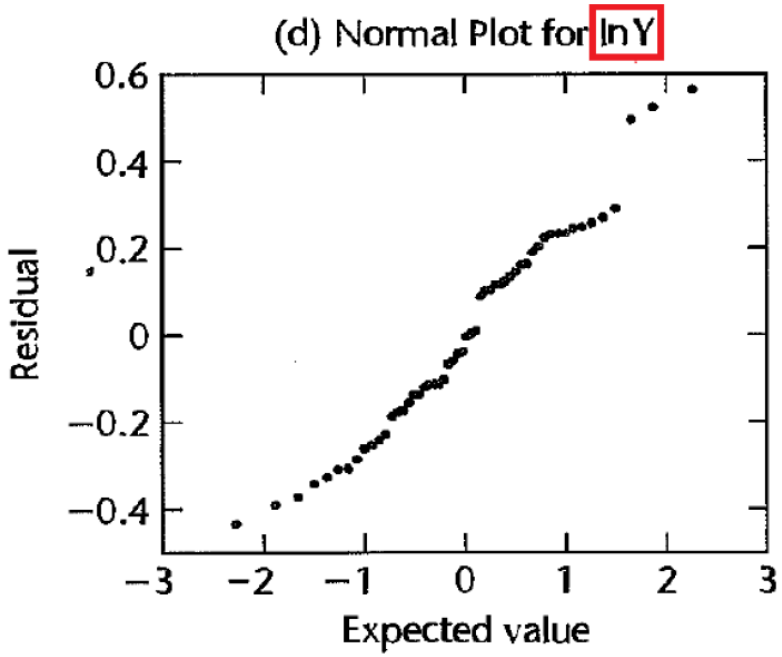
- several cases as outlying
- curvature and nonconstant error variance
- some departure from normality



Surgical Unit Example, cont'd

- To make the distribution of the error terms more nearly normal to see if the same transformation would also reduce the apparent curvature
- $Y' = \ln Y$





Surgical Unit Example, cont'd

Multivariate Correlations

	LnSurvival	Bloodclot	Progindex	Enzyme	Liver
LnSurvival	1.0000	0.2462	0.4699	0.6539	0.6493
Bloodclot	0.2462	1.0000	0.0901	−0.1496	0.5024
Progindex	0.4699	0.0901	1.0000	−0.0236	0.3690
Enzyme	0.6539	−0.1496	−0.0236	1.0000	0.4164
Liver	0.6493	0.5024	0.3690	0.4164	1.0000

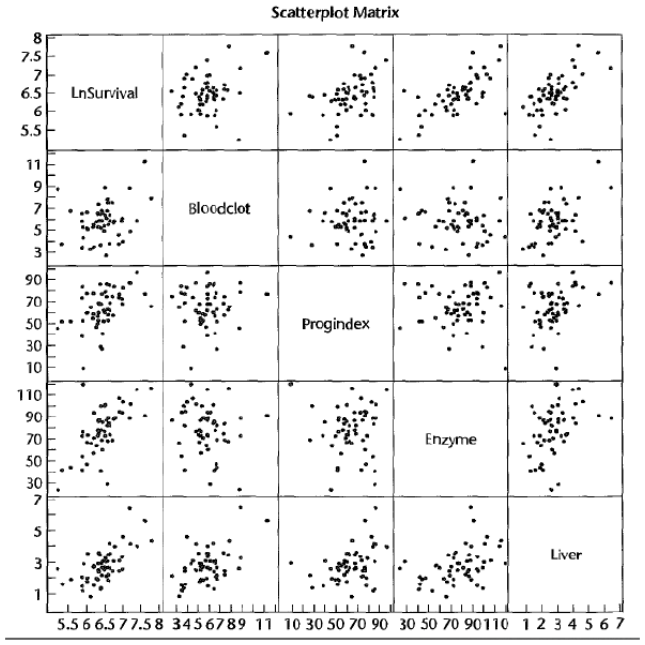


Figure : JMP Scatter Plot Matrix and Correlation Matrix when Response Variable Is Y'-Surgical Unit Example

Surgical Unit Example, cont'd

- linearly associated with Y', with X_3, X_4 showing the highest degrees of association and X_1 the lowest
- intercorrelations among the potential predictor variables (X_4 vs. X_1, X_2, X_3)
- Basic of the analyses: present the predictor variables in linear terms; not to include any interaction terms
- To examine whether all of the potential variables are needed or whether a subset of them is adequate

$\Rightarrow$ model selection

Model Selection

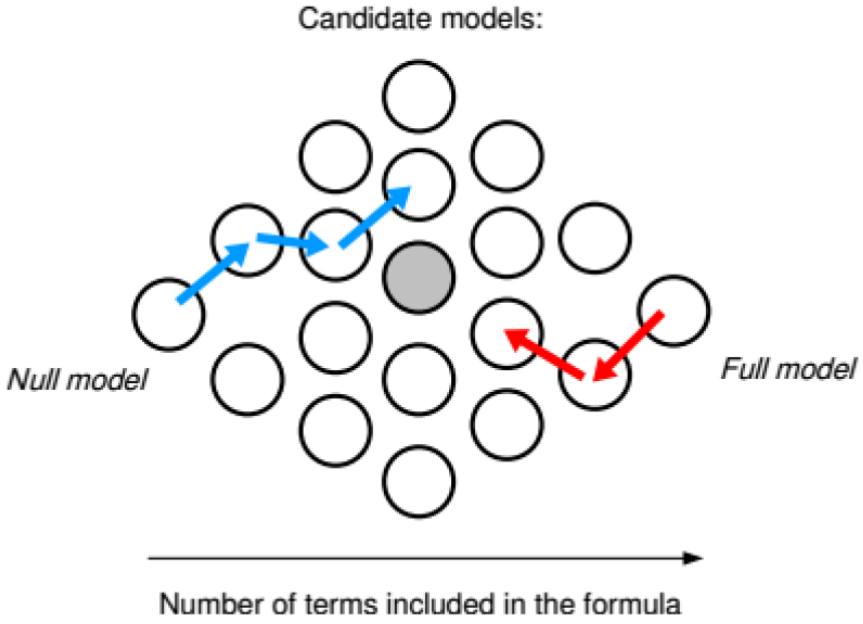


Figure 1: A schematic representation of the candidate set of models for a hypothetical "full model". There is one null model (left), one full model (right), and a range of models with some terms but not all of them (in between). Arrows represent stepwise model selection procedures: backward (red) and forward (blue). In this case both approaches would not converge to the same model, and it can be that none of them converges to the actual model that is optimal in terms of the IC used (grey circle).

- Reference: http://www.jstatsoft.org/v34/i12/paper

Criteria for Model Selection

- From any set of $p-1$ predictors: 2^{p-1} alternative models can be constructed
- There is the regression model with no X variables.

TABLE 9.2 SSE_p, R_p^2, $R_{a,p}^2$, C_p, AIC_p, SBC_p, and $PRESS_p$ Values for All Possible Regression Models—Surgical Unit Example.

X Variables in Model	(1) p	(2) SSE_p	(3) R_p^2	(4) $R_{a,p}^2$	(5) C_p	(6) AIC_p	(7) SBC_p	(8) $PRESS_p$
None	1	12.808	0.000	0.000	151.498	−75.703	−73.714	13.296
X_1	2	12.031	0.061	0.043	141.164	−77.079	−73.101	13.512
X_2	2	9.979	0.221	0.206	108.556	−87.178	−83.200	10.744
X_3	2	7.332	0.428	0.417	66.489	−103.827	−99.849	8.327
X_4	2	7.409	0.422	0.410	67.715	−103.262	−99.284	8.025
X_1, X_2	3	9.443	0.263	0.234	102.031	−88.162	−82.195	11.062
X_1, X_3	3	5.781	0.549	0.531	43.852	−114.658	−108.691	6.988
X_1, X_4	3	7.299	0.430	0.408	67.972	−102.067	−96.100	8.472
X_2, X_3	3	4.312	0.663	0.650	20.520	−130.483	−124.516	5.065
X_2, X_4	3	6.622	0.483	0.463	57.215	−107.324	−101.357	7.476
X_3, X_4	3	5.130	0.599	0.584	33.504	−121.113	−115.146	6.121
X_1, X_2, X_3	4	3.109	0.757	0.743	3.391	−146.161	−138.205	3.914
X_1, X_2, X_4	4	6.570	0.487	0.456	58.392	−105.748	−97.792	7.903
X_1, X_3, X_4	4	4.968	0.612	0.589	32.932	−120.844	−112.888	6.207
X_2, X_3, X_4	4	3.614	0.718	0.701	11.424	−138.023	−130.067	4.597
X_1, X_2, X_3, X_4	5	3.084	0.759	0.740	5.000	−144.590	−134.645	4.069

Criteria for Model Selection, cont'd

- Model selection procedure have been developed to identify a small group of regression models that are "good" according to a specified criterion
- A detailed examination can then be made of a limited number of the more promising or "candidate" models, leading to the selection of the final regression model to be employed.
- Focus on six criteria:
  1. R_p^2 or SSE_p
  2. $R_{a,p}^2$ or MSE_p
  3. C_p
  4. AIC_p
  5. SBC_p or BIC_p
  6. $PRESS_p$

Criteria for Model Selection, cont'd

Some notations for model selection

- $P - 1$: the number of potential X variables in the pool
- all regression models contain an intercept term β_0 (contains **1** in design matrix $\boldsymbol{X}$ $\Leftrightarrow$ P parameters)
- $p - 1$: the number of X variables in a subset (p parameters in the regression function for this subset of X variables)
- $1 \leq p \leq P$
- $n > P$

R_p^2 or SSE_p Selection Criterion

- the use of R^2 $\Rightarrow$ which R^2 is high
- R_p^2 criterion $\Leftrightarrow$ the error sum of squares SSE_p (subsets for which SSEp is small)

$$R_p^2 = 1 - \frac{SSE_p}{SSTO}$$

- SSTO is constant for all possible regression models
- $R^2 \uparrow$ with #{X} $\uparrow$
- The R_p^2 criterion is not intended to identify the subsets that maximize this criterion. (max R_p^2 when all $P - 1$ potential X variables are included)
- to find the point where adding more X variables is not worthwhile ($\therefore$ it leads to a very small increase in R_p^2)

$R^2_{a,p}$ or MSE_p Selection Criterion

- Take account of the number of parameters
- $R^2_{a,p}$: the adjusted coefficient of multiple determination

$$R_p^2 = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE_p}{SSTO}$$

- SSTO is constant for all possible regression models
- $R^2_{a,p} \uparrow \Leftrightarrow MSE_p \downarrow$ for the given Y observations
- $Max(R^2_{a,p})$can decrease as p increases ($\therefore$ the increase in max R^2_p becomes small) $\Rightarrow$ it is not sufficient to offset the loss of an additional df.)
- $R^2_{a,p}$ criterion: to find a few subsets for which $R^2_{a,p}$ is at the maximum or so close to the maximum

Mallow's C_p Selection Criterion

- concerned with the total mean squared error of the n fitted values
- The mean squared error for $\hat{Y}_i$:

$$(\hat{Y}_i - \mu_i)^2 = \left[\underbrace{(E\{\hat{Y}_i\} - \mu_i)}_{\substack{\text{bias} \\ \text{component}}} + \underbrace{(\hat{Y}_i - E\{\hat{Y}_i\})}_{\substack{\text{random error} \\ \text{component}}} \right]^2$$

$$\Rightarrow E\{\hat{Y}_i - \mu_i\}^2 = \left(E\{\hat{Y}_i - \mu_i\}\right)^2 + \sigma^2\{\hat{Y}_i\}$$

- The total mean squared error for all n fitted values $\hat{Y}_i$:

$$\sum_{i=1}^{n} \left[\left(E\{\hat{Y}_i - \mu_i\}\right)^2 + \sigma^2\{\hat{Y}_i\} \right] = \sum_{i=1}^{n} \left(E\{\hat{Y}_i - \mu_i\}\right)^2 + \sum_{i=1}^{n} \sigma^2\{\hat{Y}_i\}$$

Mallow's C_p Selection Criterion, cont'd

- Γ_p: the criterion measure is the total mean squared error divided by σ^2

$$\Gamma_p = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left(E\{\hat{Y}_i - \mu_i\}\right)^2 + \sum_{i=1}^{n}\sigma^2\{\hat{Y}_i\}\right]$$

- C_p: the estimator of Γ_p

$$C_p = \frac{SSE_p}{MSE(X_1, \ldots, X_{P-1})} - (n - 2p)$$

Mallow's C_p Selection Criterion, cont'd

Why C_p is an estimator of Γ_p

- $\sum_{i=1}^{n} \sigma^2\{\hat{Y}_i\} = \text{trace}(Cov(\boldsymbol{Xb})) = p\sigma^2$
- $E\{SSE_p\} = \sum(E\{\hat{Y}_i\} - \mu_i)^2 + (n-p)\sigma^2$

$$E\{SSE_p\} = \sum E\{\hat{Y}_i - Y_i\}^2$$

$$= \sum E\left\{\left(\hat{Y}_i - E\{\hat{Y}_i\} - \varepsilon_i + E\{\hat{Y}_i\} - \mu_i\right)^2\right\}$$

$$\Rightarrow \Gamma_p = \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left(E\{\hat{Y}_i\} - \mu_i\right)^2 + \sum_{i=1}^{n}\sigma^2\{\hat{Y}_i\}\right]$$

$$= \frac{E\{SSE_p\}}{\sigma^2} - (n-2p)$$

$$\Rightarrow C_p = \frac{SSE_p}{MSE(X_1, \ldots, X_{P-1})} - (n-2p)$$

Mallow's C_p Selection Criterion, cont'd

- When $E\{\hat{Y}_i\} = \mu_i$ (no bias with $p-1$ X variables)

$$\Rightarrow E\{C_p\} \approx p$$

- the regression model containing all $P-1$ X variables:

$$C_P = \frac{SSE(X_1, \ldots, X_{P-1})}{\frac{SSE(X_1, \ldots, X_{P-1})}{n-P}} - (n-2P) = P$$

- The C_p measure assumes that $MSE(X_1, \ldots, X_{P-1})$ is an unbiased estimator of σ^2.
- like to find the model:
  1. C_p is small
  2. $C_p \approx p$

AIC_p or SBC_p Criteria

- provide penalties for adding predictors:
  - Akaike's information criterion: AIC_p
  - Schwarz' Bayesian criterion: SBC_p (Bayesian information criterion (BIC))

$$AIC_p = n \ln SSE_p - n \ln n + \underbrace{2p}_{(penalty)}$$

$$BIC_p = n \ln SSE_p - n \ln n + \underbrace{[\ln n]p}_{(penalty)}$$

- Search for models that have small values of AIC_p or SBC_p
- SBC_p criterion tends to favor more parsimonious models (large penalty if $n \geq 8$)

$PRESS_p$ Criteria

- $PRESS_p$: prediction sum of squares

$$PRESS_p = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(i)})^2$$

- a measure of how well the use of the fitted values for a subset model can predict the observed responses Y_i
- $Y_{i(i)}$: the fitted value when i is being predicted from a model in which (i) was left out when the regression function was fitted (leave-one-out)
- When the prediction errors $Y_i - \hat{Y}_{i(i)}$ are small, so are $PRESS_p$
- $PRESS_p$ can be calculated without requiring n separate regression runs, each time deleting one of the n cases (Chap. 10)

Example

- Surgical unit example

Key R codes:

- leaps: performs an exhaustive search for the best subsets of the variables in x
  - method=c("Cp", "adjr2", "r2"): Calculate *Cp*, adjusted R_a^2 or R^2
  - int=TRUE: Add an intercept to the model
  - nbest: Number of subsets of each size to report
- for loop procedure

"r2"

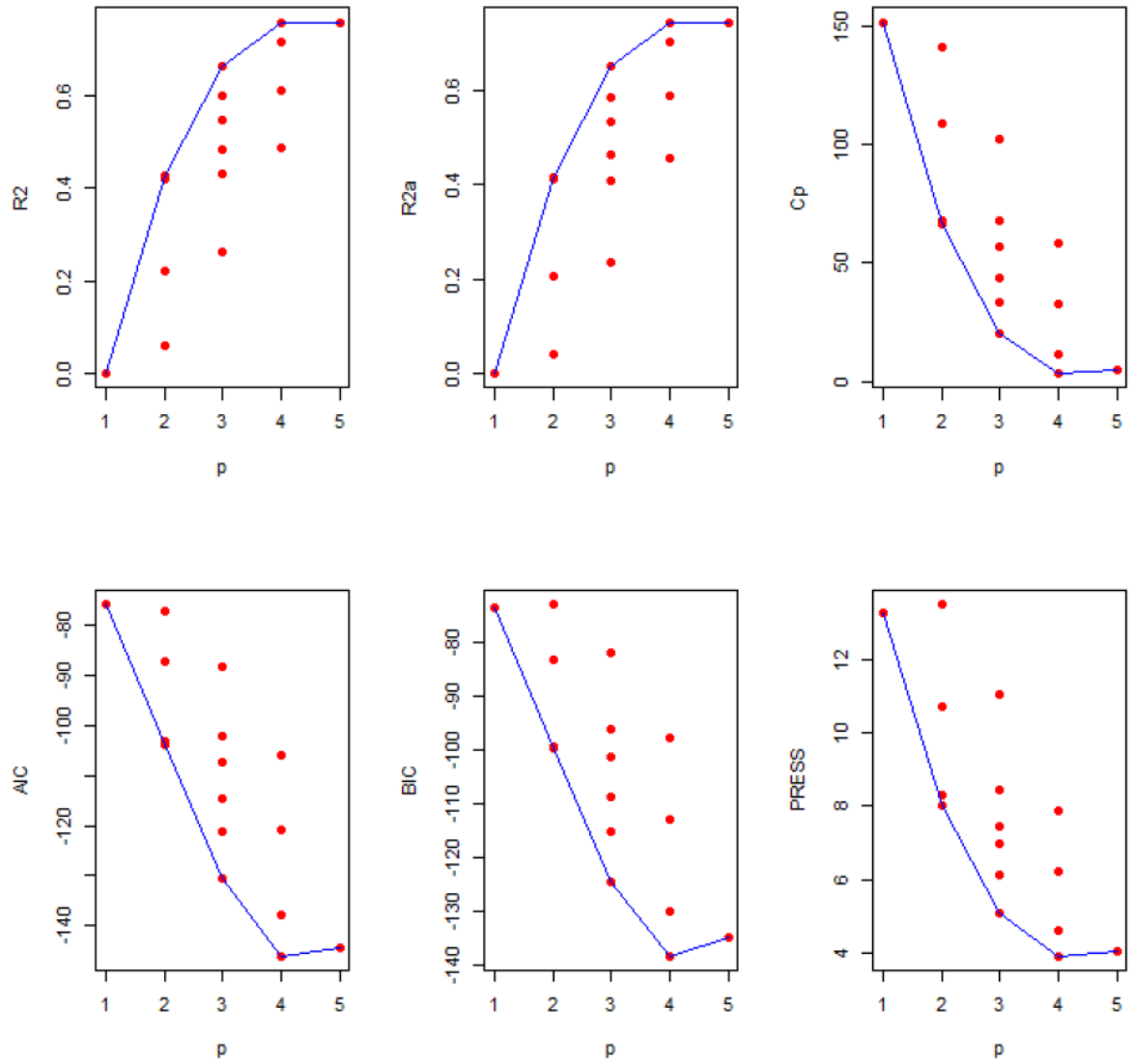
Example, cont'd

```
install.packages("leaps")
library(leaps)
attach(Dataset_09TA01)
ex<- Dataset_09TA01
ex.r2<-leaps( x=cbind(X1,X2,X3,X4,X5,X6,X7,X8),y=lnY, method='r2', nbest=6)
p<-seq( min(ex.r2$size),max(ex.r2$size) )
ind<-as.data.frame(ex.r2[c(3:4)])
ind<-ind[with(ind, order(size,r2)), ]
plot(ind[,c(1:2)] ,ylab=expression(R^2), xlab='p' ,col="red",pch=16)
Rp2 = by( data=ex.r2[4],INDICES=factor(ex.r2$size), FUN=max)
lines( Rp2 ~ p,col="blue" )
```

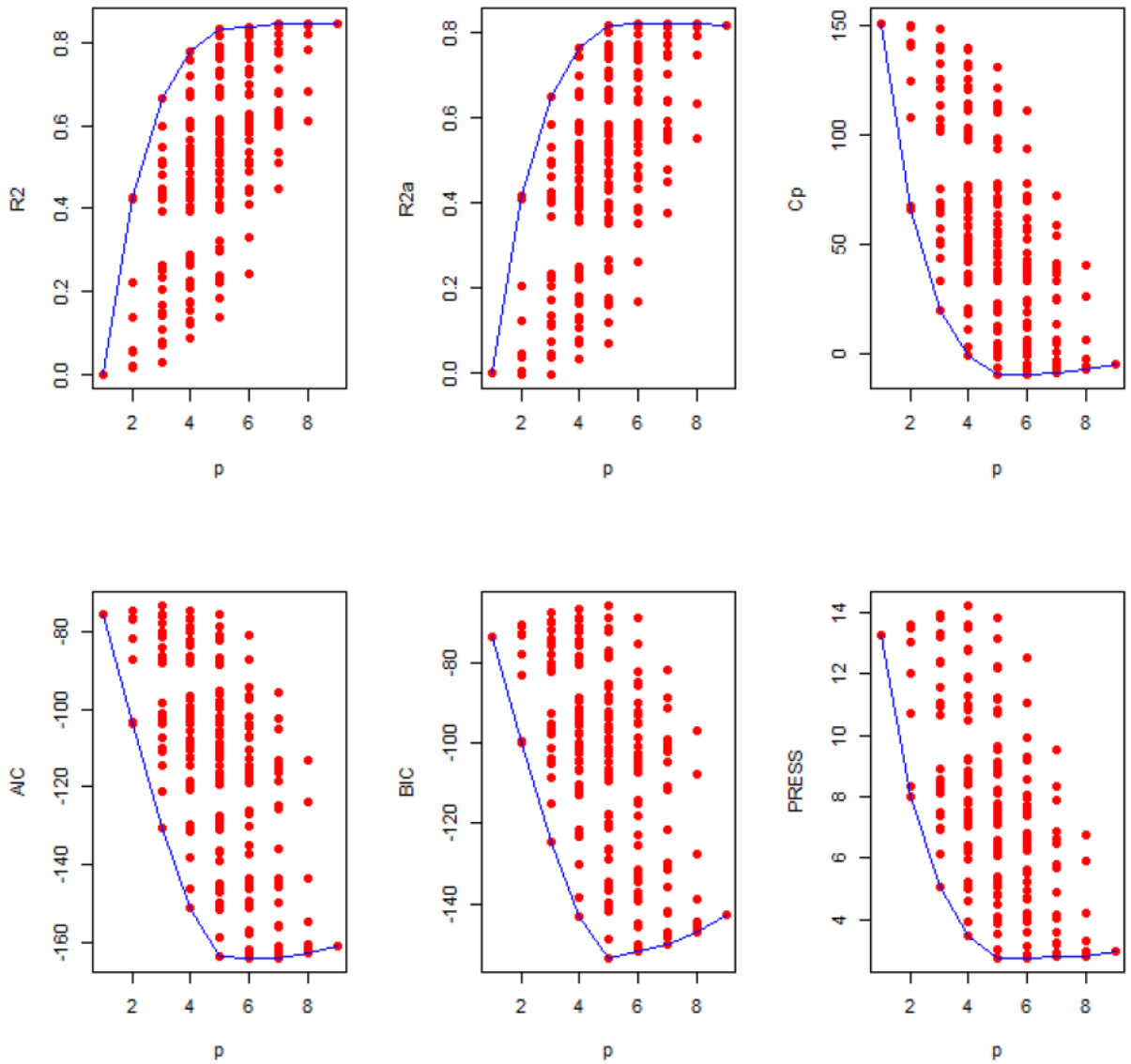
Example, cont'd

No.	Variables	p	R_p^2	$R_{a,p}^2$	C_p	AIC_p	SBC_p	$PRESS_p$
1	None	1.000	0.000	0.000	151.498	-75.703	-73.714	13.296
2	X1	2.000	0.061	0.043	141.164	-77.079	-73.101	13.512
3	X2	2.000	0.221	0.206	108.556	-87.178	-83.200	10.744
4	X3	2.000	0.428	0.417	66.489	-103.827	-99.849	8.327
5	X4	2.000	0.422	0.410	67.715	-103.262	-99.284	8.025
6	X1 + X2	3.000	0.263	0.234	102.031	-88.162	-82.195	11.062
7	X1 + X3	3.000	0.549	0.531	43.852	-114.658	-108.691	6.988
8	X1 + X4	3.000	0.430	0.408	67.972	-102.067	-96.100	8.472
9	X2 + X3	3.000	0.663	0.650	20.520	-130.483	-124.516	5.065
10	X2 + X4	3.000	0.483	0.463	57.215	-107.324	-101.357	7.476
11	X3 + X4	3.000	0.599	0.584	33.504	-121.113	-115.146	6.121
12	X1 + X2 + X3	4.000	0.757	0.743	3.391	-146.161	-138.205	3.914
13	X1 + X2 + X4	4.000	0.487	0.456	58.392	-105.748	-97.792	7.903
14	X1 + X3 + X4	4.000	0.612	0.589	32.932	-120.844	-112.888	6.207
15	X2 + X3 + X4	4.000	0.718	0.701	11.424	-138.023	-130.067	4.597
16	X1 + X2 + X3 + X4	5.000	0.759	0.740	5.000	-144.590	-134.645	4.069

Example, cont'd



Example, cont'd



Automatic Search Procedures for Model Selection

#{possible models } (2^{P-1}) ↑ with # {predictors}
Evaluating all of the possible alternatives can be a daunting

Automatic computer-search procedures:

- "Best" subsets algorithms
- stepwise regression

“Best” Subset Algorithm

- Time-saving algorithms have been developed according to a specified criterion, example: $C_p \Rightarrow$ the best subsets
- Without requiring the fitting of all the possible subset regression models
- only a small fraction of all possible regression models is required to calculate.
- Several regression models can be identified as "good" for final consideration, depending on which criteria we use.

“Best” Subset Algorithm, cont’d

The surgical unit example

- $R^2_{a,p} = 0.823$: seven- or eight-parameter model
- $\min(C_7) = 5.541$
- $\min(AIC_7) = -163.834$
- $\min(SBC_5) = -153.406$
- $\min(PRESS_5) = 2.738$

Not to identify a single best model; hope to identify a small set of promising models for further study

“Best” Subset Algorithm, cont’d

p	(1) SSE_p	(2) R_p^2	(3) $R_{a,p}^2$	(4) C_p	(5) AIC_p	(6) SBC_p	(7) $PRESS_p$
1	12.808	0.000	0.000	240.452	−75.703	−73.714	13.296
2	7.332	0.428	0.417	117.409	−103.827	−99.849	8.025
3	4.312	0.663	0.650	50.472	−130.483	−124.516	5.065
4	2.843	0.778	0.765	18.914	−150.985	−143.029	3.469
5	2.179	0.830	0.816	5.751	−163.351	−153.406	2.738
6	2.082	0.837	0.821	5.541	−163.805	−151.871	2.739
7	2.005	0.843	0.823	5.787	−163.834	−149.911	2.772
8	1.972	0.846	0.823	7.029	−162.736	−146.824	2.809
9	1.971	0.846	0.819	9.000	−160.771	−142.870	2.931

Selection Method

- The forward stepwise regression procedure is probably the most widely used of the automatic search methods.
- The search method develops a sequence of regression models, at each step adding or deleting an *X* variable. (t^* or F^* statistics)
- the stepwise search procedures end with the identification of a single regression model as "best"

Different formats:

- Forward Selection; Forward Stepwise Selection
- Backward Elimination; Backward Stepwise Elimination

Forward Stepwise Selection Method

The forward stepwise regression search algorithm: t^* statistics and the associated *P*-values

1. Fit a simple linear regression model for each of the P − 1 X variables considered for inclusion. For each compute the t statistics for testing whether or not the slope is zero:

$$t^* = \frac{b_k}{s\{b_k\}}$$

2. Pick the largest out of the P − 1 $|t_k^*|$'s and include the corresponding X variable in the regression model if $|t_k^*|$exceeds some threshold (ex: comparing the P-value with the limits in = 0.1). If no such $|t_k^*| \Rightarrow$ stop the procedure

Forward Stepwise Selection Method, cont'd

3. Using partial t statistics to add the second variable (ex: comparing the P-value with the limits $\alpha_{in} = 0.1$)
4. Test whether variables in the model need to be dropped from the model: Using partial t statistics (ex: comparing the P-value with the limits $\alpha_{out} = 0.15$)
5. Go to Step 3 and continue the procedure until stop.

Note that

- allow an X variable brought into the model at an early stage, to be dropped subsequently if it is no longer helpful
- small α_{in} are often underspecified
- $\alpha_{in} < \alpha_{out}$: avoid cycling

Forward Stepwise Selection Method, cont'd

- forward selection
- Backward Elimination: start with all $P - 1$ X-variables
  - Backward Stepwise Elimination
- R codes:
  - step(): Select a formula-based model by AIC.
  - fastbw(): Fast Backward Variable Selection (library(rms))
  - add1() or drop1(): Add or Drop All Possible Single Terms to a Model

Model Validation

- the validation of the selected regression models
- checking a candidate model against independent data

Three basic ways of validating a regression model:

- Collection of new data to check the model and its predictive ability
- Comparison of results: theoretical expectation; early results; simulation results
- Holdout sample: check the model and its predictive ability

Model Validation, cont'd

Collection of New data to Check Model

- A means of measuring the actual predictive capability of the selected regression model is to use this model to predict each case in the new data set and then to calculate the mean of the squared prediction errors, to be denoted by *MSPR*, which stands for mean squared prediction error:

$$MSPR = \frac{\sum_{i=1}^{n^*}(Y_i - \hat{Y}_i)^2}{n^*}$$

- Difficulties in replicating a study

Model Validation, cont'd

Comparing with Theory, Empirical Evidence, or Simulation Results

- Unfortunately, there is often little theory that can be used to validate regression models

Model Validation, cont'd

Data Splitting

- often called *cross-validation* : the data set is large enough to split the data into two sets
  - training sample: model-building set
  - prediction set: used to evaluate the reasonableness and predictive ability of the selected model
- If the entire data set is not large enough for making an equal split, the validation data set will need to be smaller than the model-building data set.
- Splits of the data can be made at random.
- Drawback: the variances of the estimated regression coefficients developed from the model-building data set will usually be larger than those that would have been obtained from the fit to the entire data set

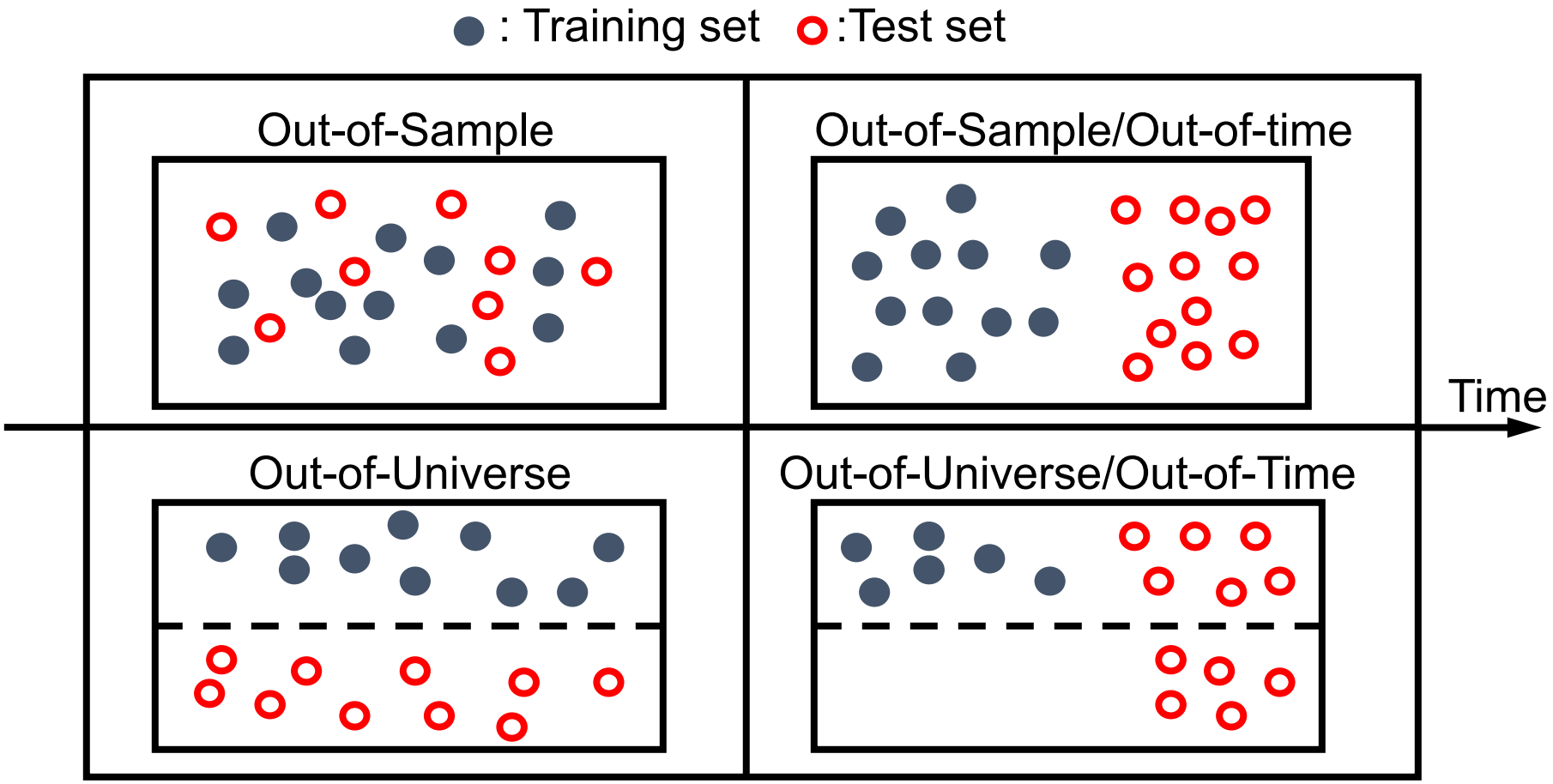
Model Validation, cont'd

- In any case, once the model has been validated, it is customary practice to use the entire data set for estimating the final regression model.

Comments:

- double cross-validation procedure
- *K*-fold cross-validating

Out-of-Sample Versus Out-of-Time versus Out-of-Universe Quantitative Validation



Validation During Model Development versus Validation During Model Usage

- Validation during model development
  - Typically Out-of-sample
- Validation during model usage
  - Automatically Out-of-sample/Out-of-time (sometimes also Out-of-universe)
- Note
  - Validation during model development: for small samples, use cross-validation methods (e.g., leave-one out cross-validation)