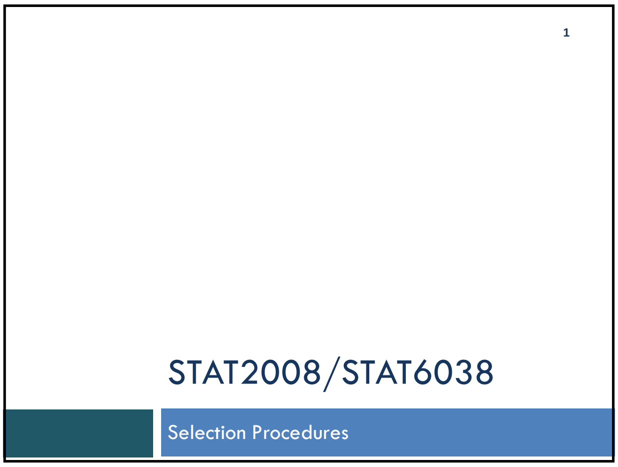
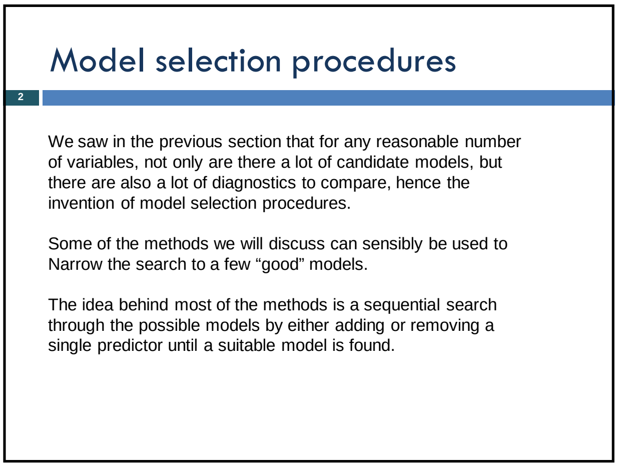


STAT2008/STAT6038

Selection Procedures

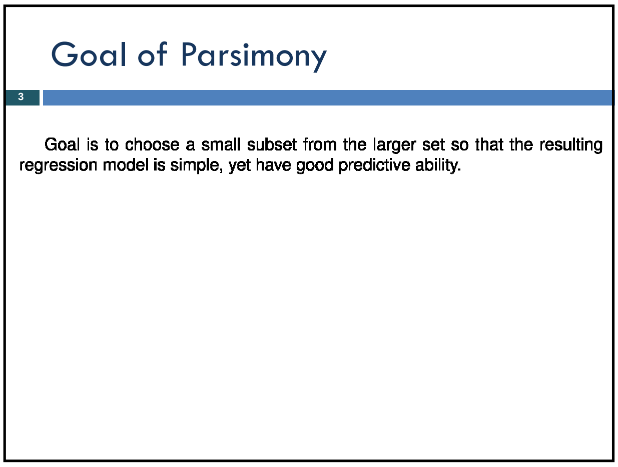


Model selection procedures

We saw in the previous section that for any reasonable number of variables, not only are there a lot of candidate models, but there are also a lot of diagnostics to compare, hence the invention of model selection procedures.

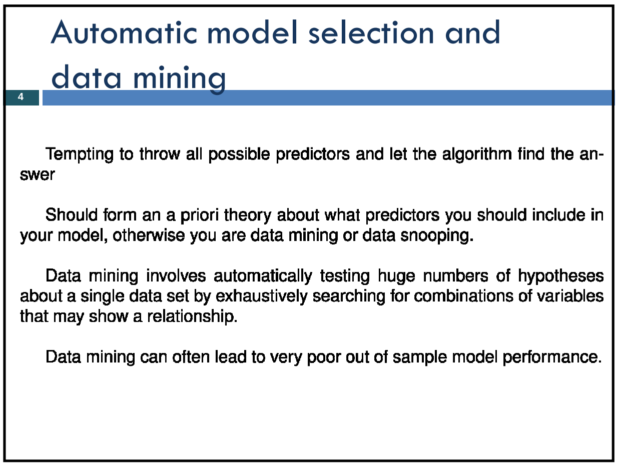
Some of the methods we will discuss can sensibly be used to Narrow the search to a few “good” models.

The idea behind most of the methods is a sequential search through the possible models by either adding or removing a single predictor until a suitable model is found.



Goal of Parsimony

Goal is to choose a small subset from the larger set so that the resulting regression model is simple, yet have good predictive ability.



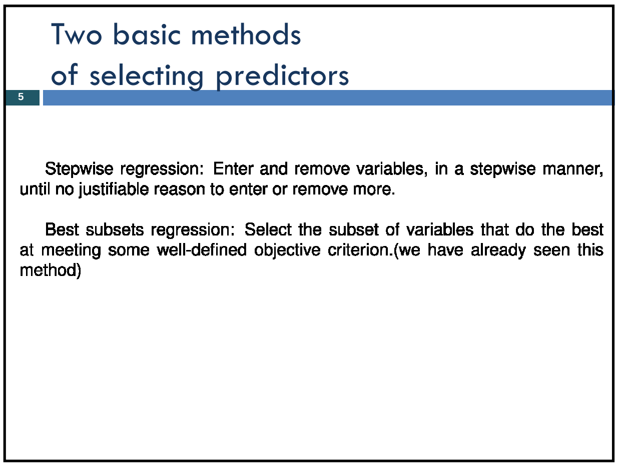
Automatic model selection and data mining

Tempting to throw all possible predictors and let the algorithm find the answer

Should form an a priori theory about what predictors you should include in your model, otherwise you are data mining or data snooping.

Data mining involves automatically testing huge numbers of hypotheses about a single data set by exhaustively searching for combinations of variables that may show a relationship.

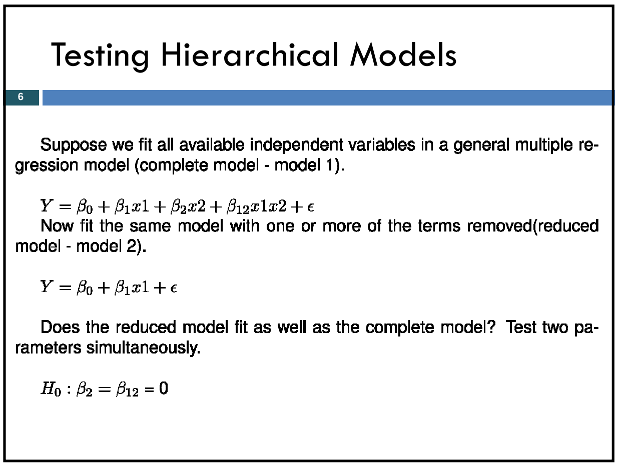
Data mining can often lead to very poor out of sample model performance.



Two basic methods of selecting predictors

Stepwise regression: Enter and remove variables, in a stepwise manner, until no justifiable reason to enter or remove more.

Best subsets regression: Select the subset of variables that do the best at meeting some well-defined objective criterion.(we have already seen this method)



Testing Hierarchical Models

Suppose we fit all available independent variables in a general multiple regression model (complete model - model 1).

$Y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_{12} x1x2 + \epsilon$

Now fit the same model with one or more of the terms removed(reduced model - model 2).

$Y = \beta_0 + \beta_1 x1 + \epsilon$

Does the reduced model fit as well as the complete model? Test two parameters simultaneously.

$H_0 : \beta_2 = \beta_{12} = 0$

Testing Hierarchical Models

7

Let SSE1 be the error sums of squares for the complete model:

$$Y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_{12} x1 x2 + \epsilon$$

Let SSE2 be the error sums of squares for the reduced model:

$$Y = \beta_0 + \beta_1 x1 + \epsilon$$

Since Model 1 includes more terms than Model 2, it should fit better, hence we have that:

SSE1 $\leq$ SSE2

The difference, SSE2 - SSE1 is a measure of the drop in the sum of squares for error attributable to the variables removed from the complete model.

Testing Hierarchical Models

8

Define the mean square drop as:

MSdrop = (SSE2 - SSE1) / (k-g)

where k is the number of terms in the complete model (Model 1) and (g < k) is the number of terms in the reduced model (Model 2).

The mean square error for the complete model is:

MSE1 = SSE1 / (n-k-1)

To test the hypothesis that the terms left out of the complete model do not contribute significantly to explaining the variability in y we use the following F statistic.

F = MSdrop/MSE1

Reject H_o: Left out parameters = 0 if $F > F_{k-g,n-k-1,\alpha}$

Sequential Models

9

With this concept of partial and full models, we can look at models in two different ways.

$Y = \beta_0 + \varepsilon$	Add constant	Mean Model
$Y = \beta_0 + \beta_1 x_1 + \varepsilon$	Add x_1	Model 1
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$	Add x_2	Model 2
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_2 x_2 + \varepsilon$	Add cross product	Model 3

Questions:

Is Model 1 better than the Mean Model?
Is Model 2 better than Model 1?
Is Model 3 better than Model 2?

All of these are tested with a drop type test.

Last In Significant (Partial) Tests

10

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$	Does x_3 add to model with x_1, x_2?
$y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_2 x_2 + \varepsilon$	Does x_2 add to model with x_1, x_3?
$y = \beta_0 + \beta_2 x_2 + \beta_3 x_3 + \beta_1 x_1 + \varepsilon$	Does x_1 add to model with x_2, x_3?

How important is a predictor in explaining variability in the response over and above what is explained by predictors already in the model. All these are tested with a drop type test.

Reduction (Drop) Sums of Squares

11

$Y = \beta_0 + \varepsilon$	Add constant	Mean Model
$Y = \beta_0 + \beta_1 x_1 + \varepsilon$	Add x_1	Model 1
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$	Add x_2	Model 2
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$	Add cross product	Model 3

SST= SSR+ SSE

$SSR(\beta_2 | \beta_0, \beta_1) = SSR_2 - SSR_1$
$= SSE_1 - SSE_2$

$SSR(\beta_{12} | \beta_0, \beta_1\ \beta_2) = SSR_3 - SSR_2$
$= SSE_2 - SSE_3$

Additional variability in Y explained by model 3, above and beyond what is already explained by model 2.

Variable/Predictor Selection

12

Find the“best” (an appropriate) subset of (predictors) for the model from among all possible candidate predictors

Problem: How do we define “best”?

Bias/Variance Tradeoff:

BIAS:The model should include as many regressors as possible so that the information content in these predictors can adequately predict Y.

VARIANCE: But we know that the variance of predictions increases as the number of regressors increases. (We also want parsimonious (simple) model for ease of interpretability).

Variance of parameter estimates increase as we add predictors

Deleting variables from the complete model actually **decreases** the variances of the parameter estimates for the remaining explanatory variables. (R^2 always increases/decreases when variables are added/deleted to/from the model.)

$$s_{\hat{\beta}_j} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{S_{x_j x_j}(1 - R^2_{x_j \bullet x_1 x_2 \cdots x_{j-1} x_{j+1} \cdots x_k})}}$$

Deleting variables from the complete model **improves** the precision of the estimates for the remaining parameters, as well as the precision of the predictions.

But, if we remove an explanatory variable which is strongly associated with Y, we produce biased estimates of the remaining parameters, the residual variance, and the predictions.

Model Selection Criteria

Choose models with high R^2. However, R^2 increases every time more predictors are added, regardless of their importance in predicting Y. Adjusted R^2.

Choose models with high R^2_{adj}. Better suited for model selection than R^2. It increases/decreases only when important/unimportant predictors are added to the model.

Choose models with low MSE.

Choose models with C_p close to k+1. Measures adequacy of predictions from reduced model relative to those from the full model. Recall that if the model is unbiased this statistic will equal p=k+1

Choose models with low PRESS statistic (Predicted Residual Sum of Squares

AIC

Akaike's Information Criterion (AIC). Measures how far the candidate model is from the "true" model. Choose models with low AIC.

For a multiple regression model with n obs and k predictors,

$$AIC_c = n\log\left(\frac{n-k-1}{n}MSE\right) + \frac{2n(k+2)}{n-k-3}$$

Require at least 10 times more observations than predictors in the candidate models, i.e. 10k<n.

Automatic Model Selection

Backward Elimination: First fit the model with all possible predictors, then sequentially eliminate those predictors that are least significant in a last-in (partial) test.

Forward Selection: First find the one predictor, call it x_1, that does the very best job of explaining variation in y. Then add to the model the predictor that is most significant (via a last-in test) when added into the model after x_1. Continue until no additional predictors contribute significantly using a last-in test.

Stepwise Selection: Begin as with the forward selection method, but each time a new predictor is added into the model, check all other predictors with a last-in test to determine if they should continue to be in the model. Drop any predictor that cannot pass the last-in test.

Stepwise is usually the preferred method

Peruvian Indian Data

Anthropologists studying long-term effects of an environmental change on systolic blood pressure, measured this and several other characteristics of 39 Indians who migrated from a primitive environment high in the Andes, into mainstream Peruvian society at a lower elevation. The variables were:

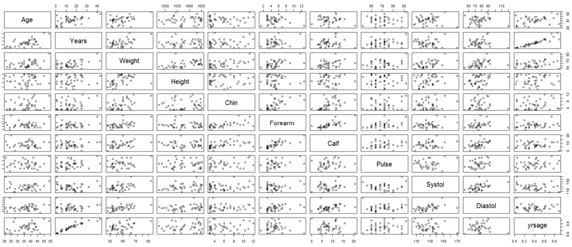
X_1 = Age (years) X_2 = Years since migration X_3 = Weight (kg)
X_4 = Height (mm) X_5 = Chin skin fold (mm) X_6 = Forearm skin fold (mm)
X_7 = Calf skin fold (mm) X_8 = Pulse rate (beats/min) ***X_9 = Systolic blood pressure***
X_{10} = Diastolic blood pressureX_{11} = Years since migration divided by age

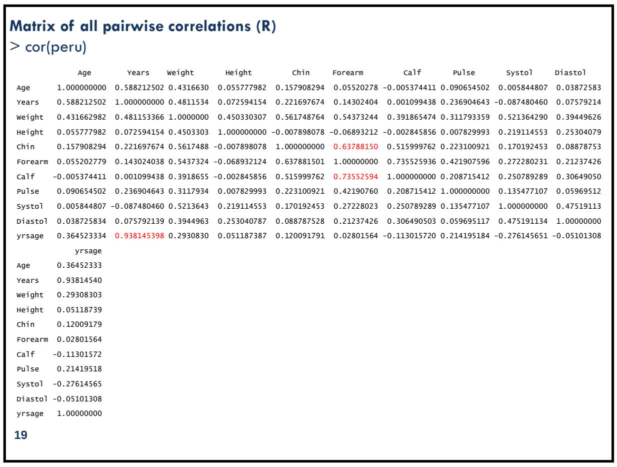
Systolic blood pressure will be our response variable

Pairwise scatterplots

R:

```
> pairs(peru)
```

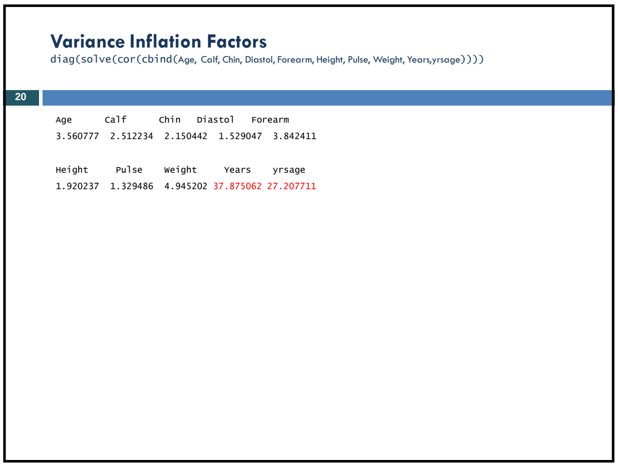




Matrix of all pairwise correlations (R)

> cor(peru)

```
               Age       Years    Weight      Height         Chin    Forearm         Calf       Pulse       Systol    Diastol
Age      1.000000000 0.588212502 0.4316630  0.055777982  0.157908294  0.05520278 -0.005374411 0.090654502  0.005844807 0.03872583
Years    0.588212502 1.000000000 0.4811534  0.072594154  0.221697674  0.14302404  0.001099438 0.236904643 -0.087480460 0.07579214
Weight   0.431662982 0.481153366 1.0000000  0.450330307  0.561748764  0.54373244  0.391865474 0.311793359  0.521364290 0.39449626
Height   0.055777982 0.072594154 0.4503303  1.000000000 -0.007898078 -0.06893212 -0.002845856 0.007829993  0.219114553 0.25304079
Chin     0.157908294 0.221697674 0.5617488 -0.007898078  1.000000000  0.63788150  0.515999762 0.223100921  0.170192453 0.08878753
Forearm  0.055202779 0.143024038 0.5437324 -0.068932124  0.637881501  1.00000000  0.735525936 0.421907596  0.272280231 0.21237426
Calf    -0.005374411 0.001099438 0.3918655 -0.002845856  0.515999762  0.73552594  1.000000000 0.208715412  0.250789289 0.30649050
Pulse    0.090654502 0.236904643 0.3117934  0.007829993  0.223100921  0.42190760  0.208715412 1.000000000  0.135477107 0.05969512
Systol   0.005844807 -0.087480460 0.5213643  0.219114553  0.170192453  0.27228023  0.250789289 0.135477107  1.000000000 0.47519113
Diastol  0.038725834 0.075792139 0.3944963  0.253040787  0.088787528  0.21237426  0.306490503 0.059695117  0.475191134 1.00000000
yrsage   0.364523334 0.938145398 0.2930830  0.051187387  0.120091791  0.02801564 -0.113015720 0.214195184 -0.276145651 -0.05101308
              yrsage
Age       0.36452333
Years     0.93814540
Weight    0.29308303
Height    0.05118739
Chin      0.12009179
Forearm   0.02801564
Calf     -0.11301572
Pulse     0.21419518
Systol   -0.27614565
Diastol  -0.05101308
yrsage    1.00000000
```

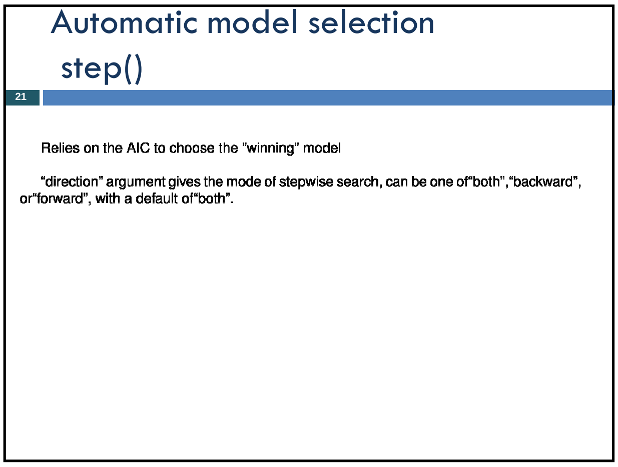


Variance Inflation Factors

diag(solve(cor(cbind(Age, Calf, Chin, Diastol, Forearm, Height, Pulse, Weight, Years,yrsage))))

```
Age       Calf      Chin    Diastol   Forearm
3.560777  2.512234  2.150442  1.529047  3.842411

Height     Pulse     Weight     Years     yrsage
1.920237  1.329486  4.945202 37.875062 27.207711
```

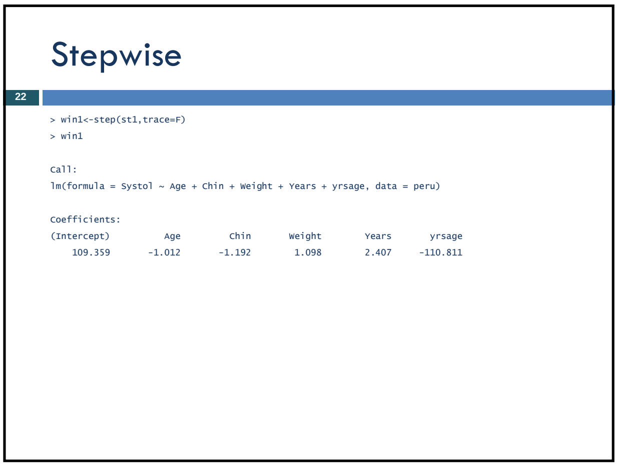


Automatic model selection

step()

Relies on the AIC to choose the "winning" model

"direction" argument gives the mode of stepwise search, can be one of"both","backward", or"forward", with a default of"both".

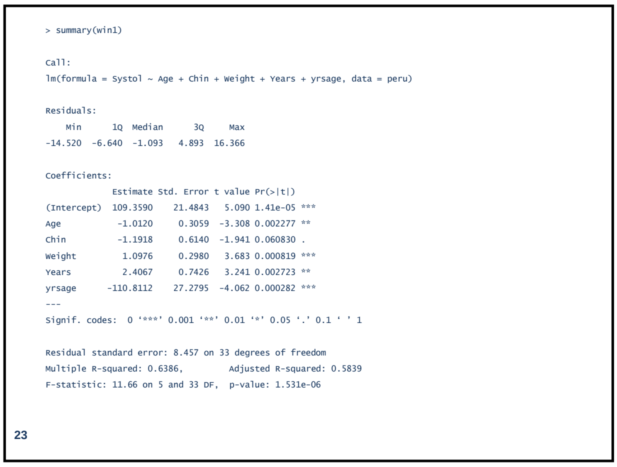


Stepwise

```
> win1<-step(st1,trace=F)
> win1

Call:
lm(formula = Systol ~ Age + Chin + Weight + Years + yrsage, data = peru)

Coefficients:
(Intercept)          Age         Chin       Weight        Years       yrsage
    109.359       -1.012       -1.192        1.098        2.407     -110.811
```



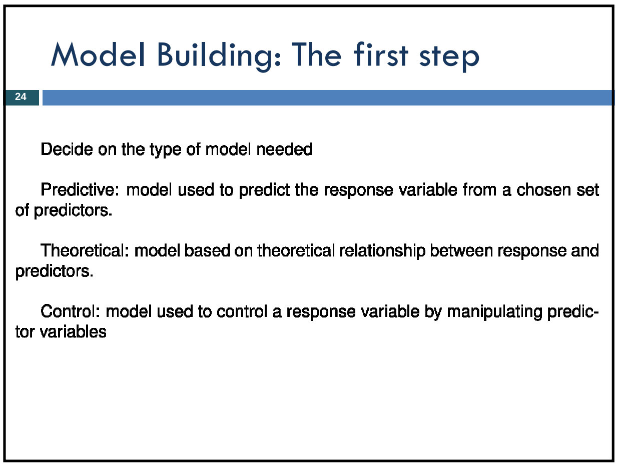
```
> summary(win1)

Call:
lm(formula = Systol ~ Age + Chin + Weight + Years + yrsage, data = peru)

Residuals:
    Min      1Q  Median      3Q     Max
-14.520  -6.640  -1.093   4.893  16.366

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.3590    21.4843   5.090 1.41e-05 ***
Age          -1.0120     0.3059  -3.308 0.002277 **
Chin         -1.1918     0.6140  -1.941 0.060830 .
Weight        1.0976     0.2980   3.683 0.000819 ***
Years         2.4067     0.7426   3.241 0.002723 **
yrsage      -110.8112    27.2795  -4.062 0.000282 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.457 on 33 degrees of freedom
Multiple R-squared: 0.6386,	Adjusted R-squared: 0.5839
F-statistic: 11.66 on 5 and 33 DF,  p-value: 1.531e-06
```



Model Building: The first step

Decide on the type of model needed

Predictive: model used to predict the response variable from a chosen set of predictors.

Theoretical: model based on theoretical relationship between response and predictors.

Control: model used to control a response variable by manipulating predictor variables

The second step

25

Decide which predictor variables and response variable on which to collect the data.

Collect the data. (often not as easy as it sounds)

The third step

26

Explore the data

Check for outliers, gross data errors, missing values on a univariate basis.

Study bivariate relationships to reveal other outliers, to suggest possible transformations, to identify possible multicollinearities.

The fourth step

27

Randomly divide the data into a training set and a test set:

The training set, rule of thumb (min of 15-20 error d.f)

The test set is used for cross-validation of the fitted model.

The fifth step

28

Using the training set, fit several candidate models:

Use best subsets regression.

Use stepwise regression (only gives one model unless specifies different alpha-to-remove and alpha-to-enter values).

The sixth step

29

Select and evaluate a few good models:

Select based on R^2_{adj}, Mallows Cp, AIC, MSE, PRESS and number (parsimony) and nature of predictors (e.g. do they make sense?).

Evaluate selected models for violation of model assumptions.

If none of the models provide a satisfactory fit, try something else, such as more data, different predictors, a different class of model, transformations