Model selection

Tuan V. Nguyen
Professor and NHMRC Senior Research Fellow
Garvan Institute of Medical Research
University of New South Wales
Sydney, Australia

Contents

- To describe some techniques for selecting the explanatory variables for a regression
- To describe the consequences of making an incorrect choice
- To apply these techniques to an example

Variable Selection

- Often there are several (perhaps a large number) of potential explanatory variables available to build a regression model. Which ones should we use?

- We could, of course, use them all. However, this turns out to be not such a good idea.

Overfitting

- If we put too many variables in the model, including some unrelated to the response, we are ***overfitting***. Consequences are:
  - Fitted model is not good for prediction of new data – prediction error is inflated
  - Model is too elaborate, models "noise" that will not be the same for new data

Underfitting

- If we put too few variables in the model, leaving out variables that could help explain the response, we are ***underfitting***. Consequences:
  - Fitted model is not good for prediction of new data – prediction is biased
  - Regression coefficients are biased
  - Estimate of error variance is too large

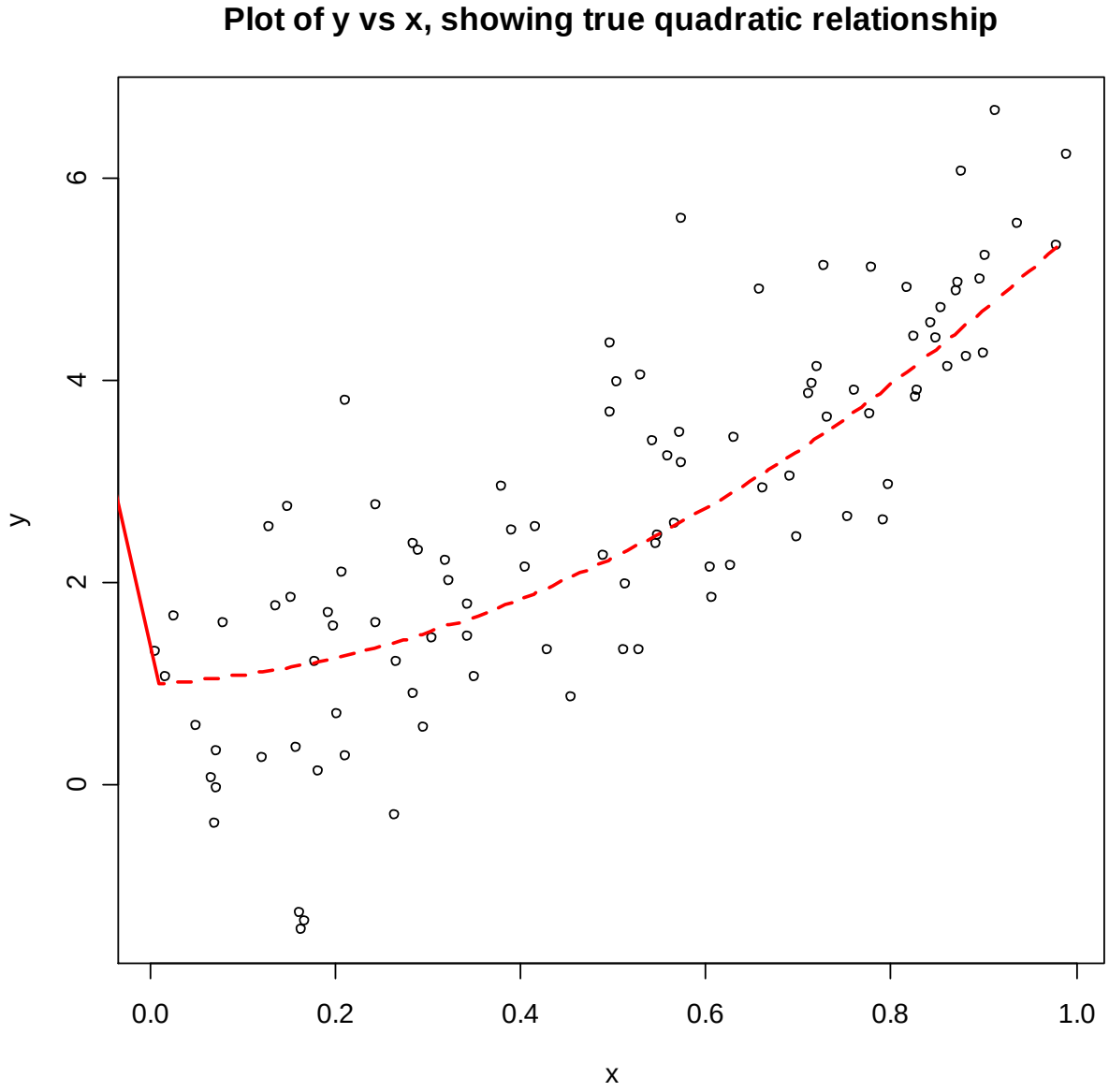
Example

- Suppose we have some data which follow a quadratic model

$$Y = 1 + 0.5\,x + 4\,x^2 + N(0,1)$$

where the x's are uniform on [0,1]

The next slide shows the data, with the true regression shown as a dotted line.

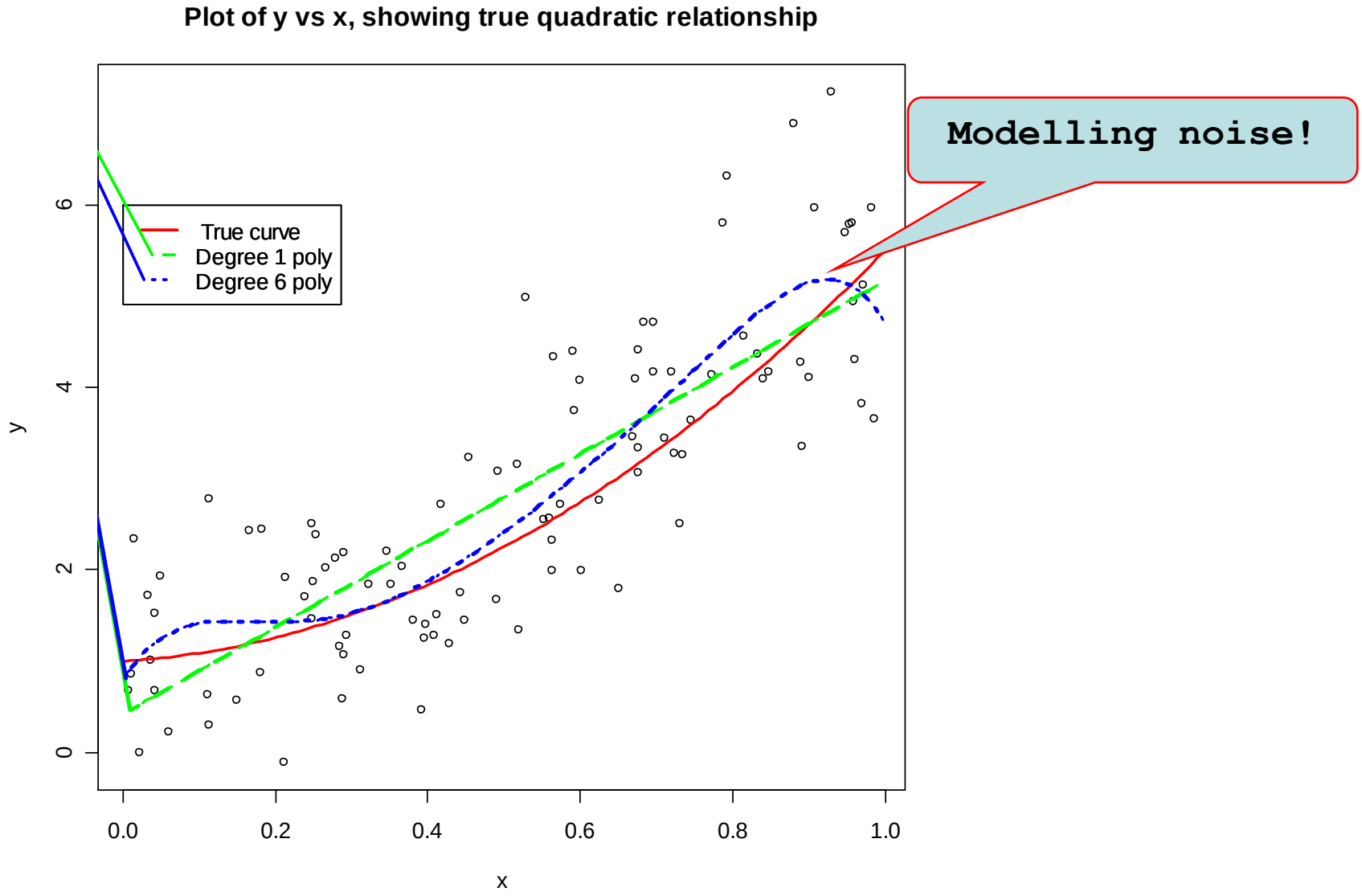

Plot of y vs x, showing true quadratic relationship
y
x
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0

Under-fitting and over-fitting

- Suppose we fit a straight line. This is underfitting, since we are not fitting the squared term. The fitted line (in green) is shown on the next slide.

- Alternatively, we could fit a 6-degree polynomial. This is overfitting, since there are unnecessary terms in x^3, x^4, x^5 and x^6. The fitted polynomial is shown in blue on the next slide. Fit using

```
lm(y ~ poly(x,6))
```


Plot of y vs x, showing true quadratic relationship
True curve
Degree 1 poly
Degree 6 poly
y
x
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0
Modelling noise!

Points to note

- Straight line is biased: can't capture the curvature in the true regression
- 6-degree line: too variable, attracted to the errors which would be different for a new set of data
- Moral: For good models we need to choose variables wisely to avoid overfitting and underfitting.

This is called ***variable selection***

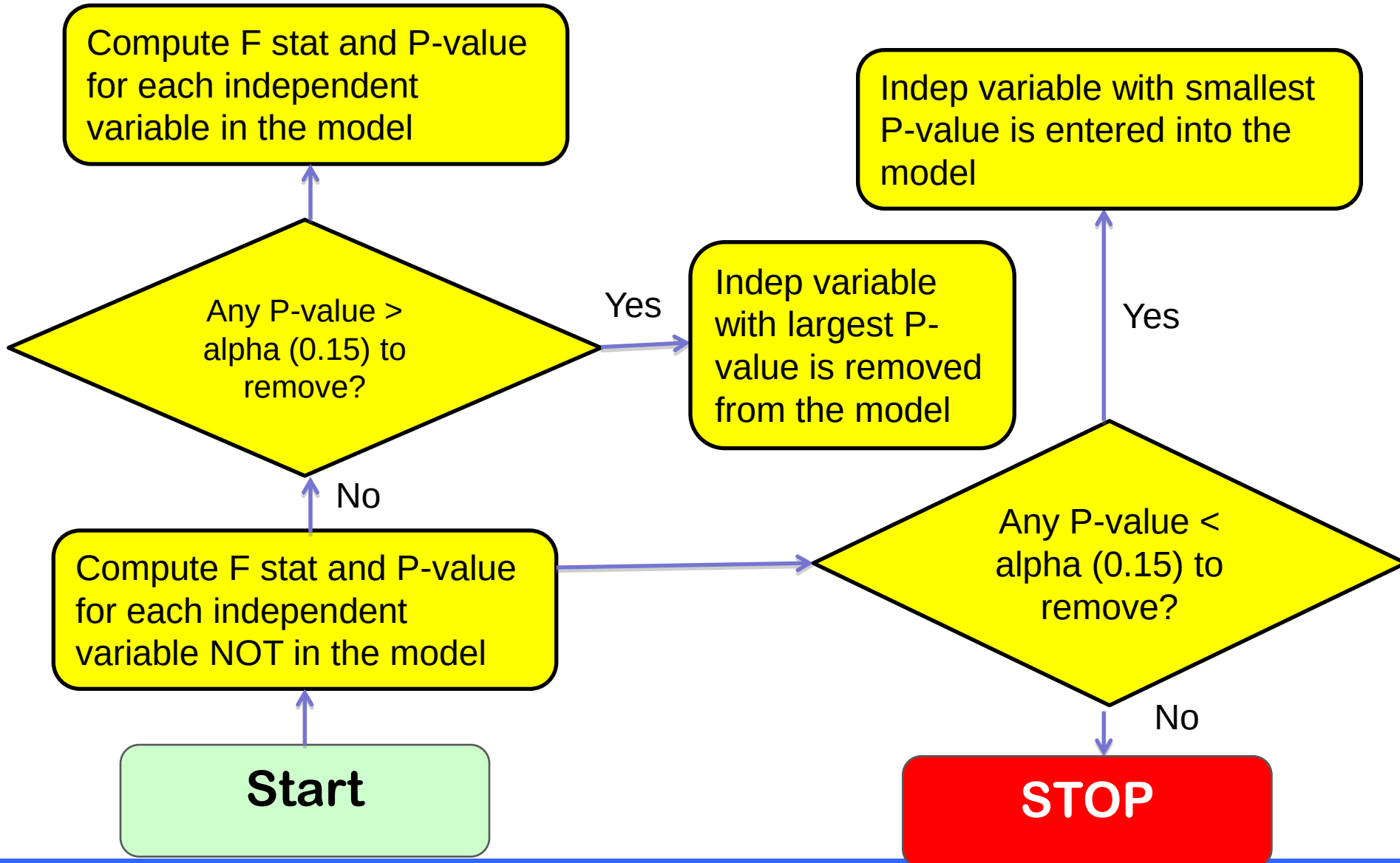
Methods for variable selection

- If we have k variables, and assuming a constant term in each model, there are 2^k-1 possible subsets of variables (not counting the null model with no variables)
- How do we select a subset for our model?

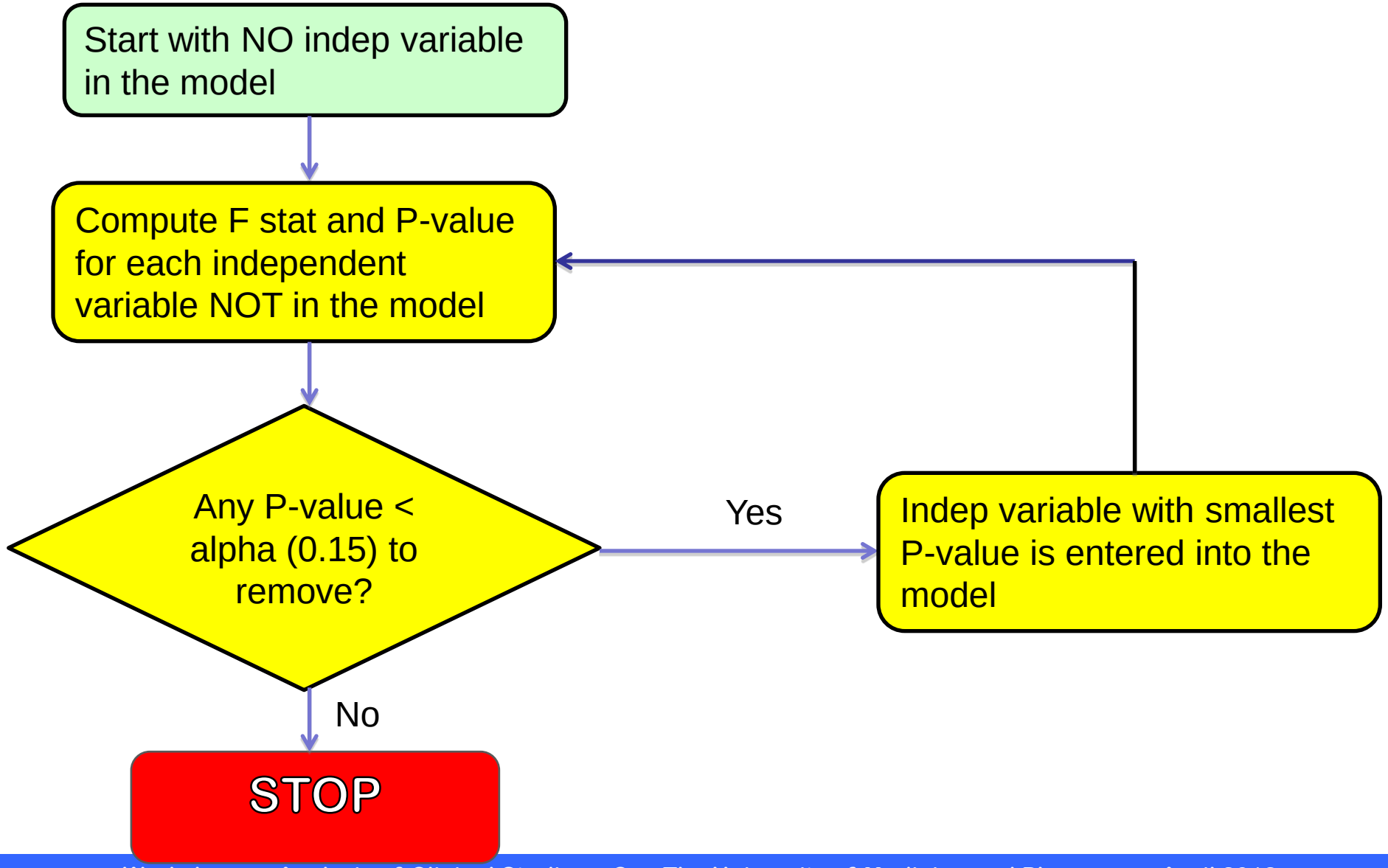
- Two main approaches: stepwise methods and all possible regressions (APR)

Stepwise Regression Procedure

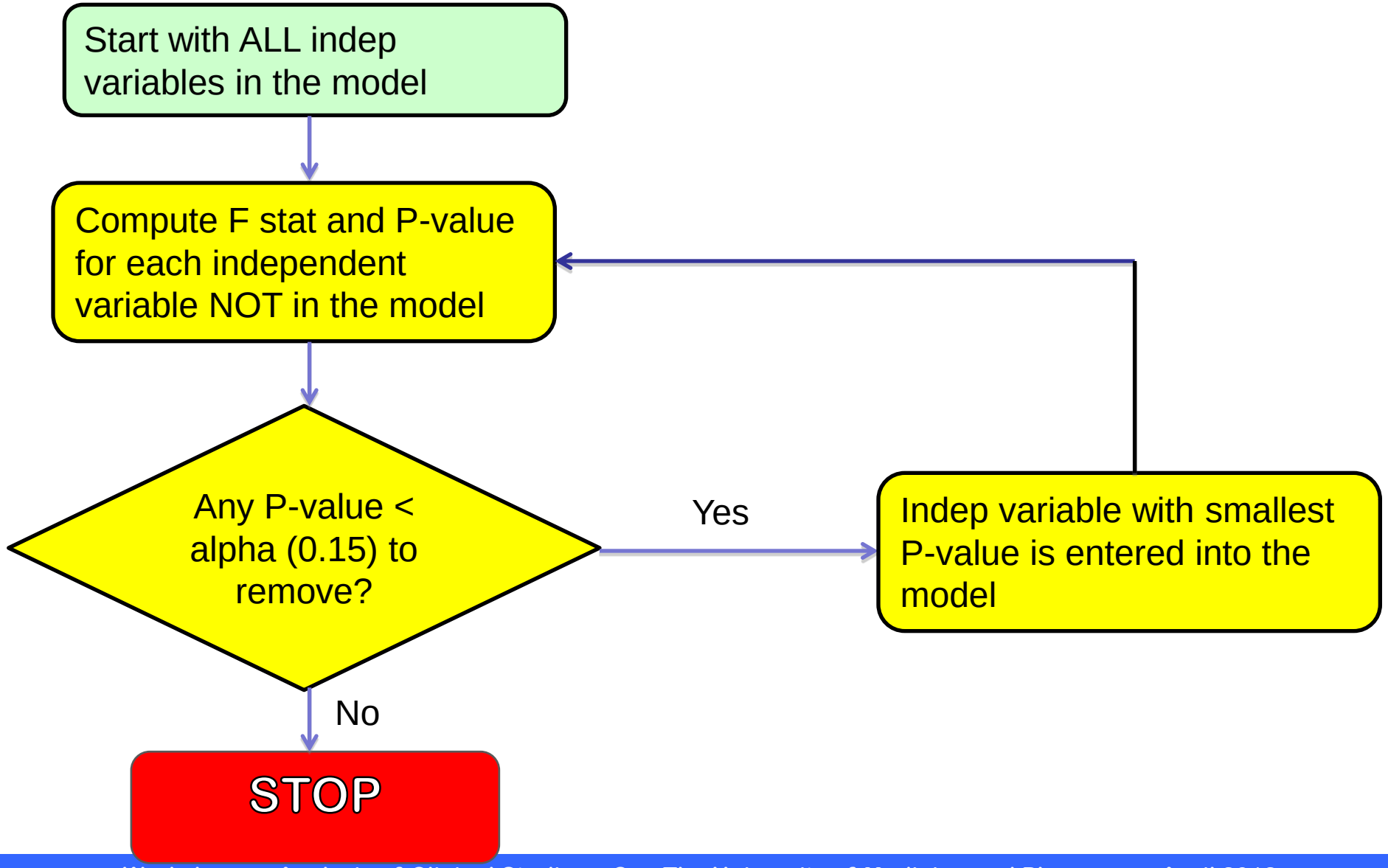
Stepwise regression algorithm



Forward selection algorithm



Backward elemination algorithm



Stepwise selection procedure

- Specify an Alpha-to-Enter significance level. Many software packages set this significance level by default to $\alpha_E = 0.15$.
- Specify an Alpha-to-Remove significance level. Again, many software packages set this significance level by default to $\alpha_R = 0.15$.
- **Step #1**. Once we've specified the starting significance levels, then we
  - Fit each of the one-predictor models — that is, regress y on $x1$, regress y on x_2, ..., and regress y on x_p-1.
  - Of those predictors whose t-test P-value is less than $\alpha_E = 0.15$, the first predictor put in the stepwise model is the predictor that has the smallest t-test P-value.
  - If no predictor has a t-test P-value less than $\alpha_E = 0.15$, stop.

Stepwise selection procedure

- **Step #2. Then:**
  - Suppose x_1 had the smallest *t*-test *P*-value below $\alpha_E = 0.15$ and therefore was deemed the "best" one predictor arising from the the first step.
  - Now, fit each of the two-predictor models that include $x1$ as a predictor — that is, regress y on x_1 and x_2, regress y on x_1 and x_3, ..., and regress y on x_1 and x_p-1.
  - Of those predictors whose *t*-test *P*-value is less than $\alpha_E = 0.15$, the second predictor put in the stepwise model is the predictor that has the smallest *t*-test *P*-value.
  - If no predictor has a *t*-test *P*-value less than $\alpha_E = 0.15$, stop. The model with the one predictor obtained from the first step is your final model.
  - But, suppose instead that x_2 was deemed the "best" second predictor and it is therefore entered into the stepwise model.
  - Now, since x_1 was the first predictor in the model, step back and see if entering x_2 into the stepwise model somehow affected the significance of the x_1 predictor. That is, check the *t*-test *P*-value for testing $\beta_1 = 0$. If the *t*-test *P*-value for $\beta_1 = 0$ has become not significant — that is, the *P*-value is greater than $\alpha_R = 0.15$ — remove x_1 from the stepwise model.

Stepwise selection procedure

- **Step #3. Then:**
  - Suppose both x_1 and x_2 made it into the two-predictor stepwise model.
  - Now, fit each of the three-predictor models that include x_1 and x_2 as predictors — that is, regress y on x_1, x_2, and x_3, regress y on x_1, x_2, and x_4, …, and regress y on x_1, x_2, and x_p-1.
  - Of those predictors whose t-test P-value is less than $\alpha_E = 0.15$, the third predictor put in the stepwise model is the predictor that has the smallest t-test P-value.
  - If no predictor has a t-test P-value less than $\alpha_E = 0.15$, stop. The model containing the two predictors obtained from the second step is your final model.
  - But, suppose instead that x_3 was deemed the "best" third predictor and it is therefore entered into the stepwise model.
  - Now, since x_1 and x_2 were the first predictors in the model, step back and see if entering x_3 into the stepwise model somehow affected the significance of the x_1 and x_2 predictors. That is, check the t-test P-values for testing $\beta_1 = 0$ and $\beta_2 = 0$. If the t-test P-value for either $\beta_1 = 0$ or $\beta_2 = 0$ has become not significant — that is, the P-value is greater than $\alpha_R = 0.15$ — remove the predictor from the stepwise model.

Stepwise selection procedure

- **Stopping the procedure**. Continue the steps as described above until adding an additional predictor does not yield a *t*-test *P*-value below $\alpha E = 0.15$.

Stepwise selection: example

- To starting our stepwise regression procedure, let's set our Alpha-to-Enter significance level at $\alpha E = 0.15$, and let's set our Alpha-to-Remove significance level at $\alpha R = 0.15$. Now, regressing *y* on *x*1, regressing *y* on *x*2, regressing *y* on *x*3, and regressing *y* on *x*4, we obtain:

Predictor	Coef	SE Coef	T	P
Constant	81.479	4.927	16.54	0.000
x1	1.8687	0.5264	**3.55**	**0.005**

Predictor	Coef	SE Coef	T	P
Constant	57.424	8.491	6.76	0.000
x2	0.7891	0.1684	**4.69**	**0.001**

Predictor	Coef	SE Coef	T	P
Constant	110.203	7.948	13.87	0.000
x3	-1.2558	0.5984	**-2.10**	**0.060**

Predictor	Coef	SE Coef	T	P
Constant	117.568	5.262	22.34	0.000
x4	-0.7382	0.1546	**-4.77**	**0.001**

Stepwise selection: example

- Now, following step #2, we fit each of the two-predictor models that include *x*4 as a predictor — that is, we regress *y* on *x*4 and *x*1, regress *y* on *x*4 and *x*2, and regress *y* on *x*4 and *x*3, obtaining:

Predictor	Coef	SE Coef	T	P
Constant	103.097	2.124	48.54	0.000
x4	-0.61395	0.04864	-12.62	0.000
x1	1.4400	0.1384	**10.40**	**0.000**

Predictor	Coef	SE Coef	T	P
Constant	94.16	56.63	1.66	0.127
x4	-0.4569	0.6960	-0.66	0.526
x2	0.3109	0.7486	**0.42**	**0.687**

Predictor	Coef	SE Coef	T	P
Constant	131.282	3.275	40.09	0.000
x4	-0.72460	0.07233	-10.02	0.000
x3	-1.1999	0.1890	**-6.35**	**0.000**

Stepwise selection: example

- Now, following step #3, we fit each of the three-predictor models that include *x*1 and *x*4 as predictors — that is, we regress *y* on *x*4, *x*1, and *x*2; and we regress *y* on *x*4, *x*1, and *x*3, obtaining

Predictor	Coef	SE Coef	T	P
Constant	71.65	14.14	5.07	0.001
x4	-0.2365	0.1733	-1.37	0.205
x1	1.4519	0.1170	12.41	0.000
x2	0.4161	0.1856	**2.24**	**0.052**

Predictor	Coef	SE Coef	T	P
Constant	111.684	4.562	24.48	0.000
x4	-0.64280	0.04454	-14.43	0.000
x1	1.0519	0.2237	4.70	0.001
x3	-0.4100	0.1992	**-2.06**	**0.070**

Stepwise selection: example

- **Now, since *x*1 and *x*4 were the first predictors in the model, we must step back and see if entering *x*2 into the stepwise model affected the significance of the *x*1 and *x*4 predictors. Indeed, it did — the *t*-test *P*-value for testing $\beta 4 = 0$ is 0.205, greater than $\alpha R = 0.15$. Therefore, we remove the predictor *x*4 from the stepwise model, leaving us with the predictors**

Predictor	Coef	SE Coef	T	P
Constant	52.577	2.286	23.00	0.000
x1	1.4683	0.1213	12.10	0.000
x2	0.66225	0.04585	14.44	0.000

Now, we proceed fitting each of the three-predictor models that include *x*1 and *x*2 as predictors — that is, we regress *y* on *x*1, *x*2, and *x*3; and we regress *y* on *x*1, *x*2, and *x*4, obtaining:

Predictor	Coef	SE Coef	T	P
Constant	71.65	14.14	5.07	0.001
x1	1.4519	0.1170	12.41	0.000
x2	0.4161	0.1856	2.24	0.052
x4	-0.2365	0.1733	**-1.37**	**0.205**

Predictor	Coef	SE Coef	T	P
Constant	48.194	3.913	12.32	0.000
x1	1.6959	0.2046	8.29	0.000
x2	0.65691	0.04423	14.85	0.000
x3	0.2500	0.1847	**1.35**	**0.209**

Stepwise selection: example

- Neither of the remaining predictors — *x*3 and *x*4 — are eligible for entry into our stepwise model, because each *t*-test *P*-value — 0.209 and 0.205, respectively — is greater than αE = 0.15. That is, we stop our stepwise regression procedure. Our final regression model, based on the stepwise procedure contains only the predictors

Predictor	Coef	SE Coef	T	P
Constant	52.577	2.286	23.00	0.000
x1	1.4683	0.1213	12.10	0.000
x2	0.66225	0.04585	14.44	0.000

Stepwise selection: example

- **Summary of steps:**

Stepwise Regression: y versus x1, x2, x3, x4

Alpha-to-Enter: 0.15 **Alpha-to-Remove:** 0.15

Response is **y** on 4 predictors, with N = 13

Step	**1**	**2**	**3**	**4**
Constant	117.57	103.10	71.65	52.58
x4	-0.738	-0.614	-0.237	
T-Value	-4.77	-12.62	-1.37	
P-Value	0.001	0.000	0.205	
x1		1.44	1.45	1.47
T-Value		10.40	12.41	12.10
P-Value		0.000	0.000	0.000
x2			0.416	0.662
T-Value			2.24	14.44
P-Value			0.052	0.000
S	8.96	2.73	2.31	2.41
R-Sq	67.45	97.25	98.23	97.87
R-Sq(adj)	64.50	96.70	97.64	97.44
C-p	138.7	5.5	3.0	2.7

Be careful!

- The final model is not guaranteed to be optimal in any specified sense.
- The procedure yields a single final model, although there are often several equally good models.
- Stepwise regression does not take into account a researcher's knowledge about the predictors. It may be necessary to force the procedure to include important predictors.
- One should not over-interpret the order in which predictors are entered into the model.
- One should not jump to the conclusion that all the important predictor variables for predicting y have been identified, or that all the unimportant predictor variables have been eliminated. It is, of course, possible that we may have committed a Type I or Type II error.
- Many t-tests for testing $\beta k = 0$ are conducted in a stepwise regression procedure. The probability is therefore high that we included some unimportant predictors or excluded some important predictors.

All Possible Regressions

All Possible Regressions

- For each subset, define a criterion of "model goodness" which tries to balance over-fitting (model too complex) with under-fitting (model doesn't fit very well).
- Calculate the criterion for each of the 2^k-1 models
- Pick the best one according to the criterion.
- One difficulty: there are several possible criteria, and they don't always agree.

Possible criteria: R^2

- Since R^2 increases as we add more variables, picking the model with the biggest R^2 will always select the model with all the variables. This will often result in overfitting.
- However, R^2 is OK for choosing between models with the same number of variables.
- We need to modify R^2 to penalize overly complicated models. One way is to use the adjusted R^2 (p = number of coefficients in model)

$$\overline{R}_p^2 = 1 - \frac{(n-1)}{(n-p)}(1 - R_p^2)$$

Interpretation

- Suppose we have 2 models: model A with p-1 variables and model B with an additional q variables (we say A is a submodel of B)
- Then the adjusted R^2 is defined so that

$$\overline{R}_p^2 < \overline{R}_{p+q}^2 \text{ if and only if } F > 1$$

where F is the F statistic for testing that model A is adequate.

Residual mean square (RMS)

- Recall the estimate of the error variance σ^2: estimated by s^2=RSS/(n-p), sometimes called the residual mean square (RMS)
- Choose model with the minimum RMS
- We can show that this is equivalent to choosing the model with the biggest adjusted R^2

AIC and BIC

- These are criteria that balance goodness of fit (as measured by RSS) against model complexity (as measured by the number of regression coefficients)

- AIC (Akaike Information Criterion) is, up to a constant depending on n , AIC = n log(RSS_p) + 2p
- Alternative version is AIC = RSS/RMS_{Full} + 2p, equivalent to Cp

- BIC (Bayesian Information Criterion) is

  n log(RSS_p) + p log n

- Small values = good model
- AIC tends to favour more complex models than BIC

Criteria based on prediction error

- Our final set of criteria use an estimate of prediction error to evaluate models
- They measure how well a model predicts *new* data

Mallow's Cp: estimating prediction error

Suppose we have a model with p regression coefficients. "Mallows C_p" provides an estimate of how well the model predicts new data, and is given by

$$Cp = \frac{RSS_p}{RMS_{FULL}} + 2p - n$$

The subscript FULL refers to the "full model" with k variables. Small values of Cp with Cp about p are good.

Warning: C_{k+1}=k+1 always, so don't take this as evidence that the full model is good unless all the other Cp's are bigger.

Mallow's Cp : Interpretation

If the p-coefficient model contains all the important explanatory variables, then RSS_p is about the same as $(n-p)\sigma^2$. Moreover, EMS_{FULL} will also be about the same as σ^2. Thus

$$Cp = \frac{RSS_p}{RMS_{FULL}} + 2p - n$$

$$\approx \frac{(n-p)\sigma^2}{\sigma^2} + 2p - n$$

$$= p$$

Cp plot

- For each model, we plot Cp against p, with the line C_p= p added.
- Points close to this line having small values of C_p correspond to good models.

Estimating prediction error: Cross-validation

- Cp is not a very good estimate of prediction error
- If we have plenty of data, we split the data into 2 parts
  - The "training set", used to fit the model and construct the predictor
  - The "test set", used to estimate the prediction error
- Test set error (=prediction error) estimated by

$$n^{-1} \sum_{test\ set} (y_i - \hat{y}_i)^2$$

Predicted value using training set predictor with new data

- Choose model with smallest prediction error

Estimating prediction error: Cross-validation (2)

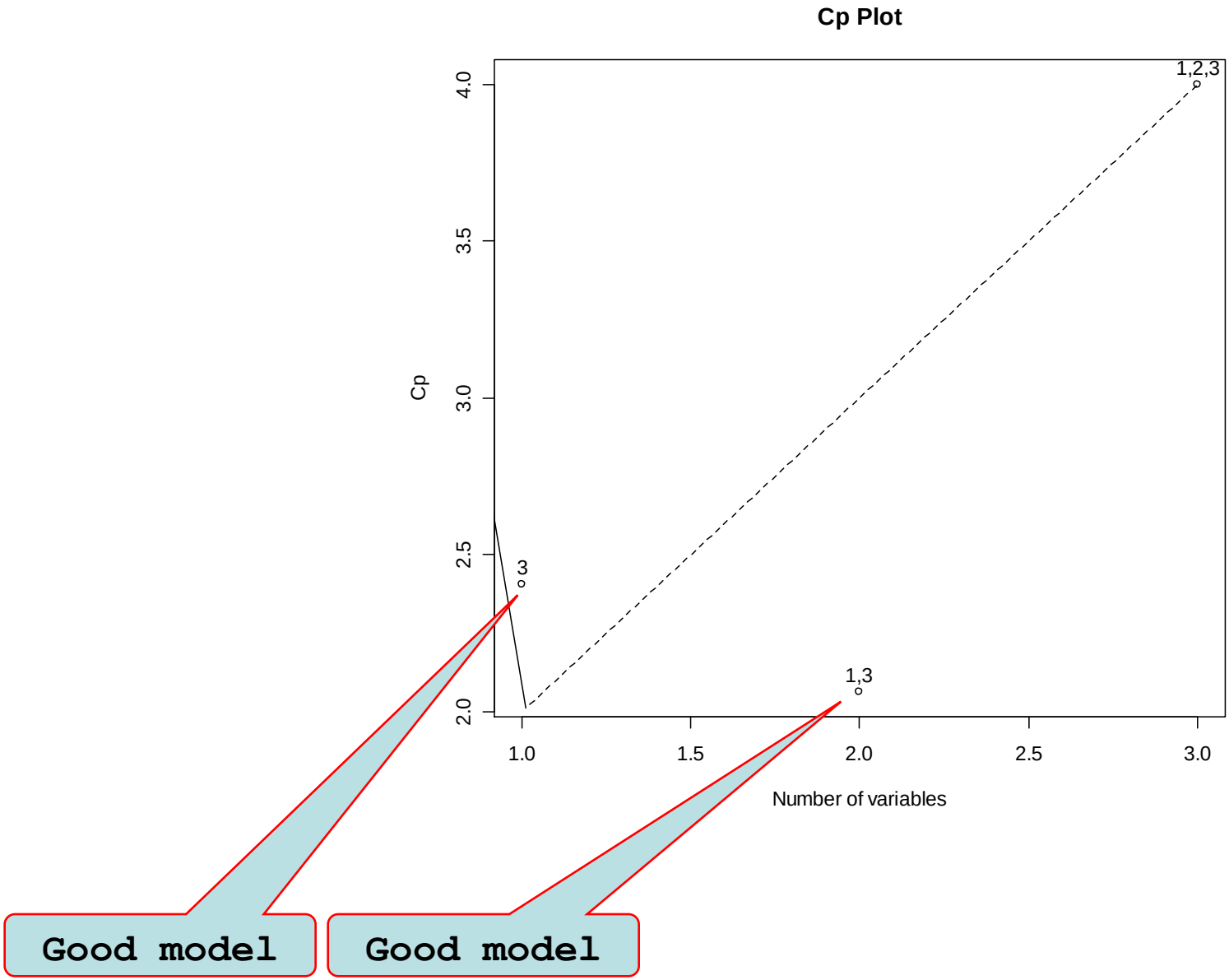
- If we don't have plenty of data, we randomly split the data into 10 parts. One part acts as a test set, the rest as the training set. We compute the prediction error from the test set as before.
- Repeat another 9 times, using a different 10^{th} as the test set each time. Average the estimates to get a good estimate of prediction error
- Repeat for different "random splits"
- This is "10-fold cross-validation". Can do 5-fold, or n-fold, but 10-fold seems to be best.

Example: the fatty acid data

The R function all.poss.regs does the business: eg for the fatty acid data *NB This function requires the package "leaps"*

```
> fatty.lm <- lm(ffa ~ age + skinfold + weight, data = fatty.df)
> library(leaps)
> all.poss.regs(fatty.lm, Cp.plot=T)
```

	rssp	sigma2	adjRsq	Cp	AIC	BIC	CV	age	weight	skinfold
1	0.910	0.051	0.380	2.406	22.406	24.397	0.114	0	1	0
2	0.794	0.047	0.427	2.062	22.062	25.049	0.107	1	1	0
3	0.791	0.049	0.394	4.000	24.000	27.983	0.117	1	1	1


Cp Plot
Cp
4.0
3.5
3.0
2.5
2.0
1.0
1.5
2.0
2.5
3.0
Number of variables
1,2,3
3
1,3
Good model
Good model

Example: the evaporation data

- This was discussed in Tutorial 2: the variables are
  - **evap:** the amount of moisture evaporating from the soil in the 24 hour period (response)
  - **maxst:** maximum soil temperature over the 24 hour period
  - **minst:** minimum soil temperature over the 24 hour period
  - **avst:** average soil temperature over the 24 hour period
  - **maxat:** maximum air temperature over the 24 hour period
  - **minat:** minimum air temperature over the 24 hour period
  - **avat:** average air temperature over the 24 hour period
  - **maxh:** maximum humidity over the 24 hour period
  - **minh:** minimum humidity over the 24 hour period
  - **avh:** average humidity over the 24 hour period
  - **wind:** average wind speed over the 24 hour period.

Variable selection

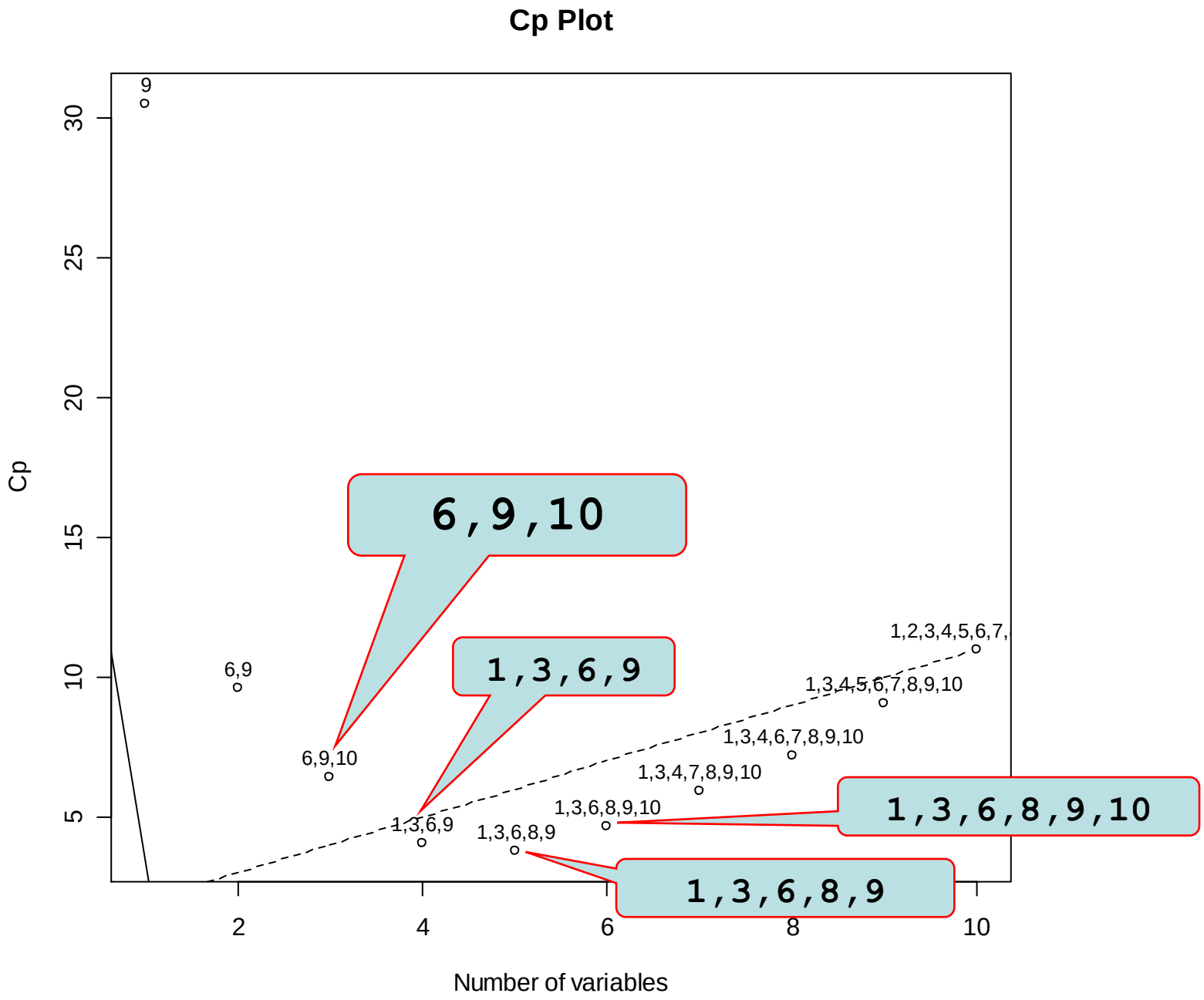
- There are strong relationships between the variables, so we probably don't need them all. We can perform an all possible regressions analysis using the code

```
evap.df = read.table(evap.txt", header=T)
evap.lm = lm(evap~.,data=evap.df)
library(leaps)
all.poss.regs(evap~.,data=evap.df)
```

```
Call:
lm(formula = evap ~ ., data = evap.df)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -54.074877 130.720826  -0.414  0.68164
avst          2.231782   1.003882   2.223  0.03276 *
minst         0.204854   1.104523   0.185  0.85393
maxst        -0.742580   0.349609  -2.124  0.04081 *
avat          0.501055   0.568964   0.881  0.38452
minat         0.304126   0.788877   0.386  0.70219
maxvat        0.092187   0.218054   0.423  0.67505
avh           1.109858   1.133126   0.979  0.33407
minh          0.751405   0.487749   1.541  0.13242
maxh         -0.556292   0.161602  -3.442  0.00151 **
wind          0.008918   0.009167   0.973  0.33733
Residual standard error: 6.508 on 35 degrees of freedom
Multiple R-Squared: 0.8463,     Adjusted R-squared: 0.8023
F-statistic: 19.27 on 10 and 35 DF,  p-value: 2.073e-11
```

```
> library(leaps)  # NB Load leaps library
> all.poss.regs(evap~., data=evap.df)
       rssp sigma2 adjRsq     Cp    AIC    BIC       CV
1  3071.255 69.801  0.674 30.519 76.519 80.177 308.052
2  2101.113 48.863  0.772  9.612 55.612 61.098 208.962
3  1879.949 44.761  0.791  6.390 52.390 59.705 191.622
4  1696.789 41.385  0.807  4.065 50.065 59.208 206.449
5  1599.138 39.978  0.813  3.759 49.759 60.731 223.113
6  1552.033 39.796  0.814  4.647 50.647 63.448 233.692
7  1521.227 40.032  0.813  5.920 51.920 66.549 260.577
8  1490.602 40.287  0.812  7.197 53.197 69.654 271.771
9  1483.733 41.215  0.808  9.034 55.034 73.321 302.781
10 1482.277 42.351  0.802 11.000 57.000 77.115 325.410
    avst minst maxst avat minat maxat avh minh maxh wind
1      0     0     0    0     0     0   0    0    1    0
2      0     0     0    0     0     1   0    0    1    0
3      0     0     0    0     0     1   0    0    1    1
4      1     0     1    0     0     1   0    0    1    0
5      1     0     1    0     0     1   0    1    1    0
6      1     0     1    0     0     1   0    1    1    1
7      1     0     1    1     0     0   1    1    1    1
8      1     0     1    1     0     1   1    1    1    1
9      1     0     1    1     1     1   1    1    1    1
10     1     1     1    1     1     1   1    1    1    1
```


Cp Plot
Cp
Number of variables
9
6,9
6,9,10
1,3,6,9
1,3,6,8,9
1,3,6,8,9,10
1,3,4,7,8,9,10
1,3,4,6,7,8,9,10
1,3,4,5,6,7,8,9,10
1,2,3,4,5,6,7,
6,9,10
1,3,6,9
1,3,6,8,9,10
1,3,6,8,9

```
> sub.lm = lm(evap~avat + avh + wind,data=evap.df)
> summary(sub.lm)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 123.901800  24.624411   5.032 9.60e-06 ***
avat          0.222768   0.059113   3.769 0.000506 ***
avh          -0.342915   0.042776  -8.016 5.31e-10 ***
wind          0.015998   0.007197   2.223 0.031664 *
Residual standard error: 6.69 on 42 degrees of freedom
Multiple R-Squared: 0.805,      Adjusted R-squared: 0.7911
F-statistic:  57.8 on 3 and 42 DF,  p-value: 5.834e-15
```

Full model was 0.8463

Model building strategy

- **Step1: Determine your goal:**
  - For **predictive** reasons — that is, the model will be used to predict the response variable from a chosen set of predictors.
  - For **theoretical** reasons — that is, the researcher wants to estimate a model based on a known theoretical relationship between the response and predictors.
  - For **control** purposes — that is, the model will be used to control a response variable by manipulating the values of the predictor variables.
  - For **inferential** reasons — that is, the model will be used to explore the strength of the relationships between the response and the predictors.
  - For **data summary** reasons — that is, the model will be used merely as a way to summarize a large set of data by a single equation.

Model building strategy

- **Step 2:** Decide which predictor variables and response variable on which to collect the data. Collect the data.
- **Step 3: Exploration of data**
  - On a univariate basis, check for outliers, gross data errors, and missing values.
  - Study bivariate relationships to reveal other outliers, to suggest possible transformations, and to identify possible multicollinearities.

Model building strategy

- **Step 4: Randomly divide the data into a training set and a test set:**
  - The training set, with at least 15-20 error degrees of freedom, is used to estimate the model.
  - The test set is used for cross-validation of the fitted model.
- **Step 5: Using the training set, identify several candidate models:**
  - Use best subsets regression.
  - Use stepwise regression, which of course only yields one model unless different alpha-to-remove and alpha-to-enter values are specified.

Model building strategy

- **Step 6: Select and evaluate a few "good" models:**
  - Select the models based on the four criteria we learned, as well as the number and nature of the predictors.
  - Evaluate the selected models for violation of the model conditions.
  - If none of the models provide a satisfactory fit, try something else, such as collecting more data, identifying different predictors, or formulating a different type of model.
- **Step 7 (final): Select the final model**
  - Compare the competing models by cross-validating them against the test data.
  - The model with a larger cross-validation $R2$ is abetter predictive model.
  - Consider residual plots, outliers, **parsimony**, relevance, and ease of measurement of predictors.