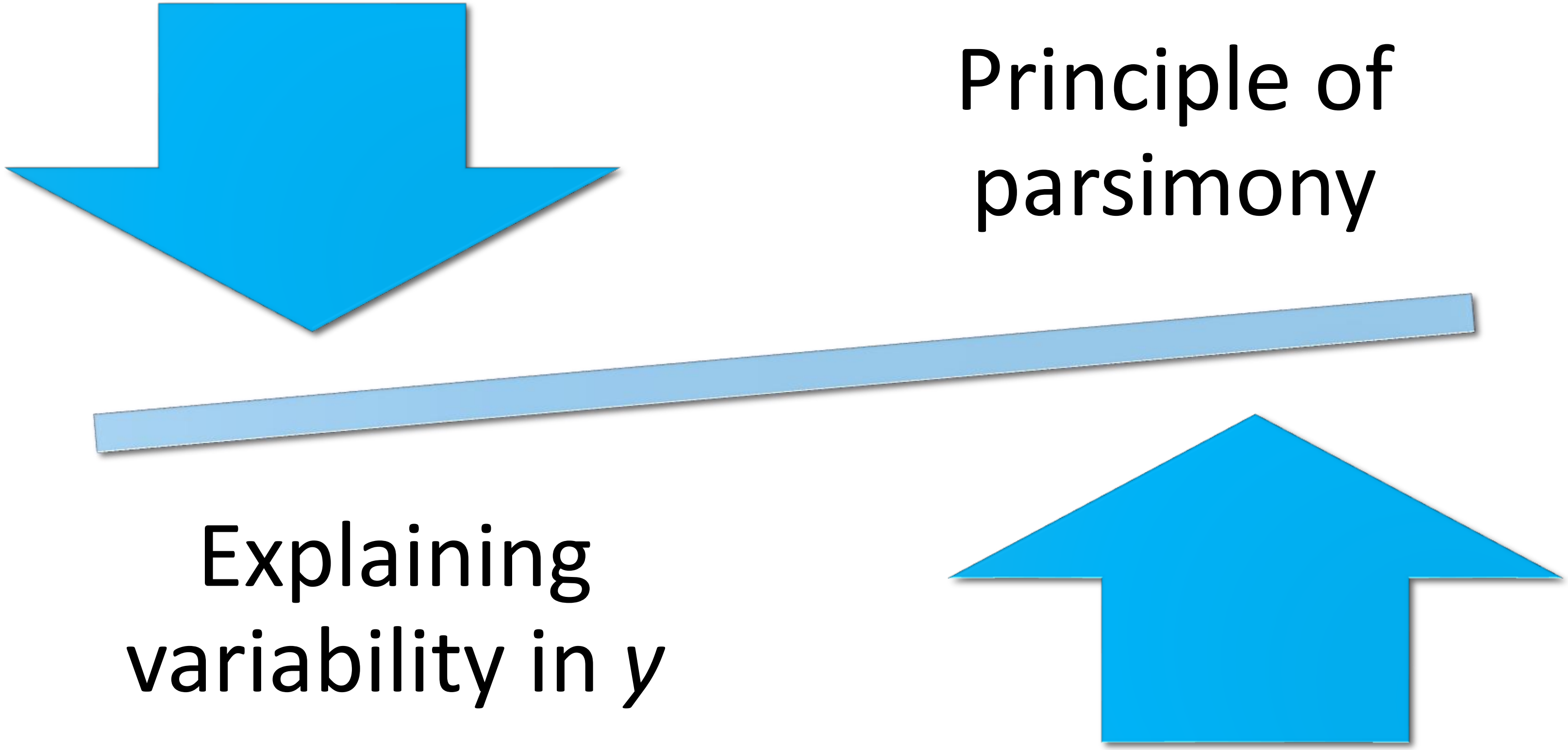
Problems fitting Multiple Linear Regression Models

CHRIS SUTTON, MARCH 2024

Re-cap of automated approaches to model building from last week

Conflicting objectives



Principle of parsimony

Explaining variability in *y*

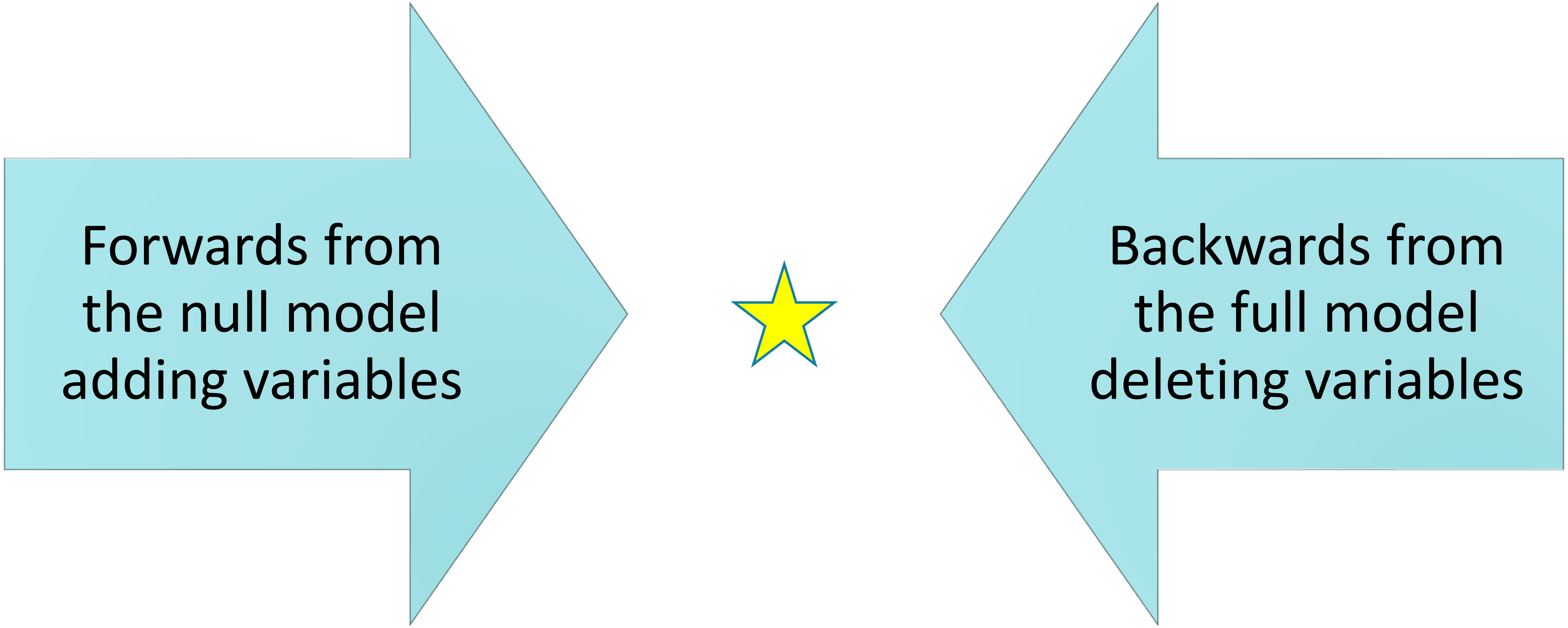
Automatic methods

In response to the challenge of larger number of explanatory variables, a number of so-called *Automatic* regression model selection procedures have been devised

They all have advantages and disadvantages

They generally involve a sequence of statistical tests

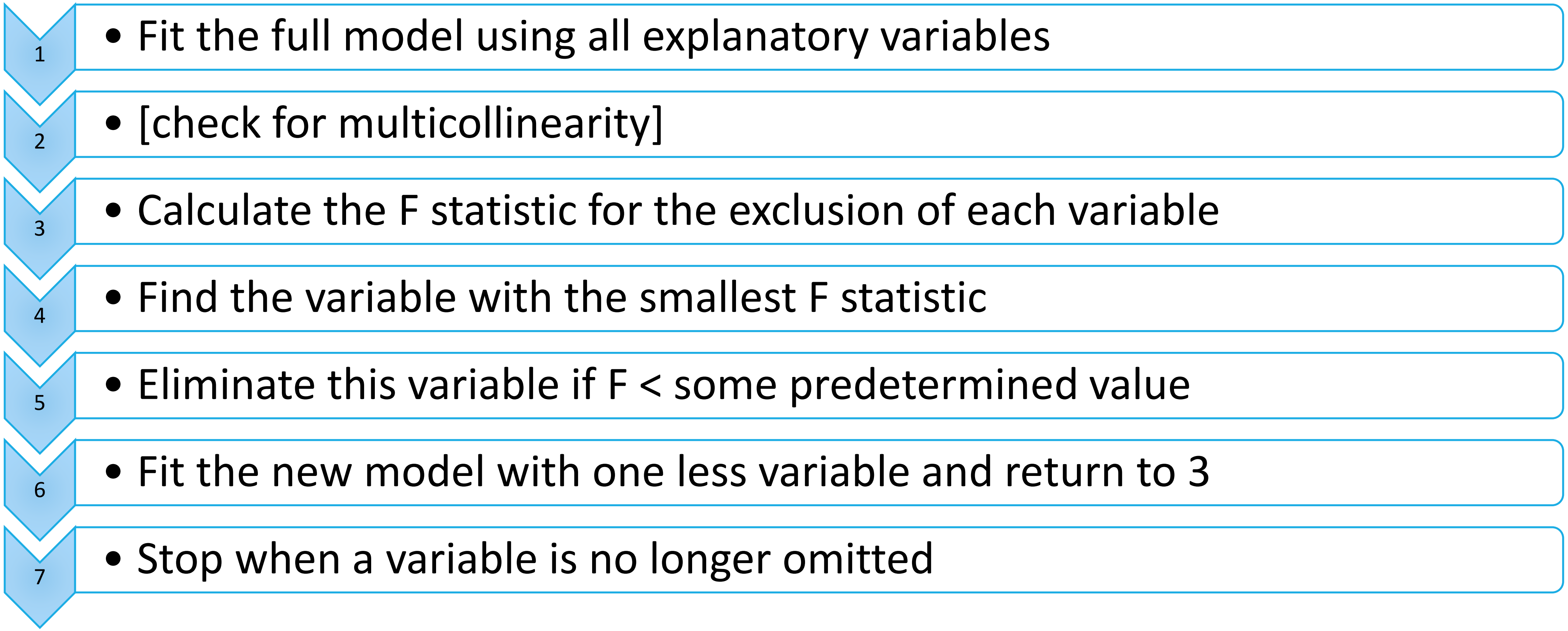
Two routes to chosen model



Forwards from the null model adding variables

Backwards from the full model deleting variables

Backward Elimination



1
- Fit the full model using all explanatory variables

2
- [check for multicollinearity]

3
- Calculate the F statistic for the exclusion of each variable

4
- Find the variable with the smallest F statistic

5
- Eliminate this variable if F < some predetermined value

6
- Fit the new model with one less variable and return to 3

7
- Stop when a variable is no longer omitted

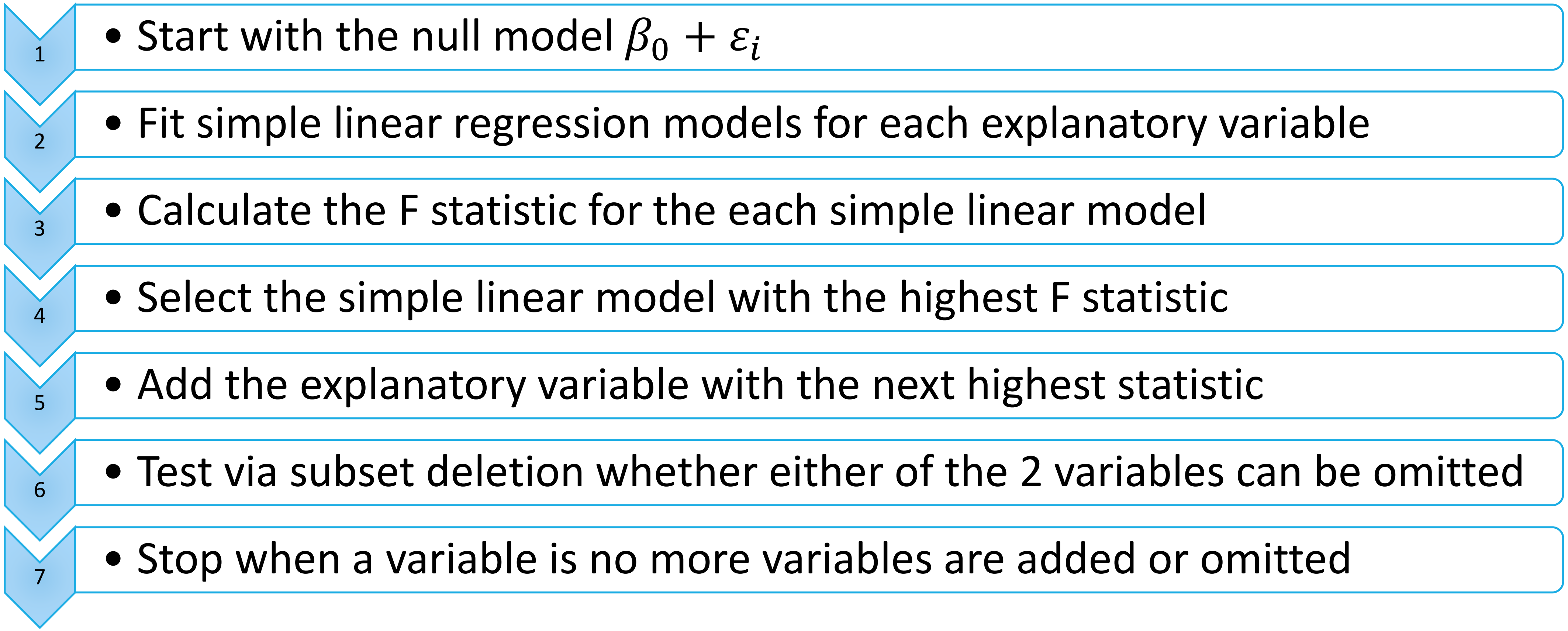
Key disadvantage - Multicollinearity

We will take a closer look at this problem in regression later

The temptation with Backward Elimination is to start with as many explanatory variables as possible thinking insignificant ones are sure to be deleted

As the number of variables increases and we risk including variables that are essentially themselves linear combinations of other variables in the model

Stepwise Regression



1

- Start with the null model $\beta_0 + \varepsilon_i$

2

- Fit simple linear regression models for each explanatory variable

3

- Calculate the F statistic for the each simple linear model

4

- Select the simple linear model with the highest F statistic

5

- Add the explanatory variable with the next highest statistic

6

- Test via subset deletion whether either of the 2 variables can be omitted

7

- Stop when a variable is no more variables are added or omitted

Key disadvantage – Variance estimate

We need to estimate σ^2 in the F tests

Our usual method is to use $\mathrm{MS_E}$

But starting with simpler models, $\mathrm{MS_E}$ is likely to be higher in the early rounds of stepwise regression and then fall as more variables are added

This distorts early round F tests compared to later ones

- So the bar for adding or deleting variables is not the same across the process

One work around is to use full model $\mathrm{MS_E}$ in all the F tests including considering the first simple linear regression models

False positives in F tests

One of the main issues with automatic methods that rely on F (or t) tests

Risk that we fail to reject H_0: $\beta_j = 0$ for some j when we should have rejected it

Means we would include (or fail to eliminate) an explanatory variable whose parameter value was really zero (and therefore a variable with no statistically significant explanatory power)

Akaike’s Information Criterion (AIC) can help with this

AIC uses some of the Maximum Likelihood methods of week 6

Akaike’s Information Criteria (AIC)

$$AIC = 2(p+1) - 2logL$$

Where:

- p = the number of regression parameters (so p – 1 explanatory variables)
- L is the Likelihood function evaluated at the maximum likelihood estimates of each of the parameters

Using AIC in model selection

We seek the regression model that minimises AIC

- because of the $-2logL$ this is equivalent to the model that maximises likelihood balanced against the number of parameters

- We can look to do this through backward or forward type processes

Backwards Elimination using AIC

1. Construct the full model and calculate its AIC
2. Construct all possible models that omit one variable
3. Calculate the AIC for each of these models
4. If the full model has lowest AIC use that and stop
5. If another has lowest AIC move on to that model and repeat
6. Stop once AIC cannot be lowered by removing a variable

Backwards using AIC in R

This process can be automated in R programming using the `step()` function

If the full model is constructed with `lm()` and stored as `full_model` (say)

A backwards route to the reduced model with lowest AIC is given by

```
reduced_model <- step(full_model, direction = “backward”)
```

Forward regression alterative using AIC

We start with the null model

In R this is found by `null_model <- lm(y ~ 1)`

say we have six explanatory variables to consider `x1 x2 x3 x4 x5 x6`

This is done in R with

```
 forward_model <- step(null_model, scope =
x1+x2+x3+x4+x5+x6, direction = “forward”)
```

Different results

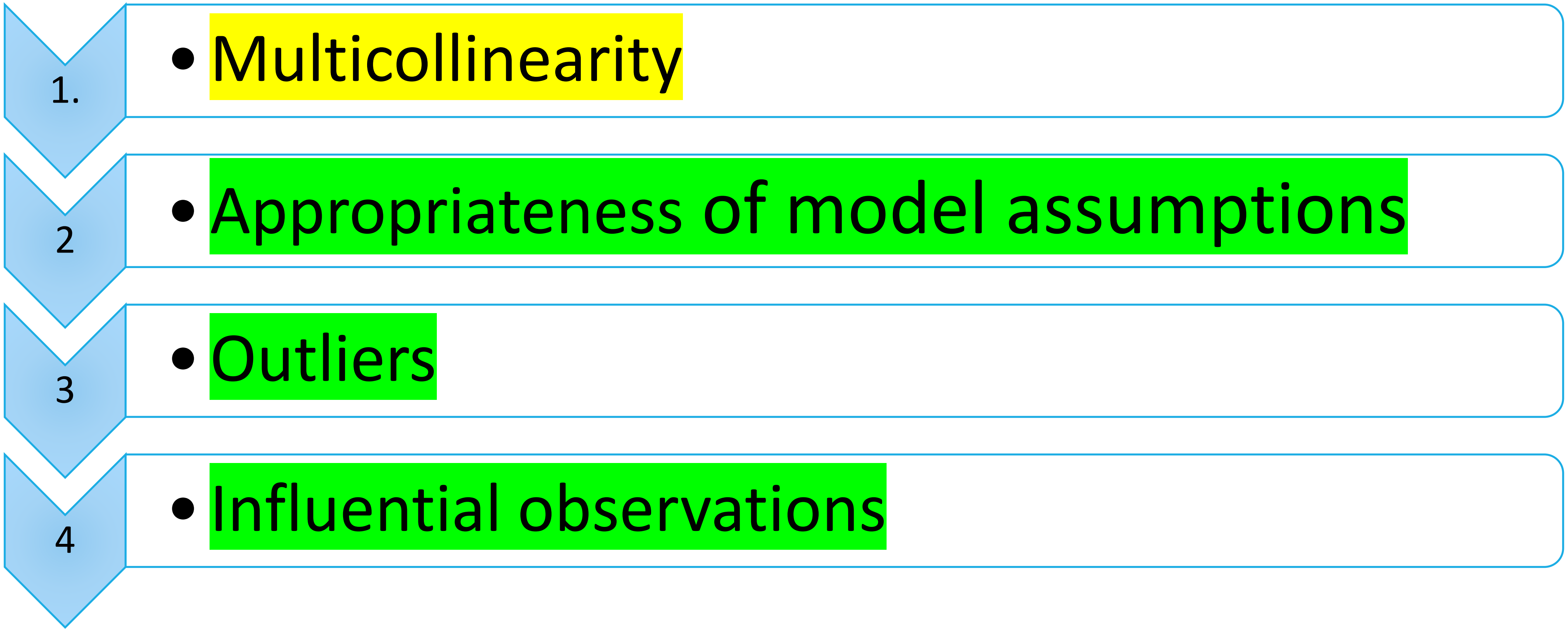
Backwards and Forwards methods using AIC may lead to different recommended models

A third alternative is to set `direction = “both”` inside `step()`

- This has the effect of adding additional variables from the null model and then later allowing deletion of one or more of those variables once others are added

Problems fitting multiple linear regression models

Potential problems to identify and overcome (3 old ones and 1 new)



1.
- Multicollinearity

2
- Appropriateness of model assumptions

3
- Outliers

4
- Influential observations

Multicollinearity

Multicollinearity

- We mentioned this problem before with automated model building techniques, particularly Backwards Elimination procedures

- Becomes more likely to occur when we have a large number of explanatory variables

❑ How do we know whether two or more of these variables are measuring the same effect?

❑ Does it matter if they are?

Singular X^TX

Mathematically the problem arises when we have singular $\boldsymbol{X^TX}$ matrix

- then its determinant is zero
- so it cannot be inverted
- and we cannot find a unique solution to the *Normal Equations*
- and therefore no unique least squares beta estimates

When does this occur?

The problem of singularity occurs when there is linear dependence between two or more of the explanatory variables

- when two variables are equal
- when one variable is a linear combination of other variables

e.g. consider $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

Extreme case $x_{1i} = \ x_{2i} = \ 1$ all i

Then $\boldsymbol{X}$ = $\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}$

And multiplying out $\boldsymbol{X}^T\boldsymbol{X}$ = $\begin{pmatrix} 3 & 3 & 3 \\ 3 & 3 & 3 \\ 3 & 3 & 3 \end{pmatrix}$

So det($\boldsymbol{X}^T\boldsymbol{X}$) = 0 meaning ($\boldsymbol{X}^T\boldsymbol{X}$) cannot be inverted and the $\widehat{\beta}$ found

Two variables equal

If $\boldsymbol{X} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 3 & 3 \end{pmatrix}$

then again $\det(\boldsymbol{X}^T\boldsymbol{X}) = 0$

and again we cannot solve the normal equations

Variable linear combination of others

e.g. if $x_{2i} = 2x_{1i}$

X = $\begin{pmatrix} 1 & 2 & 4 \\ 1 & 2 & 4 \\ 1 & 2 & 4 \end{pmatrix}$

then again det(***X**^T**X***) = 0

But these types of situations ought to be quite easy to identify and eliminate

The more usual problem in modelling

In modelling scenarios a more common situation is where one explanatory variable's observation set is very close to that of another (or a linear transformation of other variables)

Then the determinant of $\boldsymbol{X}^T\boldsymbol{X}$ will be close to (but not equal to) zero

Example $\boldsymbol{X} = \begin{pmatrix} 1 & 0.95 & 1.04 \\ 1 & -0.98 & -1.01 \\ 1 & 1.03 & 0.96 \end{pmatrix}$ then $\det(\boldsymbol{X}^T\boldsymbol{X}) = 0.1014$

Example (continued)

det($\boldsymbol{X}^T\boldsymbol{X}$) = 0.1014

($\boldsymbol{X}^T\boldsymbol{X}$)$^{-1}$ = $\begin{pmatrix} 78.23 & 0.13 & -78.04 \\ 0.12 & 0.38 & -0.01 \\ -78.04 & -0.01 & 78.23 \end{pmatrix}$

we know that $Var(\hat{\beta}_j) = \sigma^2 c_{jj}$ where the c_{jj} from the diagonal of ($\boldsymbol{X}^T\boldsymbol{X}$)$^{-1}$

$Var(\hat{\beta}_2) = 78.23\,\sigma^2$

which is very large in comparison to σ^2

The real problem with multicollinearity

- Parameters with very large variances
- especially where one column of ***X*** close to a linear combination of other columns
- Can even lead to a parameter estimate with the wrong sign
  - β_i should be positive (Y increases as x_{ij}increases)
  - But $\widehat{\beta_i}$ is negative because of multicollinearity and a large estimation variance
  - or vice versa
- Multicollinearity can be difficult to spot by examining the data if we have a large number of explanatory variables

Variance Inflation Factor (VIF)

- one way of detecting multicollinearity

if our p – 1 variable model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{p-1,i} + \varepsilon_i$$

- we calculate VIF for each x_j j = 1, 2, … p – 1
- perform a multiple linear regression of x_j against the p – 2 other $x's$
- which will have a different set of β’s to the original model for y_i
- calculate the R^2 as [0,1] not a % for this new regression and call it R_j^2

Variance Inflation Factor

Then

$$VIF_j = \frac{1}{1 - R_j^2}$$

- High R_j^2 indicates a strong linear relationship between x_j and the other $x's$
- Which results in a large VIF for x_j
- We usually take VIF > 10 as indication of a multicollinearity problem
- We would need to reduce the set of explanatory variables to remove linear combinations