# STA302 METHODS OF DATA ANALYSIS 1

## MODULE 9: MODEL SELECTION TOOLS

PROF. KATHERINE DAIGNAULT

# MODULE 9 OUTLINE

1. Numerical Measures of Goodness
2. All Possible Subsets Selection
3. Automated Selection Methods
4. Cautions for Automated Selection Tools

# MODEL SELECTION MOTIVATION

- We've encountered idea of "comparing models" in previous modules
  - Partial F Test for a subset of predictors
  - Adjusted $R^2$ for models of different sizes
- These are part of a group of tools used for model selection
- Used to help determine "best" model for a given purpose
  - more predictors $\rightarrow$ predictions with low bias but high variance
  - too many predictors $\rightarrow$ over-fitted model
- Need to consider purpose of model:
  - Prediction: extra predictors help explain more variation, better accuracy if not over-fitted
  - Description: too many predictors or complicated transformations hurt interpretability
  - In both, some model selection will need to occur.
- Additional measures can be used to help determine a preferred model.
  - These are called likelihood-based measures of goodness
  - Used in conjunction with diagnostics and assessing assumptions to select "best" model.

# REVIEW OF LIKELIHOODS

- Maximum likelihood is a method used to determine estimators of parameters.
- Likelihood is a function of the parameters given the observed data
  - it is a probability distribution that is maximized to find estimators.
- Depends on Normality assumption:
  - $y_i | x_{i1}, \ldots, x_{ip} \sim N(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \sigma^2)$
- Assuming independence/uncorrelated errors, likelihood is product of $n$ such Normals:

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - \{\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}\}\right]^2\right)$$

- Natural log of likelihood easier to work with:

$$\ln(L(\boldsymbol{\beta}, \sigma^2|Y)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

RSS if $\beta_j$'s replaced with $\hat{\beta}_j$'s

- If we replace $\sigma^2$ with the MLE $\hat{\sigma}^2_{MLE} = {}^{RSS}/_{n}$, we simplify to:

$$\ln(L(\widehat{\boldsymbol{\beta}}, \hat{\sigma}^2\,|Y)) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}ln\left(\frac{RSS}{n}\right) - \frac{n}{2}$$

- Since we want a model with RSS as small as possible, this means the log-likelihood will also be small
- Can use this to create measures of goodness.

# 3 LIKELIHOOD MEASURES

- Like $R^2$, log-likelihood only measures goodness without accounting for effect of additional predictors
- The 3 likelihood criteria introduce a penalty term
  - but each criteria penalizes differently
- Akaike's Information Criteria (AIC):

$$AIC = -2\left[\ln\left(L\left(\widehat{\boldsymbol{\beta}}, \hat{\sigma}^2 \,\middle|\, \boldsymbol{Y}\right)\right) - (p+2)\right] \propto n\, ln\left(\frac{RSS}{n}\right) + 2p$$

- Goodness: $n\, ln\left(\frac{RSS}{n}\right)$ where smaller is better
- Penalty for complexity: $2p$ to control for added X's
- A smaller AIC indicates a better model.

- When $n$ is small or $p$ is large fraction of $n$ (i.e., $n/(p+2) \leq 40$), a corrected AIC is used:

$$AIC_c = AIC + \frac{2(p+2)(p+3)}{n-p-1}$$

- Once again, smaller $AIC_c$ indicates better model.
- Finally, a Bayesian Information Criteria (BIC) uses a harsher penalty than AIC to favour simpler models:

$$BIC = -2\ln\left(L\left(\widehat{\boldsymbol{\beta}}, \hat{\sigma}^2 \,\middle|\, \boldsymbol{Y}\right)\right) + (p+2)\ln(n)$$
$$\propto n\, ln\left(\frac{RSS}{n}\right) + (p+2)\ln(n)$$

- As always, smaller BIC indicates a better model.

# USING LIKELIHOOD MEASURES FOR SELECTION

- Each criteria measures goodness with a different penalty for unnecessary complexity.
  - so, it's not reasonable to expect them all to always agree on the "best" model.
- For complete picture, consider all measures:
  - Adjusted $R^2$: look for largest or close to largest value
  - AIC, corrected AIC, BIC: look for smallest or close to smallest value
- Looking at models that are close to having smallest/largest values also helpful
  - if difference is very small, can opt for either option
- All measures depend on model assumptions holding
  - consider the assumptions of each model when using measures to explain slight differences or to select "suboptimal" model.
- Consider context of data and research question too
  - literature already highlights possible important variables that should possibly remain in "best" model
  - avoid accidently removing predictor(s) of interest even if measures indicate model without is preferred
- Diagnostics and measures of goodness should be used together to select preferred model.

# EXAMPLE BY HAND & IN R

```
> model <- lm(Defective ~ Temperature + Density + Rate, data=d)
> summary(model)

Call:
lm(formula = Defective ~ Temperature + Density + Rate, data = d)

Residuals:
     Min       1Q   Median       3Q      Max 
-12.7367  -4.1116  -0.5755   2.7617  16.3279 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  10.3244    65.9265   0.157   0.8768  
Temperature  16.0779     8.2941   1.938   0.0635 .
Density      -1.8273     1.4971  -1.221   0.2332  
Rate          0.1167     0.1306   0.894   0.3797  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.11 on 26 degrees of freedom
Multiple R-squared:  0.8797,	Adjusted R-squared:  0.8658 
F-statistic: 63.36 on 3 and 26 DF,  p-value: 4.371e-12
```

```
> model2 <- lm(Defective ~ Temperature, data=d)
> summary(model2)

Call:
lm(formula = Defective ~ Temperature, data = d)

Residuals:
     Min       1Q   Median       3Q      Max 
-18.5952  -4.9203  -0.6253   4.2133  15.1861 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -40.938      5.298  -7.727 2.04e-08 ***
Temperature   30.904      2.327  13.279 1.32e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 7.312 on 28 degrees of freedom
Multiple R-squared:  0.863,	Adjusted R-squared:  0.8581 
F-statistic: 176.3 on 1 and 28 DF,  p-value: 1.317e-13
```

```
> p = length(coef(model))-1
> n=nrow(d)
> cbind(summary(model)$adj.r.squared, extractAIC(model, k=2)[2],
+       extractAIC(model, k=log(n))[2],
+       extractAIC(model, k=2)[2]+ (2*(p+2)*(p+3)/(n-p-1)))
          [,1]     [,2]     [,3]     [,4]
[1,] 0.8657897 121.3989 127.0036 123.7066
>
```

BIC

AIC

AICc

for simple model:

```
> p = length(coef(model2))-1
> n=nrow(d)
> cbind(summary(model2)$adj.r.squared, extractAIC(model2, k=2)[2],
+       extractAIC(model2, k=log(n))[2],
+       extractAIC(model2, k=2)[2]+ (2*(p+2)*(p+3)/(n-p-1)))
          [,1]     [,2]     [,3]     [,4]
[1,] 0.8580803 121.2977 124.1001 122.1548
```

likelihood criteria all smaller so simple model preferred

$$RSS = 26 \times 7.11^2 = 1314.35$$
$$p = 3; n = 26 + 4 = 30$$

$$AIC = n ln\left(\frac{RSS}{n}\right) + 2p = 30\ ln\left(\frac{1314.35}{30}\right) + 2(3) = 119.40$$

$$AIC = 30 ln\left(\frac{28 \times 7.312^2}{30}\right) + 2(1) = \boxed{119.30}$$

smaller, so preferred model

# MODULE 9 OUTLINE

1. Numerical Measures of Goodness
2. All Possible Subsets Selection
3. Automated Selection Methods
4. Cautions for Automated Selection Tools

# COMPARING SETS OF MODELS

- The four main measures to compare models are:
  - Adjusted $R^2$: $R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$
  - AIC: $AIC \propto n\, ln\left(\frac{RSS}{n}\right) + 2p$
  - Corrected AIC: $AIC_c \propto n\, ln\left(\frac{RSS}{n}\right) + 2p + \frac{2(p+2)(p+3)}{n-p-1}$
  - BIC: $BIC \propto n\, ln\left(\frac{RSS}{n}\right) + (p+2)\ln(n)$
- The RSS term drives the idea of goodness, while the rest are penalties for number of predictors.
  - we know that models with more predictors automatically have lower RSS even if not useful predictors
- Only a problem for comparing models with different number of predictors
  - if $p$ is fixed between models, then only RSS changes so we'd prefer model with smallest RSS.
  - e.g., for all possible models with 3 predictors, the best one explains the most variation and so has smallest RSS
  - all four criteria would agree on the preferred model
- Criteria are then most helpful in comparing models with different $p$.
- Idea of comparing all models of the same size to each other is the root of our next tool.

# ALL POSSIBLE SUBSETS METHOD OF MODEL SELECTION

All possible subsets works in two steps:

1. Compare models of each size using adjusted $R^2$
2. Use all four numerical criteria to pick the best of the best $(R^2_{adj}, AIC, AIC_c, BIC)$

**All one-predictor models**

- $\boxed{X_1 \text{ has } R^2_{adj} = 0.84}$
- $X_2$ has $R^2_{adj} = 0.8$
- $X_3$ has $R^2_{adj} = 0.83$
- $X_4$ has $R^2_{adj} = 0.75$

**All two-predictor models**

- $X_1, X_2$ has $R^2_{adj} = 0.87$
- $X_1, X_3$ has $R^2_{adj} = 0.88$
- $X_1, X_4$ has $R^2_{adj} = 0.86$
- $\boxed{X_2, X_3 \text{ has } R^2_{adj} = 0.89}$
- $X_2, X_4$ has $R^2_{adj} = 0.81$
- $X_3, X_4$ has $R^2_{adj} = 0.84$

**All three-predictor models**

- $X_1, X_2, X_3$ has $R^2_{adj} = 0.87$
- $X_1, X_2, X_4$ has $R^2_{adj} = 0.88$
- $\boxed{X_1, X_3, X_4 \text{ has } R^2_{adj} = 0.89}$
- $X_2, X_3, X_4$ has $R^2_{adj} = 0.85$

**All four-predictor models**

- $\boxed{X_1, X_2, X_3, X_4 \text{ has } R^2_{adj} = 0.9}$

# EXAMPLE IN R

```
> #install.packages("leaps")
> library(leaps)
>
> best <- regsubsets(Defective ~ Temperature + Density + Rate, data=d,
+                    nbest = 1, nvmax=3)
```

load new library

maximum # predictors

how many best models in each subset

```
> summary(best)
Subset selection object
Call: regsubsets.formula(Defective ~ Temperature + Density + Rate,
    data = d, nbest = 1, nvmax = 3)
3 Variables  (and intercept)
            Forced in Forced out
Temperature     FALSE      FALSE
Density         FALSE      FALSE
Rate            FALSE      FALSE
1 subsets of each size up to 3
Selection Algorithm: exhaustive
         Temperature Density Rate
1  ( 1 ) "*"         " "     " "
2  ( 1 ) "*"         "*"     " "
3  ( 1 ) "*"         "*"     "*"
```

nothing forced to be excluded or included

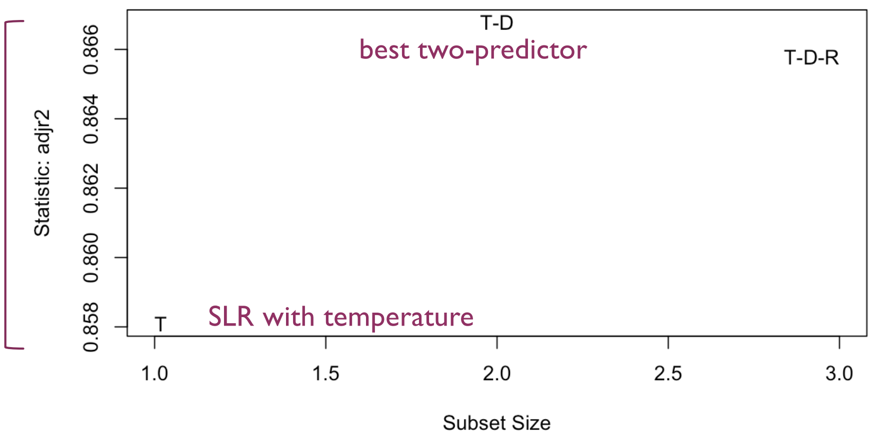
best SLR uses Temperature; best two-predictor

```
> # install car package
> #packageurl <- "https://cran.r-project.org/src/contrib/Archive/pbkrtest/pbkrtest_0.4-4.tar.gz"
> #install.packages(packageurl, repos=NULL, type="source")
> #install.packages("car", dependencies=TRUE)
> library(car)
>
> subsets(best, statistic = "adjr2", legend=FALSE)
```

load the car library

plot each best model against the $R^2_{adj}$

see how different the $R^2_{adj}$ is between best models



# PROS & CONS OF ALL POSSIBLE SUBSETS

Pros:

- All possible models are fit and compared
  - gives you the opportunity to investigate the best ones more directly
- Can be flexible in how we define "best" (e.g., maybe picking best 2 from each subset)
  - allows overall more flexibility in how you select your overall preference
- Nice systematic way to select possible "best" models
  - allows you to compare using many measures while gaining efficiencies by removing unneeded models

Cons:

- Can be very impractical for large numbers of predictors
  - means you still need to compare at least $p$ different models
- Best of each subset does not consider model issues
  - can perform this method without checking model assumptions
  - doesn't account for effect of multicollinearity or problematic observations
  - may mean decisions are not reliable unless manually check

# MODULE 9 OUTLINE

1. Numerical Measures of Goodness
2. All Possible Subsets Selection
3. Automated Selection Methods
4. Cautions for Automated Selection Tools

# WHAT ARE AUTOMATED SELECTION METHODS?

- All possible subsets is cumbersome but provides a list of models
- Automated selection methods is another tool for model selection.
  - sequentially explores models from a specific starting model
  - uses AIC or BIC instead of $R^2_{adj}$ to decide between models
- Three different automated selection methods:
  - forward selection
  - backward selection
  - stepwise selection
- Differ in the order in which models are created:
  - forward: start with intercept model and add predictors
  - backward: start at full model and delete predictors
  - stepwise: iterate between forward and backward
- At each step, model from previous step is taken and each possible predictor available is added/deleted
  - AIC or BIC computed for each, and smallest value is chosen
  - chosen model becomes the starting model for next step
  - once no smaller AIC/BIC value is obtained, last model is the preferred model

# FORWARD & BACKWARDS SELECTION PROCEDURES

Start with $\hat{y} = \hat{\beta}_0$, with e.g., AIC=41

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$ with AIC=40

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_2$ with AIC=42

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_3$ with AIC=44

$\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_4 \text{ with AIC=38}}$

Starting with $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_4$ with AIC=38

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_4 + \hat{\beta}_2 x_1$ (AIC=36)

$\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_4 + \hat{\beta}_2 x_2 \text{ (AIC=35)}}$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_4 + \hat{\beta}_2 x_3$ (AIC=37)

Starting with $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_4 + \hat{\beta}_2 x_2$ with AIC=35

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_4 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1$ (AIC=37)

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_4 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$ (AIC=36)

End process and keep $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_4 + \hat{\beta}_2 x_2$ with AIC=35

Start with $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$ with AIC=43

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$ (AIC=45)

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_4 x_4$ (AIC=44)

$\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 \text{ (AIC=42)}}$

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$ (AIC=43)

Starting with $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$ (AIC=42)

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_3 x_3$ (AIC=44)

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_4 x_4$ (AIC=46)

$\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 \text{ (AIC=41)}}$

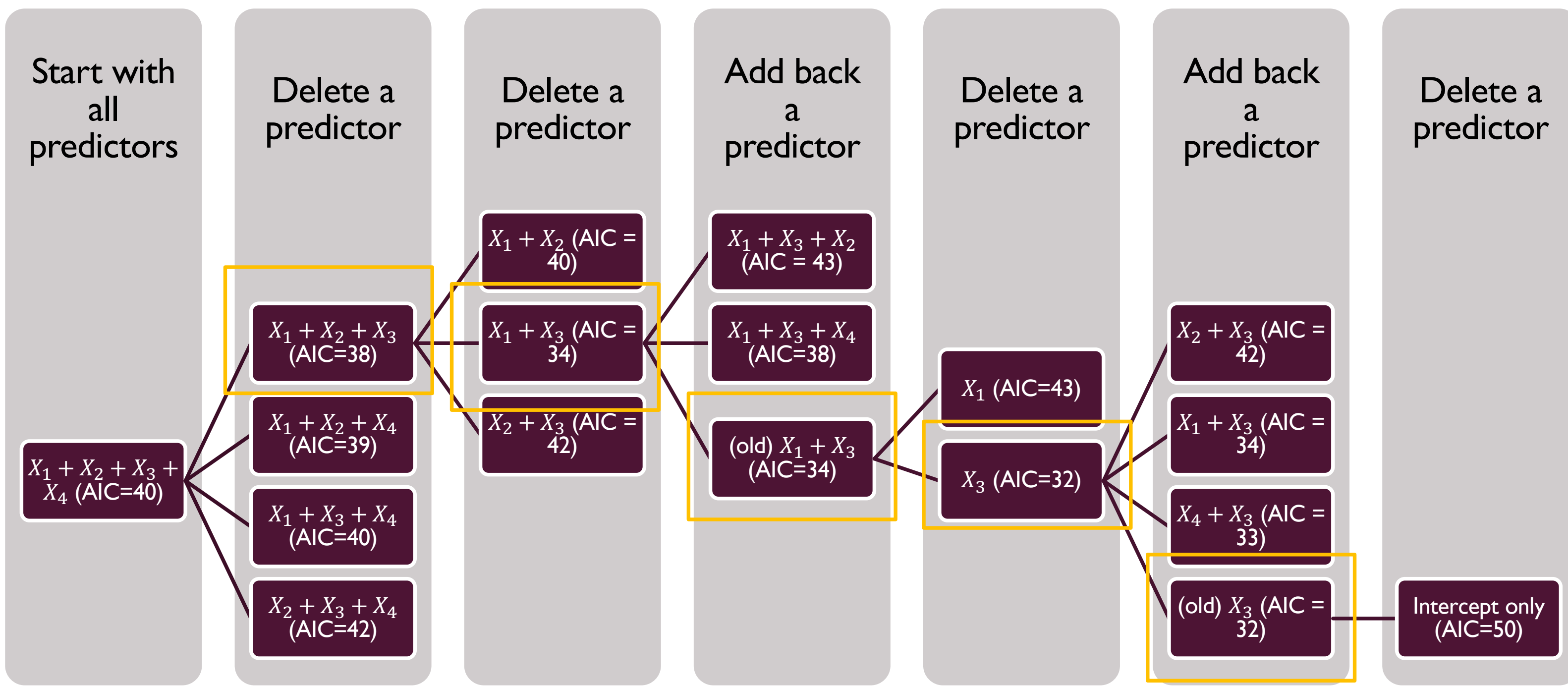
Starting with $\hat{y} = \hat{\beta}_0 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$ (AIC=41)

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_3 x_3$ (AIC=45)

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_4 x_4$ (AIC=44)

End process and keep $\hat{y} = \hat{\beta}_0 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4$ (AIC=41)

# STEPWISE SELECTION PROCEDURE



# EXAMPLE OF FORWARD AND BACKWARDS SELECTION

- Use same function (stepAIC)
- Specify starting model
  - intercept for forward, full for backward
- Add end point
  - upper = full model in forward
  - lower = intercept in backward
- specify direction: "forward" or "backward"
- set k=2 for AIC or k=log(n) for BIC

short for full model

```
> library(MASS)
>
> stepAIC(lm(Defective ~ 1, data=d[,-1]),
+         scope=list(upper=lm(Defective ~ ., data=d[,-1])),
+         direction = "forward", k=2)

Start:  AIC=178.93
Defective ~ 1

              Df Sum of Sq     RSS    AIC
+ Temperature  1    9427.0  1496.9 121.30
+ Density      1    9311.4  1612.5 123.53
+ Rate         1    8559.6  2364.3 135.01
<none>                     10923.9 178.93

Step:  AIC=121.3
Defective ~ Temperature

          Df Sum of Sq    RSS    AIC
+ Density  1    142.05 1354.8 120.31
+ Rate     1    107.10 1389.8 121.07
<none>                 1496.8 121.30

Step:  AIC=120.31
Defective ~ Temperature + Density

       Df Sum of Sq    RSS    AIC
<none>              1354.8 120.31
+ Rate  1    40.372 1314.4 121.40

Call:
lm(formula = Defective ~ Temperature + Density, data = d[, -1])

Coefficients:
(Intercept)  Temperature      Density
     46.238       18.050       -2.327
```

AIC of intercept model

all 3 SLRs better

intercept model

best SLR now reference

both 2-predictor models better

best SLR

best 2-predictor model

2-predictor model better than full model

selected model

change starting model

```
> stepAIC(lm(Defective ~ ., data=d[,-1]),
+         scope=list(lower=lm(Defective ~ 1, data=d[,-1])),
+         direction = "backward", k=2)

Start:  AIC=121.4
Defective ~ Temperature + Density + Rate

              Df Sum of Sq    RSS    AIC
- Rate         1    40.372 1354.8 120.31
- Density      1    75.318 1389.8 121.07
<none>                     1314.4 121.40
- Temperature  1   189.970 1504.4 123.45

Step:  AIC=120.31
Defective ~ Temperature + Density

              Df Sum of Sq    RSS    AIC
<none>                     1354.8 120.31
- Density      1    142.05 1496.8 121.30
- Temperature  1    257.67 1612.5 123.53

Call:
lm(formula = Defective ~ Temperature + Density, data = d[, -1])

Coefficients:
(Intercept)  Temperature      Density
     46.238       18.050       -2.327
```

specify direction

adjust end point

AIC of full model

only 2 two-predictor models better than full

no SLRs better than two-predictor model

selected model

# EXAMPLE OF STEPWISE SELECTION

specify full model as starting point

AIC of full model

indicate you will work in both directions

```
> stepAIC(lm(Defective ~ ., data=d[,-1]),
+          direction="both", k=2)
Start:  AIC=121.4
Defective ~ Temperature + Density + Rate

              Df Sum of Sq    RSS    AIC
- Rate         1    40.372 1354.8 120.31
- Density      1    75.318 1389.8 121.07
<none>                     1314.4 121.40
- Temperature  1   189.970 1504.4 123.45

Step:  AIC=120.31
Defective ~ Temperature + Density

              Df Sum of Sq    RSS    AIC
<none>                     1354.8 120.31
- Density      1   142.050 1496.8 121.30
+ Rate         1    40.372 1314.4 121.40
- Temperature  1   257.670 1612.5 123.53

Call:
lm(formula = Defective ~ Temperature + Density, data = d[, -1])

Coefficients:
(Intercept)  Temperature      Density
     46.238       18.050       -2.327
```

all deletion

best two-predictor model

AIC of best two-predictor model

considers both deleting (to get SLRs) and adding back what was removed

current model is best

# MODULE 9 OUTLINE

1. Numerical Measures of Goodness
2. All Possible Subsets Selection
3. Automated Selection Methods
4. Cautions for Automated Selection Tools

# PROS & CONS OF AUTOMATED SELECTION

Pros:

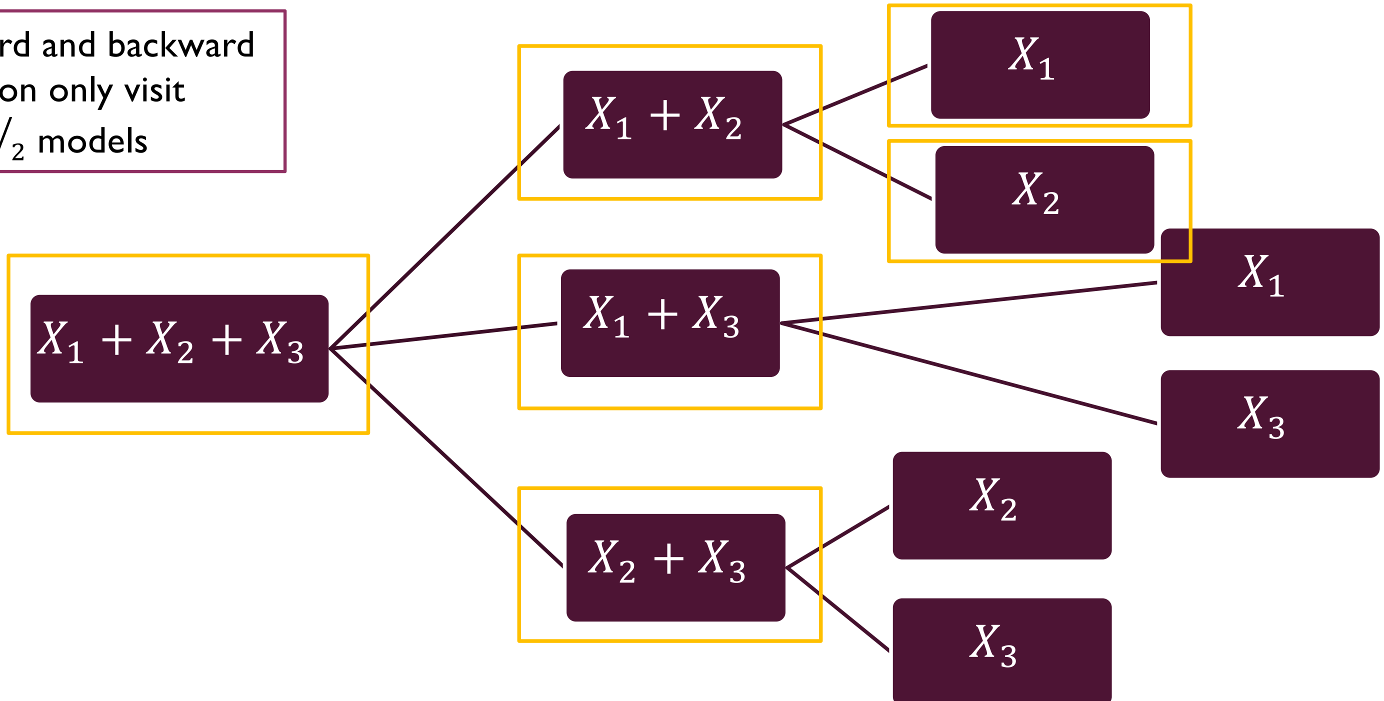
- All less intensive than all possible subsets
  - with systematic way to select model from large number of predictors
- Give an idea of preferred model, although may not actually be the best one
- Stepwise selection accounts for conditional nature of linear regression
  - allows addition and removal of predictors
  - acknowledges importance of variable may fluctuate in presence of others

Cons:

- All methods may not agree on preferred model
  - further, AIC and BIC may also disagree with each other
- Will run even in presence of model violations or other issues (e.g., multicollinearity)
  - can provide unreliable or even incorrect results
- Do not consider context of data or question in decision-making
  - easily select a model that omitted important evidence-based predictors.

# NOT ALL MODELS CONSIDERED

Forward and backward selection only visit $^{p(p+1)}/_{2}$ models



# AUTOMATION IS IGNORANT

- Automated methods can certainly be time-saving when $p$ is large
- Blindly using them can cause you to select a model far from the "best"
  - this is because automated methods (and your computer) is ignorant of potential problems with the model/data
- Ignores context and purpose of the model
  - may remove predictor of interest or predictor known to be relevant from literature
  - extra predictors may aid with predictions
- Ignores model violations and other issues
  - does not check model assumptions at intermediate steps
  - does not perform diagnostics or address any such issues
- Actually creates bias in the model (Loeb & Potscher, 2005)
  - similar to p-hacking where we adjust methods until significant result occurs
  - caused by searching the data for significance rather than collecting data to test specific hypothesis
- Advice for using:
  - look for agreement between methods, perform all checks before and after, consider using intermediate models

# MODULE TAKE-AWAYS

1. How are the likelihood criteria of goodness computed?
2. What are the differences and similarities between the likelihood criteria of goodness?
3. How is a “best”/”preferred” model selected using likelihood criteria?
4. How does all possible subsets and the automated selection procedures select a “best” model?
5. What are the advantages and disadvantages of using automated and all possible subsets model selection?