THE UNIVERSITY OF WESTERN AUSTRALIA

STAT2401 Analysis of Experiments

Lecture Week 9 Dr Darfiana Nur

Aims of this lecture

- Aim 1 Introduction to variable selection
- Aim 2 All subsets selection
- Aim 3 Stepwise, or sequential methods
  - 3.1 Forward
  - 3.2 Backward

Aim 1 Variable Selection

- In many situations we have multiple potential explanatory variables to choose from, not all of which are related to the response. In addition, many of these variables may be relatedto each other and multi-collinearity may be a problem.
- We need a method to select an optimal model from these variables and choose the most significant ones effecting the response.
- We need to select among the variables that are collinear or strongly related to each other so parameter variances will not be inflated.
- There are many such methods around.

Variable subset selection methods

- We’ve seen that even in small examples (e.g., fuel consumption in 50 states and DC example), finding the ‘best’ model is not straightforward
  - When explanatory variables are related, the significance of a variable in a model depends on what terms are already in the model
  - Can use partial F-tests to assess the significance of subsets of coefficients, but doing so manually is tedious
- Using observational data, prediction of the response is often the principal objective
  - Even though causal attribution isn’t possible, we shouldn’t ignore *why* a variable might be in a model

Variable subset selection methods

- With cheap computing, automatic variable selection methods have been developed to choose a subset of predictors that are 'best' in a given sense
- Unfortunately,
  - There are lots of criteria for defining what might be 'best' …
  - The number of models to assess gets large very quickly: if there are m potential predictors, there will be 2^m potential regression equations
    - When $m = 100$, $2^{100} \approx 1.27 \times 10^{30}$
  - Adding additional variables decreases RSS (SSE), but it doesn't mean the the predictive capability of the model will necessarily increase
    - Need some criteria that allow us to assess the trade-off between model complexity and 'goodness-of-fit'
- Subset/variable selection methods help us identify a handful of models that we might want to examine and assess further, e.g., their predictive ability

Classes of subset selection methods

- Brute-force
  - All subsets selection
- Stepwise methods
  - Forward, backward, and ‘both directions’
- Regularization methods (not covered)
  - Shrinkage (no variable selection) and shrinkage and selection

Aim 2 All subsets selection

- All subsets selection can be thought of as a 'brute-force' method in which we evaluate all possible 2^m subsets of m variables; if m is too large, we evaluate all possible 2^q subsets, where $q \ll m$
  - First determine candidate models containing $1, 2, \ldots, p$ predictors based on RSS
  - Then evaluate these subsets based on information criteria to determine *which* subset(s) to consider; information criteria include:

SSE = RSS

- $R^2_{\text{adj}} = 1 - \frac{\mathrm{RSS}/(n-p-1)}{\mathrm{SST}/(n-1)}$ ~~$\frac{\mathrm{SSReg}}{\mathrm{SST}}$~~ RSS: Residual Sum of Squares; SSE: Sum of Squares Error; RSS=SSE
- $\mathrm{AIC} = n \log\left(\frac{\mathrm{RSS}}{n}\right) + 2p$ (Akaike Information Criterion)
- $\mathrm{BIC} = n \log\left(\frac{\mathrm{RSS}}{n}\right) + (p+2)\log(n)$ (Bayesian Information Criterion)
- $C_{p'} = \frac{RSS}{\hat{\sigma}^2} + 2p' - n$ where p'=p+1

  - Each of these criteria can be considered as a compromise between 'goodness-of-fit' (small $RSS(SSE)$) and the number of variables in the model

Example 1 Model with ALL factors

Used car price with all 13 possible explanatory variables

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	10112.212	1366.193		7.402	.000	7271.057	12953.367
	SUNROOF	3134.285	525.186	.380	5.968	.000	2042.102	4226.468
	AGE	-1230.970	141.912	-.725	-8.674	.000	-1526.093	-935.848
	ODOMETER	1.073E-03	.011	.009	.102	.920	-.021	.023
	AUTO	395.309	507.610	.052	.779	.445	-660.324	1450.942
	AIRCON	-612.622	952.770	-.038	-.643	.527	-2594.016	1368.773
	NOCYL	262.403	147.955	.127	1.774	.091	-45.286	570.091
	GTMODEL	2559.814	591.873	.332	4.325	.000	1328.948	3790.681
	RED	-677.744	983.048	-.051	-.689	.498	-2722.104	1366.616
	BLUE	-443.159	899.939	-.052	-.492	.628	-2314.686	1428.367
	BLACK	-518.086	866.782	-.049	-.598	.556	-2320.657	1284.485
	WHITE	-346.609	859.088	-.041	-.403	.691	-2133.181	1439.963
	SILVER	707.890	1381.272	.032	.512	.614	-2164.622	3580.402
	BURGUNDY	159.589	908.919	.015	.176	.862	-1730.612	2049.790

a. Dependent Variable: PRICE

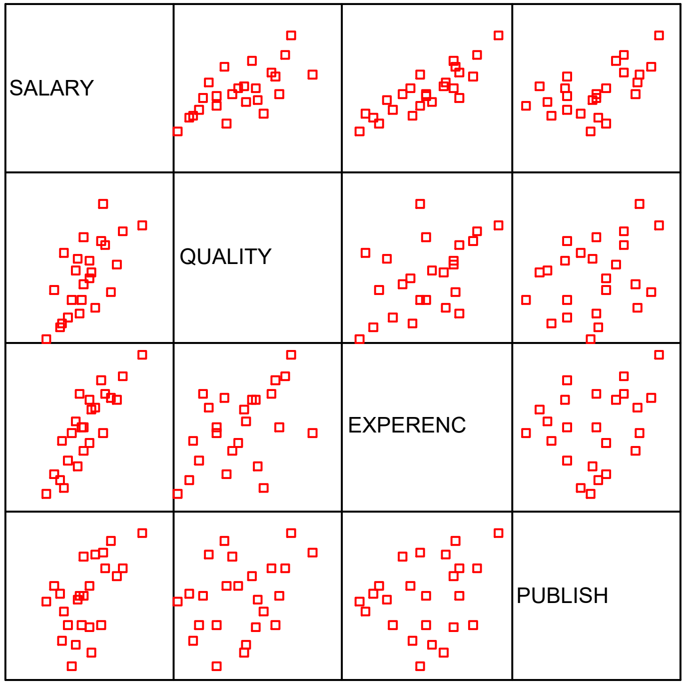
SPSS OUTPUT

Some of these variables seem to have insignificant effect on price

Surely a subset of these variables will model price almost as well

Example 2 All possible subsets regression

Consider the mathematics lecturers data. There are 3 explanatory variables meaning $2^3 = 8$ possible models



$$X_1 = Quality; X_2 = Experience;$$

$$X_3 = publications$$

The possible models are

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$$
$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$
$$Y = \beta_0 + \beta_2 X_2 + \varepsilon$$
$$Y = \beta_0 + \beta_3 X_3 + \varepsilon$$
$$Y = \beta_0 + \varepsilon$$

This only includes the linear models!

We can fit each of these models. How do we choose the best model? We need a criterion.

1. R_A^2 or adjusted R^2

$$R_{\text{adj}}^2 = 1 - \frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}$$

The adjusted figure takes account of the number of variables in the model.

R^2 always increases when we add another variable to the model. Why?

2. MSE = $\hat{\sigma}^2 = \frac{SSE}{n-p-1}$

MSE also takes account of the number of variables in the model

3. Mallows $C_{p'}$

p’=p+1,
p be the number of predictors

$$C_{p'} = \frac{SSE}{MSE_{full}} + 2p' - n$$

where MSE is calculated for the model with ALL possible explanatory variables included. It’s possible to show that for good models

$$E(C_{p'}) \approx p'$$

If there are *K-1* variables (*K* parameters to be estimated) in the full model, then $C_K = K$

R^2-adjusted

- We have seen that $R^2 = \frac{\text{SSReg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}}$
- Adding irrelevant predictor variables to regression equation often increases R^2
- To compensate for the number of variables, define an adjusted coefficient of determination, R^2_{adj}, as

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-p-1)}{\text{SST}/(n-1)}$$

- R^2_{adj} not immune to including irrelevant variables, so often used in conjunction with other criteria

Information criteria

- AIC (Akaike Information Criterion)

$$\text{AIC} = n \log\left(\frac{\text{RSS}}{n}\right) + 2p$$

- BIC (Bayesian Information Criterion)

$$\text{BIC} = n \log\left(\frac{\text{RSS}}{n}\right) + (p + 2) \log(n)$$

- Mallows' $C_{p'}$ statistic (p'=p+1)

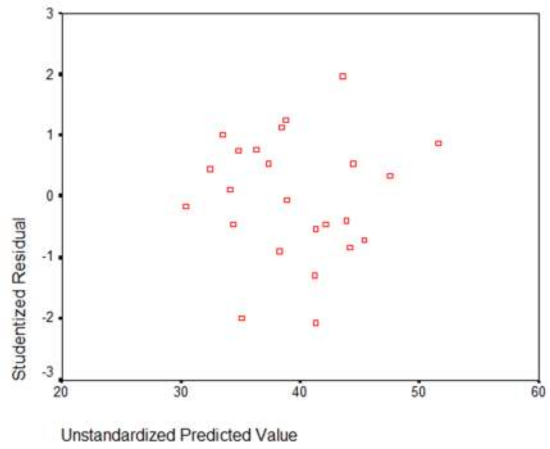
$$C_{p'} = \frac{RSS}{\hat{\sigma}^2} + 2p' - n$$

Example 2 Salary data: All possible subsets regression

Model	C(p')	MSE	Adj. R2
1 (p'=4)	4	3.072	0.897
2 (p'=3)	20.65	5.654	0.811
3 (p'=3)	77.07	13.908	0.536
4 (p'=3)	13.21	4.565	0.848
5 (p'=2)	104.52	17.388	0.42
6 (p'=2)	38.98	8.236	0.725
7 (p'=2)	134.47	21.57	0.28
8 (p'=1)	202.4	29.968	0

From this analysis, it seems only models 1 and 4 are worth looking at.

The full model seems to be the 'best'



The residuals from this model seem ok

Example 3: Highway accident data

Variable	Description
log(*Rate*)	Base-two logarithm of 1973 accident rate per million vehicle miles, the response
log(*Len*)	Base-two logarithm of the length of the segment in miles
log(*ADT*)	Base-two logarithm of average daily traffic count in thousands
log(*Trks*)	Base-two logarithm of truck volume as a percent of the total volume
Slim	1973 speed limit
Lwid	Lane width in feet
Shld	Shoulder width in feet of outer shoulder on the roadway
Itg	Number of freeway-type interchanges per mile in the segment
log(*Sigs1*)	Base-two logarithm of (number of signalized interchanges per mile in the segment + 1)/(length of segment)
Acpt	Number of access points per mile in the segment
Hwy	A factor coded 0 if a federal interstate highway, 1 if a principal arterial highway, 2 if a major arterial, and 3 otherwise

variable Lane

10 numerical, 1 categorical - 3 indicator variable

Example 3: Highway accident data

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.047344	2.623516	2.305053	0.029746
logLen	-0.214470	0.099986	-2.145000	0.041859
logADT	-0.154625	0.111893	-1.381900	0.179227
logTrks	-0.197560	0.239812	-0.823816	0.417835
logSigs1	0.192322	0.075367	2.551806	0.017211
slim	-0.039327	0.024236	-1.622645	0.117210
shld	0.004291	0.049281	0.087076	0.931305
lane	-0.016061	0.082264	-0.195235	0.846787
acpt	0.008727	0.011687	0.746730	0.462192
itg	0.051536	0.350312	0.147115	0.884221
lwid	0.060769	0.197391	0.307860	0.760739
hwyMA	-0.550063	0.515724	-1.066585	0.296352
hwyMC	-0.342705	0.576821	-0.594127	0.557766
hwyPA	-0.755001	0.418441	-1.804316	0.083244

R OUTPUT

Some of these variables seem to have insignificant effect on price

```
print(load(".RData"))
lm1 <- lm(logRate ~ ., data =
Highway1)
summary(lm1)$coefficients
```

Example 3: Highway accident data

	logLen	logADT	logTrks	logSigs1	slim	shld	lane	acpt	itg	lwid	hwyMA	hwyMC	hwyPA
1					*								
2	*				*								
3				*	*								*
4	*			*	*								*
5	*	*		*	*								*
6	*	*		*	*						*		*
7	*	*	*	*	*						*		*
8	*	*	*	*	*			*			*		*
9	*	*	*	*	*			*			*	*	*
10	*	*	*	*	*			*		*	*	*	*

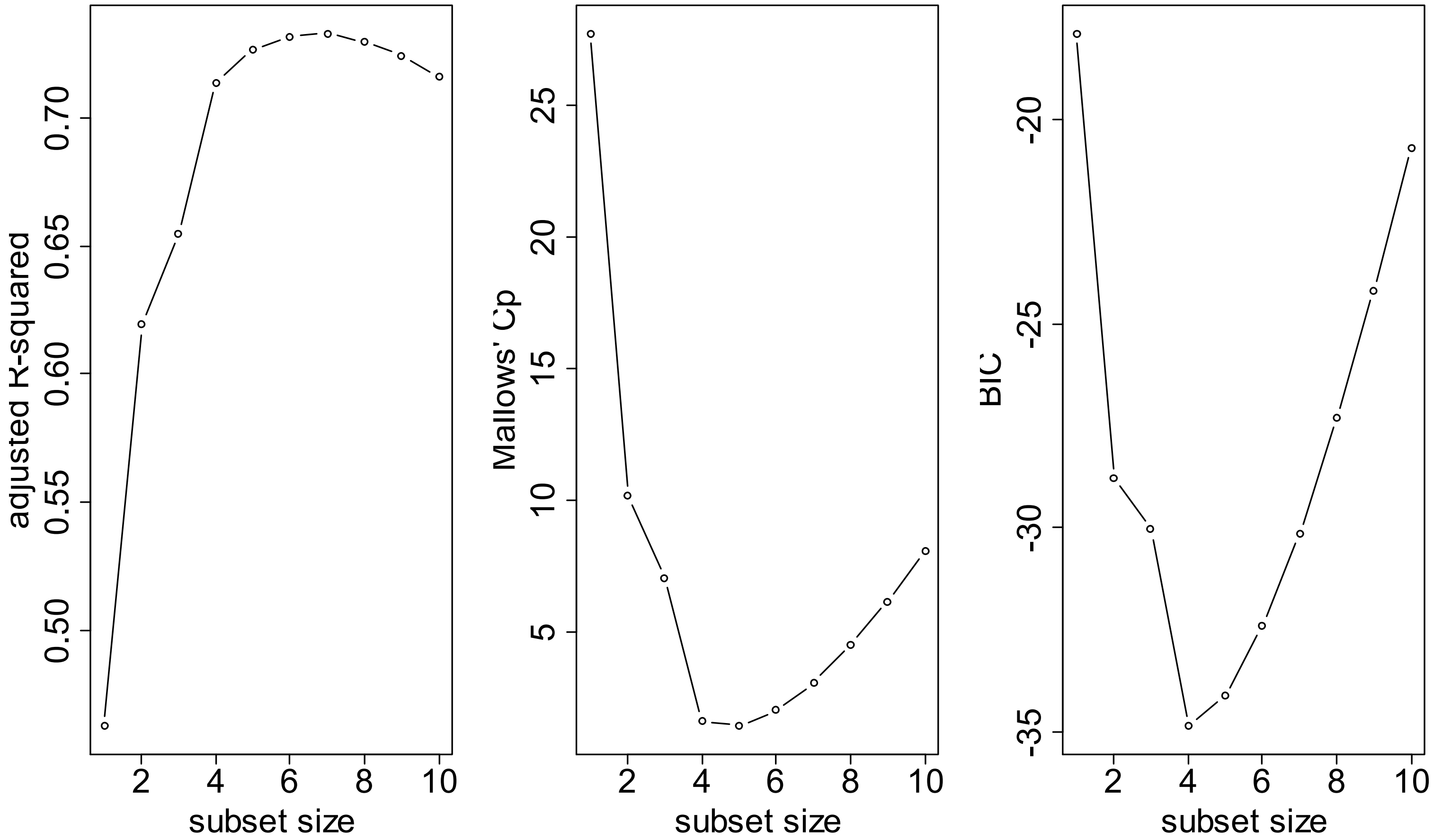
all subset models 2^13 = 8192,
dont wanna do so use library leaps

```
require(leaps)
AllSubsets <-
regsubsets(logRate
~ ., nvmax = 10,
data = Highway1)
AllSubsets.summar
y <-
summary(AllSubset
s)
```

nbest - to choose best models 1or more

nvmax - max size of subset we want to work on

Example 3: Highway accident data




```
par(mfrow = c(1, 3))
par(cex.axis = 1.5)
par(cex.lab = 1.5)
plot(1:10,
AllSubsets.summary$adjr2, xlab =
"subset size", ylab = "adjusted R-
squared", type = "b")
plot(1:10,
AllSubsets.summary$cp, xlab =
"subset size", ylab = "Mallows'
Cp", type = "b")
plot(1:10,
AllSubsets.summary$bic, xlab =
"subset size", ylab = "BIC", type =
"b")
par(mfrow = c(1, 1))
par(cex.axis = 1)
par(cex.lab = 1.5)
```

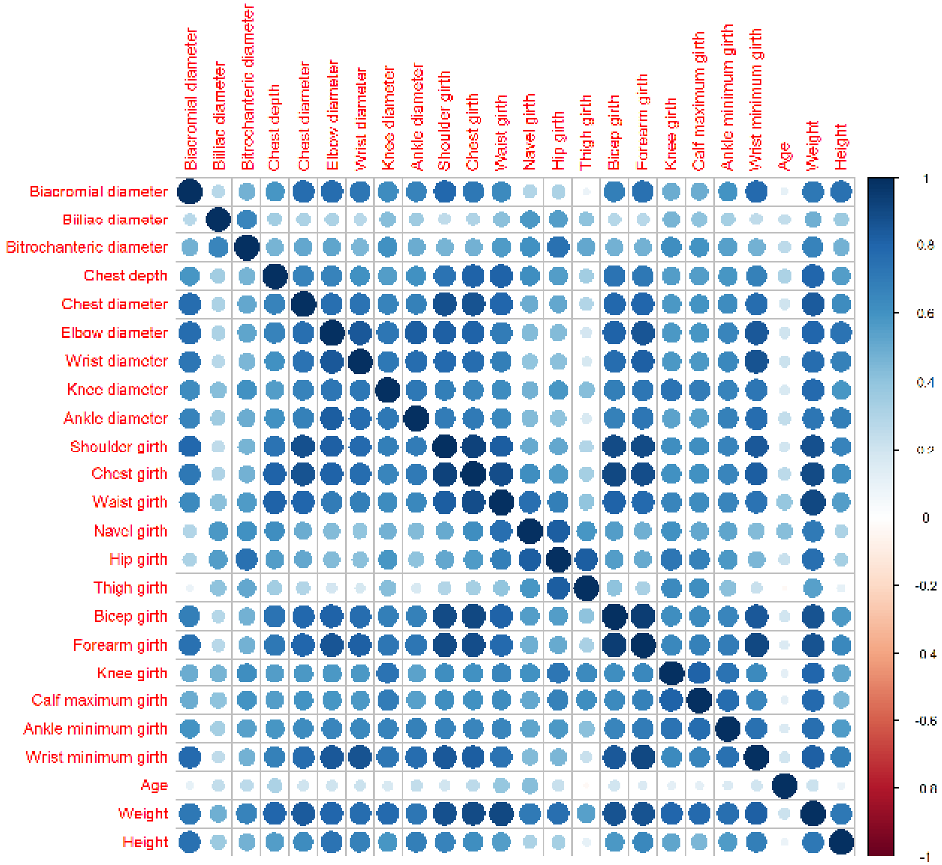
Example 3: Highway accident data

- (Mallows' C_p equivalent to BIC in this case.)
- As the number of variables in the model increases:
  - R^2_{adj} increases and then decreases
  - Mallows' C_p and BIC decrease and then start to increase
  - BIC increases more quickly than C_p (or AIC) because it penalizes more severely
- Criteria don't necessarily agree on which models might be the 'best'
- **Next steps**: might push ahead further in the model evaluation with models consisting of between 4 and 6 variables

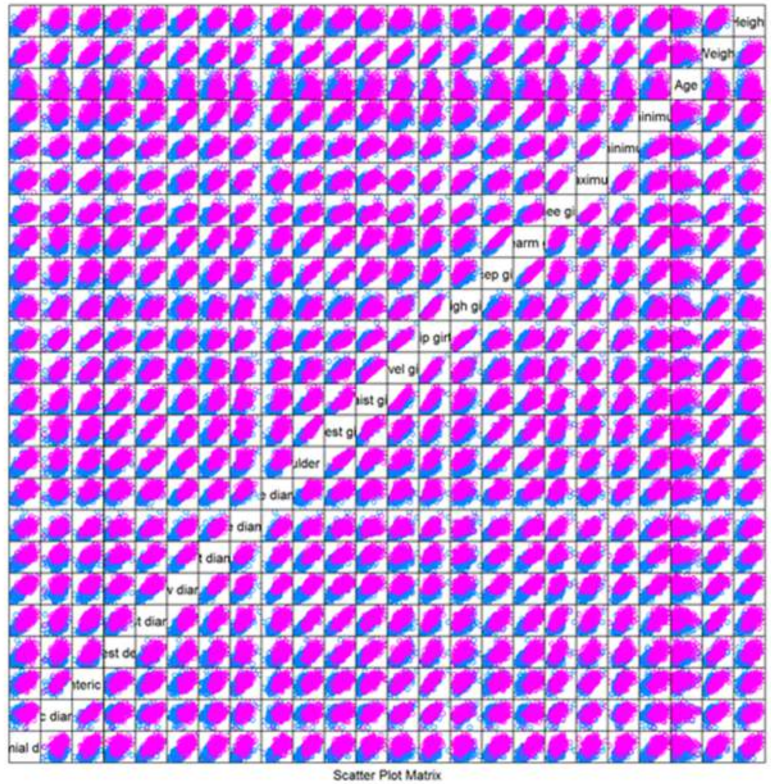
Example 4: Body weight

- Objective is to predict body weight from using 24 potential covariates:
  - Chest depth
  - Chest diameter
  - Knee diameter
  - Shoulder girth
  - …
  - Age
  - Height
  - Gender
- Covariates are highly correlated

Example 4: Body weight - Exploratory



require(corrplot)
corrplot(cor(BodyMeasurements[, -25])) # remove gender from display



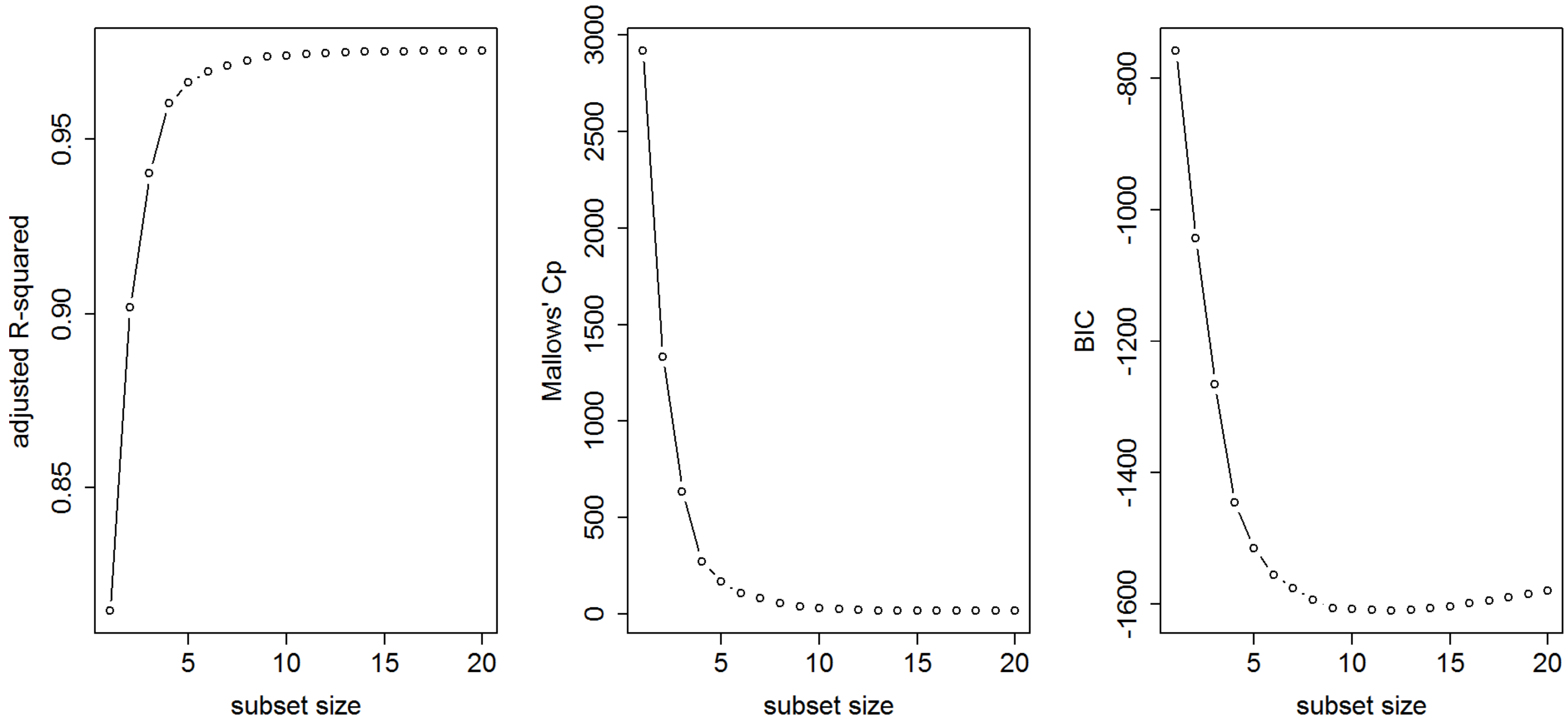
require(lattice)
splom(~BodyMeasurements[, -25], groups = Gender, data = BodyMeasurements, pscales = 0, varname.cex = 0.5)

Example 4: Body weight

```
require(leaps)
AllSubsets <- regsubsets(Weight ~ ., nvmax = 20,
data = BodyMeasurements)
AllSubsets.summary <- summary(AllSubsets)
AllSubsets.outmat <-
AllSubsets.summary$outmat
```

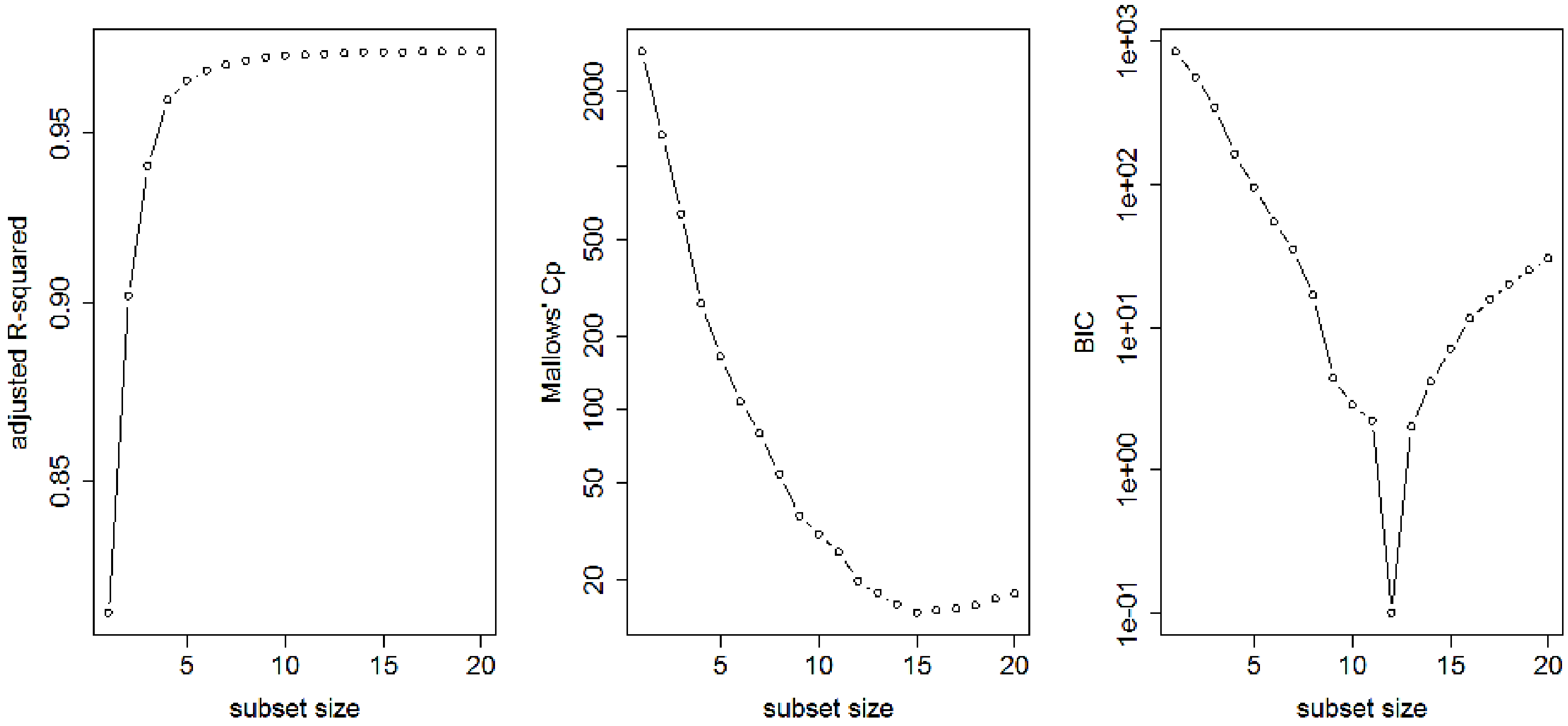
	BiaDia	BiiDia	BitDia	CheDep	CheDia	ElbDia	WriDia	KneDia	AnkDia	ShoGir	CheGir	WaiGir	NavGir	HipGir	ThiGir	BicGir	ForGir	KneGir	CalGir	AnkGir	WriGir	Age	Height	Sex
1												*												
2											*							*						
3												*			*								*	
4												*			*		*						*	
5											*	*			*				*				*	
6											*	*			*		*		*				*	
7											*	*		*	*		*		*				*	
8								*			*	*		*	*		*		*				*	
9								*			*	*		*	*		*		*			*	*	
10				*				*			*	*		*	*		*		*			*	*	

Example 4: Body weight




```
par(mfrow = c(1, 3))
par(cex.axis = 1.5)
par(cex.lab = 1.5)
plot(1:20,
AllSubsets.summary$adjr2,
xlab = "subset size", ylab =
"adjusted R-squared", type
= "b")
plot(1:20,
AllSubsets.summary$cp,
xlab = "subset size", ylab =
"Mallows' Cp", type = "b")
plot(1:20,
AllSubsets.summary$bic,
xlab = "subset size", ylab =
"BIC", type = "b")
par(mfrow = c(1, 1))
par(cex.axis = 1)
par(cex.lab = 1.5)
```

Example 4: Log(body weight)




```
par(mfrow = c(1, 3))
par(cex.axis = 1.5)
par(cex.lab = 1.5)
plot(1:20, AllSubsets.summary$adjr2, xlab = "subset size", ylab = "adjusted R-squared", type = "b", log = "y")
plot(1:20, AllSubsets.summary$cp, xlab = "subset size", ylab = "Mallows' Cp", type = "b", log = "y")
plot(1:20, AllSubsets.summary$bic - min(AllSubsets.summary$bic) + 0.1, xlab = "subset size", ylab = "BIC", type = "b", log = "y")
par(mfrow = c(1, 1))
par(cex.axis = 1)
par(cex.lab = 1.5)
```

Example 5: Body weight (12 variables)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -122.26113    2.52608 -48.400  < 2e-16 ***
CheDep         0.26584    0.06879   3.864 0.000126 ***
KneDia         0.64053    0.11727   5.462 7.47e-08 ***
ShoGir         0.08655    0.02825   3.063 0.002308 ** 
CheGir         0.16188    0.03385   4.782 2.30e-06 ***
WaiGir         0.38580    0.02499  15.440  < 2e-16 ***
HipGir         0.23328    0.03843   6.070 2.55e-09 ***
ThiGir         0.25782    0.04873   5.290 1.84e-07 ***
ForGir         0.59434    0.09648   6.160 1.51e-09 ***
CalGir         0.40568    0.05797   6.998 8.49e-12 ***
Age           -0.05331    0.01181  -4.515 7.93e-06 ***
Height         0.32247    0.01553  20.769  < 2e-16 ***
Gender        -1.57950    0.48321  -3.269 0.001155 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.114 on 494 degrees of freedom
Multiple R-squared:  0.9755,	Adjusted R-squared:  0.9749 
F-statistic:  1639 on 12 and 494 DF,  p-value: < 2.2e-16
```

```
# Model with 12 variables
# Don't worry about how this next line is constructed
lm.as <- lm(formula(paste("Weight ~",
paste(names(which(AllSubsets.outmat[12, ] == "*")), collapse = " + "))),
data = BodyMeasurements)
summary(lm.as)
```

Strategies for dealing with many explanatory variables

1. Identify the main objectives of the analysis
2. Justify the potential inclusion of each variable in the model
3. Exploratory and graphical analysis using scatterplots and correlations. correlation will only among numerical vars, not categorical
   - Remove one of each pair of highly collinear variables.
   - Consider possible transformations of explanatory variables and/or response variable (*Y*).
4. Find a suitable subset of explanatory variables.

All subsets selection: Summary

- Brute-force method
- Can also consider top-2 models (smallest RSS or SSE) containing $1, 2, \ldots$ variables
- Gives us a smaller candidate set of models that we can take forward for further investigation
- Can be unrealistic to computer when we have lots of potential explanatory variables

Aim 3 Stepwise Regression

- Stepwise methods carry out a **sequential** search of the 2^m possible regression models that involves evaluating many fewer models
- Stepwise methods **not guaranteed** to find the candidate subset that is **optimal** according to any overall criterion, but produce results in practice
- Models are often evaluated during the search procedure using F statistic, AIC or BIC
- Forward, backward, or 'both-directions'

Aim 3.1 Stepwise regression – Forward using F Statistic

- We can use methods learnt so far in the unit to develop a technique to select a significant subset of explanatory variables

FORWARD SELECTION

1. Start with the constant mean model $Y = \beta + \varepsilon$.

2. Consider all possible models with 1 explanatory variable. For each of these models, calculate the F statistic of the hypothesis test comparing

$$H_0 : Y = \beta_0 + \varepsilon \qquad (\beta_1 = 0)$$
$$H_A : Y = \beta_0 + \beta_1 X + \varepsilon \qquad (\beta_1 \neq 0)$$

Do this test for each variable separately

3. Add the variable to the model with the largest F statistic **IF this F stat > 4 or greater than $F_{0.95,1,n-2}$**

4. Start with the model $Y = \beta_0 + \beta_1 X + \varepsilon$ including the variable just added. Consider each 2 variable model with each of the remaining explanatory variables. Calculate each F statistic for each added regression parameter.

5. Add the variable to the model with the largest F statistic IF this F stat > 4 or $F_{0.95,1,n-3}$

6. Continue adding variables to the model in this fashion until no more variables are significant (have F stat > 4 or $F_{0.95,1,n-3}$).

7. **Analyse the selected model, find parameter estimates, diagnose the model using residuals, make required inferences to answer objectives.**

Forward selection: using AIC

STEP 1

- Start with a base model, e.g., with intercept only
- Fit all possible models $y = \beta_0 + \beta_j x_j, + \epsilon, j = 1, 2, \ldots, p$, and keep the variable (say it's x_2) that yields the smallest AIC

STEP 2

- Fit $y = \beta_0 + \beta_2 x_2 + \beta_j x_j + \epsilon, j = 1, 3, \ldots, p$, and keep the model with the smallest AIC as long as it's less than AIC in Step 1

$\vdots$

STEP n

- Continue until the addition of an extra term increases the value of AIC

Example 6 Highway: forward selection

STEP 1:

- Fit model with intercept only

STEP 2:

- Fit all models with one explanatory variable, and select the one which minimizes the information criterion
- Keep this model and continue

```
lm.0 <- lm(logRate ~ 1, data = Highway1)
lm.forward <-   step(lm.0, scope = ~ logLen + logADT + logTrks + logSigs1 + slim +
shld + lane + acpt + itg + lwid + hwy, direction = "forward")
```

```
Start:  AIC=-30.5
logRate ~ 1
```

Y = X(beta)

'1' represents the first column of matrix X

	Df	Sum of Sq	RSS	AIC
+ slim	1	8.077	8.874	-53.74
+ acpt	1	7.434	9.517	-51.01
+ logSigs1	1	6.174	10.777	-46.16
+ logLen	1	5.537	11.414	-43.92
+ logTrks	1	5.042	11.909	-42.26
+ shld	1	2.754	14.197	-35.41
<none>			16.951	-30.50
+ hwy	3	1.816	15.135	-28.92
+ lane	1	0.014	16.937	-28.53
+ logADT	1	0.013	16.938	-28.53
+ itg	1	0.012	16.939	-28.52
+ lwid	1	0.008	16.943	-28.52

Example 6 Highway: forward selection

STEP 3:

- Fit all possible models with intercept, `slim`, and an additional variable and select the one with the smallest AIC
- If it is less than the AIC of previous model, continue; if not, stop

```
Step:  AIC=-53.74
logRate ~ slim

            Df Sum of Sq   RSS    AIC
+ logLen     1    2.7618 6.112 -66.28
+ logTrks    1    2.0098 6.864 -61.75
+ logSigs1   1    1.7430 7.131 -60.27
+ acpt       1    1.1646 7.709 -57.22
<none>                   8.874 -53.74
+ lane       1    0.4327 8.441 -53.69
+ logADT     1    0.3579 8.516 -53.34
+ itg        1    0.3543 8.520 -53.33
+ shld       1    0.1699 8.704 -52.49
+ lwid       1    0.1392 8.735 -52.35
+ hwy        3    0.3626 8.511 -49.36
```

Example 6 Highway: forward selection, final step

```
Step:  AIC=-68.31
logRate ~ slim + logLen + acpt

           Df Sum of Sq   RSS    AIC
+ logTrks   1    0.3600 5.152 -68.94
<none>                  5.512 -68.31
+ logSigs1  1    0.2499 5.262 -68.12
+ shld      1    0.0720 5.440 -66.82
+ logADT    1    0.0316 5.480 -66.53
+ lane      1    0.0310 5.481 -66.53
+ itg       1    0.0281 5.484 -66.51
+ lwid      1    0.0263 5.485 -66.50
+ hwy       3    0.4527 5.059 -65.65
```

```
Step:  AIC=-68.94
logRate ~ slim + logLen + acpt +
logTrks

           Df Sum of Sq   RSS    AIC
<none>                  5.152 -68.94
+ shld      1    0.1359 5.016 -67.99
+ logSigs1  1    0.1053 5.047 -67.75
+ logADT    1    0.0650 5.087 -67.44
+ hwy       3    0.5401 4.612 -67.26
+ lwid      1    0.0396 5.112 -67.24
+ itg       1    0.0228 5.129 -67.12
+ lane      1    0.0069 5.145 -67.00
```

Example 6 Highway
Forward selection- ‘final’ model

```
Call:
lm(formula = logRate ~ slim + logLen + acpt + logTrks, data = Highway1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.011048   1.069130   5.622 2.67e-06 ***
slim        -0.045953   0.014805  -3.104  0.00383 ** 
logLen      -0.235735   0.084897  -2.777  0.00887 ** 
acpt         0.015876   0.009622   1.650  0.10815    
logTrks     -0.329037   0.213484  -1.541  0.13251    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3893 on 34 degrees of freedom
Multiple R-squared: 0.6961, Adjusted R-squared: 0.6603 
F-statistic: 19.47 on 4 and 34 DF,  p-value: 2.067e-08
```

In R

```
lm.0 <- lm(logRate ~ 1, data = Highway1)

lm.forward <-   step(lm.0, scope = ~ logLen + logADT + logTrks + logSigs1 + slim +  shld + lane + acpt + itg + lwid + hwy, direction = "forward")
```

Aim 3.2 BACKWARD SELECTION – using F statistic

1. Start with the full model containing all *p* explanatory variables $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$

2. Consider all possible (p-1) variable models, calculating the F statistic for each variable removed from the model.

3. **Remove** the variable with the smallest F statistic, IF this F stat < 2 (or $F_{0.9,1,n-K-1}$)

4. Continue this until no variables can be further removed from the model.

A higher level of 10% is used here to allow for some multi-collinearity in the full model

Backward selection – using AIC

STEP 1

- Start with the full model $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$

STEP 2

- Consider all possible subsets obtained by removing one variable, and keep the subset that yields the largest AIC

$\vdots$

STEP n

- Continue until the next deletion increases the value of the criterion, or until all terms have been deleted

Example 7: Backward selection

STEP 1:

- Fit full model

STEP 2:

- Delete one variable at a time, and then select the model that minimizes the information criterion
- Keep this model and continue

```
Start:  AIC=-65.61
logRate ~ logLen + logADT + logTrks + logSigs1 +
slim + shld + lane + acpt + itg + lwid + hwy

           Df Sum of Sq   RSS    AIC
- shld      1    0.0011 3.538 -67.60
- itg       1    0.0031 3.540 -67.58
- lane      1    0.0054 3.542 -67.55
- lwid      1    0.0134 3.550 -67.46
- acpt      1    0.0789 3.616 -66.75
- logTrks   1    0.0960 3.633 -66.57
<none>                  3.537 -65.61
- hwy       3    0.6253 4.162 -65.26
- logADT    1    0.2702 3.807 -64.74
- slim      1    0.3725 3.909 -63.71
- logLen    1    0.6509 4.188 -61.02
- logSigs1  1    0.9213 4.458 -58.58
```

Example 7 Highway: Backward selection

STEP 1:

- Fit full model

STEP 2:

- Delete one variable at a time, and then select the model that minimizes the information criterion
- Keep this model and continue

```
Start:  AIC=-65.61
logRate ~ logLen + logADT + logTrks + logSigs1 +
slim + shld + lane + acpt + itg + lwid + hwy

           Df Sum of Sq   RSS    AIC
- shld      1    0.0011 3.538 -67.60
- itg       1    0.0031 3.540 -67.58
- lane      1    0.0054 3.542 -67.55
- lwid      1    0.0134 3.550 -67.46
- acpt      1    0.0789 3.616 -66.75
- logTrks   1    0.0960 3.633 -66.57
<none>                  3.537 -65.61
- hwy       3    0.6253 4.162 -65.26
- logADT    1    0.2702 3.807 -64.74
- slim      1    0.3725 3.909 -63.71
- logLen    1    0.6509 4.188 -61.02
- logSigs1  1    0.9213 4.458 -58.58
```

Example 7: Backward selection, final step

```
Step:  AIC=-74.21
logRate ~ logLen + logADT + logTrks
+ logSigs1 + slim + hwy

           Df Sum of Sq   RSS    AIC
- logTrks   1    0.1429 3.810 -74.71
<none>                  3.667 -74.21
- logADT    1    0.3106 3.977 -73.03
- logLen    1    0.9437 4.611 -67.27
- hwy       3    1.5129 5.180 -66.73
- logSigs1  1    1.1598 4.827 -65.49
- slim      1    1.2071 4.874 -65.11
```

```
Step:  AIC=-74.71
logRate ~ logLen + logADT + logSigs1
+ slim + hwy

           Df Sum of Sq   RSS    AIC
<none>                  3.810 -74.71
- logADT    1    0.2882 4.098 -73.87
- hwy       3    1.6857 5.495 -66.43
- slim      1    1.1595 4.969 -66.35
- logLen    1    1.2489 5.059 -65.66
- logSigs1  1    1.5637 5.373 -63.30
```

Example 7 Backward selection -'final' model

```
Call:
lm(formula = logRate ~ logLen + logADT + logSigs1 + slim + hwy, data = Highway1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  6.45541    0.98737   6.538 2.68e-07 ***
logLen      -0.26161    0.08206  -3.188  0.00327 ** 
logADT      -0.12691    0.08287  -1.531  0.13581    
logSigs1     0.20836    0.05841   3.567  0.00120 ** 
slim        -0.04290    0.01397  -3.072  0.00441 ** 
hwyMA       -0.38446    0.36526  -1.053  0.30067    
hwyMC       -0.17862    0.48529  -0.368  0.71533    
hwyPA       -0.71475    0.28662  -2.494  0.01819 *  
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3506 on 31 degrees of freedom
Multiple R-squared: 0.7753, Adjusted R-squared: 0.7245
F-statistic: 15.28 on 7 and 31 DF, p-value: 1.835e-08
```

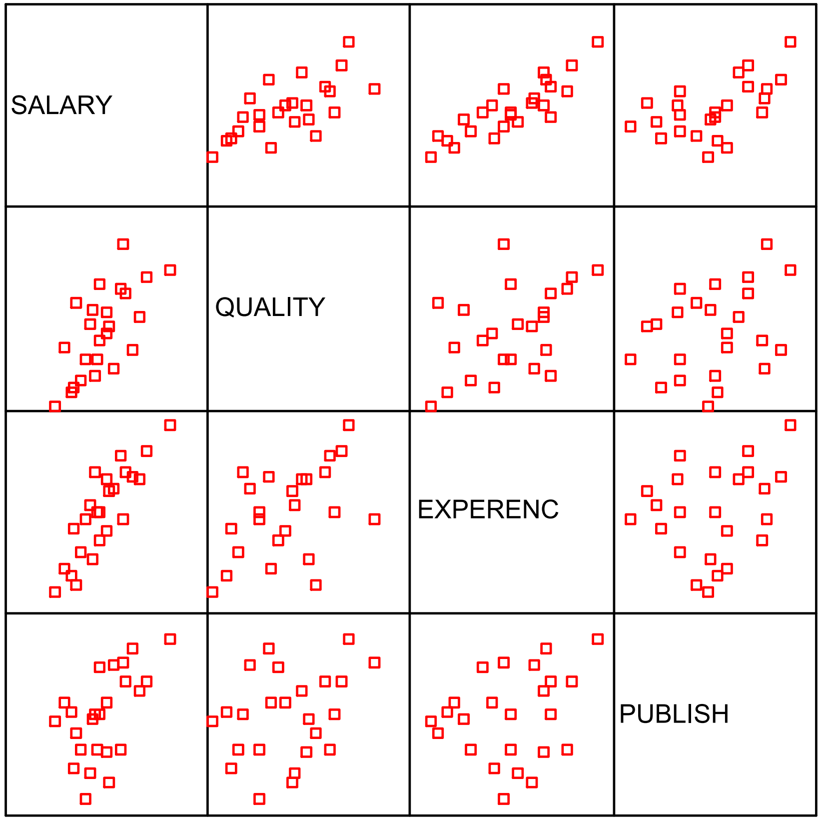
In R

```
lm.all <-
lm(logRate ~ .,
data =
Highway1)

lm.backward <-
step(lm.all,
direction =
"backward" )
```

Example 8: Mathematicians salaries

Objective: Identify factors affecting salary level and build model predicting salary level



The explanatory variables are not strongly related to each other and all 3 make sense to include in the model

Forward selection: SPSS Output – F statistic

Variables Entered/Removed [a]

Model	Variables Entered	Variables Removed	Method
1	EXPERENC	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
2	PUBLISH	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
3	QUALITY	.	Forward (Criterion: Probability-of-F-to-enter <= .050)

a. Dependent Variable: SALARY

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.859[a]	.737	.725	2.8698
2	.928[b]	.861	.848	2.1365
3	.954[c]	.911	.897	1.7528

a. Predictors: (Constant), EXPERENC

b. Predictors: (Constant), EXPERENC, PUBLISH

c. Predictors: (Constant), EXPERENC, PUBLISH, QUALITY

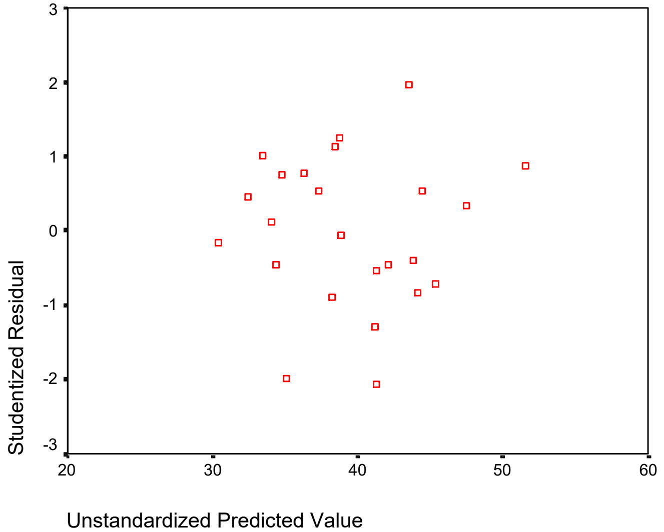
Experience is added first, then *Publish* and then *Quality*. All three significantly affect salary.

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	Upper Bound
1	(Constant)	29.048	1.454		19.978	.000	26.032	32.063
	EXPERENC	.419	.053	.859	7.854	.000	.308	.529
2	(Constant)	21.025	2.148		9.788	.000	16.558	25.493
	EXPERENC	.374	.041	.766	9.107	.000	.288	.459
	PUBLISH	1.528	.353	.364	4.324	.000	.793	2.262
3	(Constant)	17.847	2.002		8.915	.000	13.671	22.023
	EXPERENC	.322	.037	.659	8.664	.000	.244	.399
	PUBLISH	1.289	.298	.307	4.318	.000	.666	1.912
	QUALITY	1.103	.330	.260	3.347	.003	.416	1.791

a. Dependent Variable: SALARY

Once we have selected a model, we should diagnostically check it out using the studentised or deleted residuals.



Summary of models chosen plus final model.

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	PUBLISH, EXPERENC, QUALITY[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: SALARY

Backward selection gives us the same model in this case.

The linear model seems fine.

Backward selection using F-Test in R

- We illustrate the implementation using Cheese Tasting Data
  - Data on production of cheddar cheese from the LaTrobe Valley of Victoria
  - Taste of the final product is related to the concentration of several chemicals in the cheese.
  - 30 samples of cheese were tasted by experts, and the following variables: Tasters' ratings (taste), Acetic acid in cheese (Acetic), Hydrogen sulphide in cheese (H2S), and Lactic acid in the cheese (Lactic) are recorded.

```
> cheese = read.table(file="cheese.txt",header=T)
> str(cheese)
'data.frame': 30 obs. of  4 variables:
 $ taste : num  12.3 20.9 39 47.9 5.6 25.9 37.3 21.9 18.1 21 ...
 $ Acetic: num  4.54 5.16 5.37 5.76 4.66 ...
 $ H2S   : num  3.13 5.04 5.44 7.5 3.81 ...
 $ Lactic: num  0.86 1.53 1.57 1.81 0.99 1.09 1.29 1.78 1.29 1.58 ...
```

Backward selection using F-Test in R

- Backward model selection starts with the full model (i.e. with all predictors):

```
> cheese.lm.full=lm(taste~Acetic+H2S+Lactic,data=cheese)
> summary(cheese.lm.full)

Call:
lm(formula = taste ~ Acetic + H2S + Lactic, data = cheese)

Residuals:
    Min      1Q  Median      3Q     Max
-17.390  -6.612  -1.009   4.908  25.449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
Acetic        0.3277     4.4598   0.073  0.94198
H2S           3.9118     1.2484   3.133  0.00425 **
Lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared:  0.6518,Adjusted R-squared:  0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

Backward selection using F-Test in R

- Consider effect of dropping each single variable using drop1:

```
> drop1(cheese.lm.full,test="F")
Single term deletions

Model:
taste ~ Acetic + H2S + Lactic
       Df Sum of Sq    RSS    AIC F value   Pr(>F)
<none>               2668.4 142.64
Acetic  1      0.55 2669.0 140.65  0.0054 0.941980
H2S     1   1007.66 3676.1 150.25  9.8182 0.004247 **
Lactic  1    533.32 3201.7 146.11  5.1964 0.031079 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- R output tells us that both H2S and Lactic should not be dropped, since the models without these terms have a considerably worse fit than the full model (as evidenced by the *p*-values of 0.004 and 0.031 respectively).
- However, deletion of Acetic from the model makes little difference in terms of model fit (*p*-value of 0.942 in comparison with full model), so we should omit this variable.
- If there had been more than one variable with *p*-value greater than 0.05, then we would have removed the variable with largest corresponding *p*-value.

Backward selection using F-Test in R

- We can create a new model without `Acetic` using `update`:

```
> cheese.lm.A = update(cheese.lm.full,.~.-Acetic,data=cheese)
> summary(cheese.lm.A)

Call:
lm(formula = taste ~ H2S + Lactic, data = cheese)

Residuals:
    Min      1Q  Median      3Q     Max
-17.343  -6.530  -1.164   4.844  25.618

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -27.592      8.982  -3.072  0.00481 **
H2S            3.946      1.136   3.475  0.00174 **
Lactic        19.887      7.959   2.499  0.01885 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.942 on 27 degrees of freedom
Multiple R-squared:  0.6517,Adjusted R-squared:  0.6259
F-statistic: 25.26 on 2 and 27 DF,  p-value: 6.551e-07
```

Note

- The general syntax for updating models is

  `update(old.model,new.formula)`

- Note that full stops in the updated formula stand for "whatever was in the comparison position in the old formula".

Dangers of Stepwise regression

- The final model selected does NOT optimise any criterion function. For instance, it doesn’t minimise SSE, MSE OR maximise R_A^2

- Forward and backward selection may give different models

- Multi-collinearity can cause wrong choices to be made.

- In large databases, typically too many explanatory variables are chosen in the final model.

Notes on sequential selection

- Forward and backward don’t always end up at the same ‘final’ model!

```
Forward:   logRate ~ slim + logLen + acpt + logTrks, AIC=-68.94
Backward:  logRate ~ logLen + logADT + logSigs1 + slim + hwy, AIC=-74.71
```

- Compromise between forward and backward selection is known (confusingly) as **stepwise**
  - Additional dropping/adding of terms at each stage to ensure the continued effectiveness of variables that have been added at an earlier stage
  - Stepwise is the default in the *R* function `step`, e.g.,

```
step(lm.all, direction = "both")
```

  - Gives the same result as

```
step(lm.all)
```

Summary for Aims 1-3

- All subsets and sequential methods search the model space to find parsimonious models that optimize some criterion
- Depending on the search method and the criterion, slightly different models can result
- Unlikely there will be one ‘best’ model, but slightly different models that will yield similar performance – after all, we generally use these methods on observational data
- We want to carry forward a small number of candidate models to the next step: evaluating predictive ability