THE UNIVERSITY OF WESTERN AUSTRALIA

STAT2401 Analysis of Experiments

Lecture Week 10 **Dr Darfiana Nur**

Aims of this lecture

1. Predictive ability of linear models (Sheather Ch 7.3)

- Internal measures
- External measures

2. Putting it all together: strategies for model building and evaluation

- Bias-variance tradeoff
- A case study

3. From EDA to Prediction

4. Multicollinearity

Aim 1 Predictive ability of linear models

- When our principal purpose is to construct a linear model that will be used to make future predictions, we assume that the:
  - Relationship between the response variable and explanatory variables will continue into the future
  - Problem: we can't test our model now on data that we might encounter in the future!
- Two possible strategies
  - Use 'internal' measures of predictive ability (PRESS)
  - Divide the data you have into two:
    - a **training set** that we use to construct the model using all subsets, variable selection, etc., and
    - a **test set** that we use to evaluate those models

PRESS – Predicted residual sum of squares

- How might we evaluate the predictive ability of a candidate model with explanatory variables $\boldsymbol{x}_C$?
  - Leave out the ith observation y_i and fit a model containing $\boldsymbol{x}_C$
  - Predict the value of the y_i that has been left out, which we denote by $\hat{y}_{-i}$
  - Calculate the residual $y_i - \hat{y}_{-i}$, and square it
  - Repeat the procedure so that every observation is left out once

DAAG pakage in r will do this

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{-i})^2$$

PRESS – Predicted residual sum of squares

- After a bit of algebra, it is possible to show that

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{-i})^2 = \sum_{i=1}^{n} \left(\frac{\hat{e}_i}{1 - h_{ii}} \right)^2$$

hii is the leverage

- So, for a linear regression, we don't actually have to iteratively remove an observation, fit a model, and predict the left-out observation
- For a linear regression, PRESS depends only on quantities from the model fitted to **all** the data:
  - residuals $\hat{e}_i$ and
  - leverage h_{ii},

both of which we have already seen!

PRESS for all subsets regression

- The disadvantage to calculate PRESS in `leaps` since resulting object does not contain the fitted models
  - If it did, we could use PRESS along with the other criteria
  - Once you have identified a small subset of models from all subsets regression, you can fit those manually and calculate PRESS
- Function *`press`* in package DAAG

Example 1 (Body Weight): PRESS

- PRESS is most useful when we do not have a lot of observations in the data at hand
- Let's compare the PRESS of the different models selected by all subsets (12 variables) and forward/backward selection
- Model selected by forward selection has smallest PRESS, even smaller than model with all variables
  - Having too many variables can worsen predictive ability

```
press(lm.as) # all subsets, 12 variables
```

```
[1] 2352.882
```

```
press(lm.forward) # forward selection, 16 variables
```

```
[1] 2329.183
```

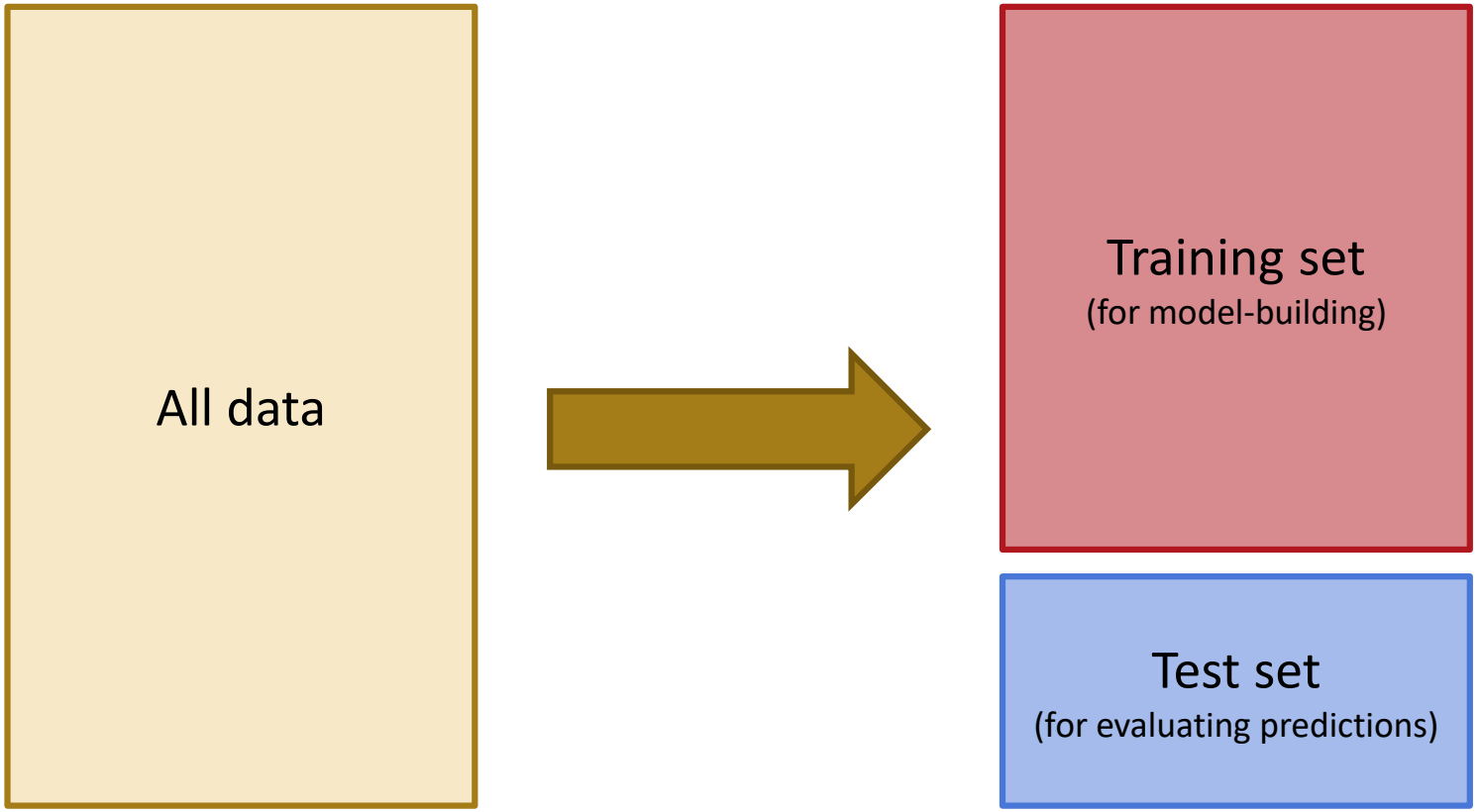
```
press(lmall) # all 25 variables
```

```
[1] 2383.939
```

Predictive ability: 'future data'

- A standard approach to assessing the predictive ability of different regression models is to evaluate their performance on a new data set
- New data set is one that has not been used for developing models – but where is it going to come from?
- In practice, this can be achieved (*if we have a large enough data set to work with*) by **randomly splitting** the existing data set into a **training set** and a **test set**
- Carry out model-building solely on training set and calculate squared prediction error on test set

Predictive ability: 'future data'



Constructing training/test set

- Only rules of thumb for the fraction of data to set aside
  - Need to ensure a compromise between having enough observations to construct models, and enough observations to assess predictive ability
- Set aside roughly 20% of the observations
  - e.g., `sample(250, 50)` will generate a vector of 50 numbers drawn at random from the numbers 1 – 250
  - These are the observations that will constitute the test set; the rest will constitute the training (model-building) set
- Once you have constructed a small set of candidate models, predict the observations in the test set, calculate the squared prediction error, and construct a plot of the predicted against actual values

Prediction error

- Suppose that we have constructed a model on a training set and have made predictions $\hat{y}_i^t$ for n_t test set observations with observed values y_i^t
- Then, the root mean squared error of prediction is defined as

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_t}\left(y_i^t - \hat{y}_i^t\right)^2}{n_t}}$$

- Keep in mind, however, that random splits of the data set will yield slightly different $RMSEP$s

Example 2: body weight data

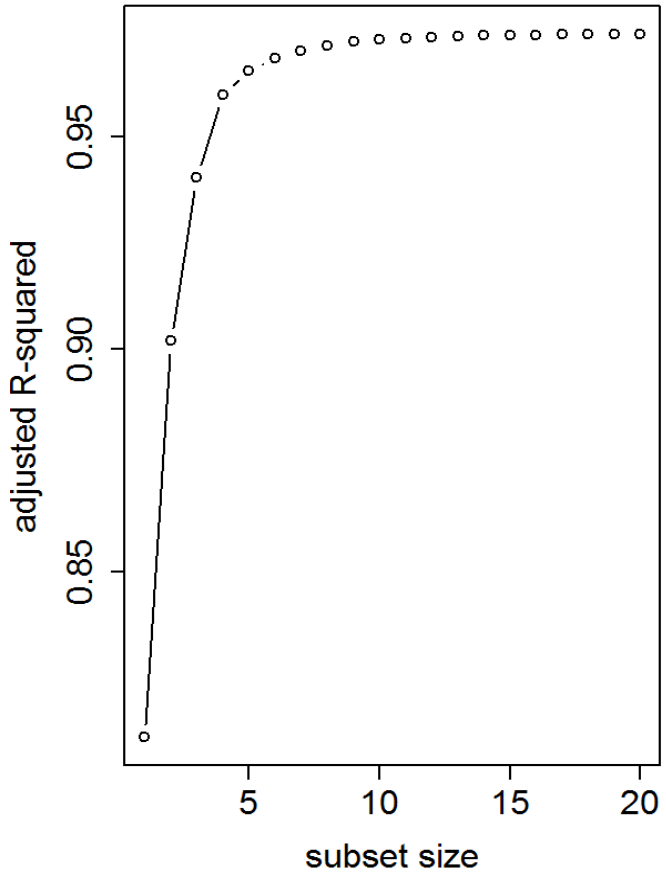
- Objective is to predict body weight from using 24 potential covariates:
  - Chest depth
  - Chest diameter
  - Knee diameter
  - Shoulder girth
  - …
  - Age
  - Height
  - Gender
- Covariates are highly correlated
- 507 observations in data set

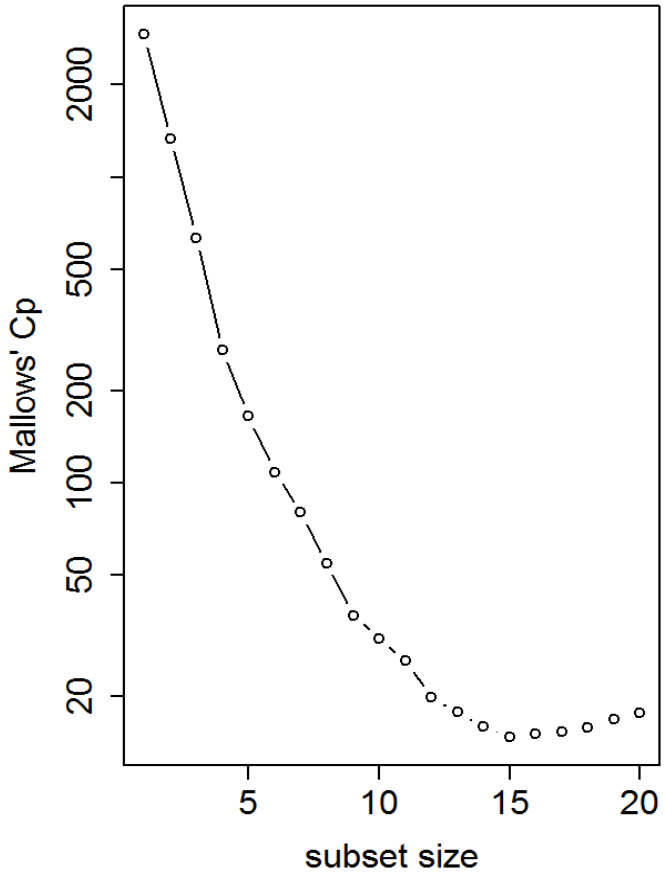
All subsets: body weight

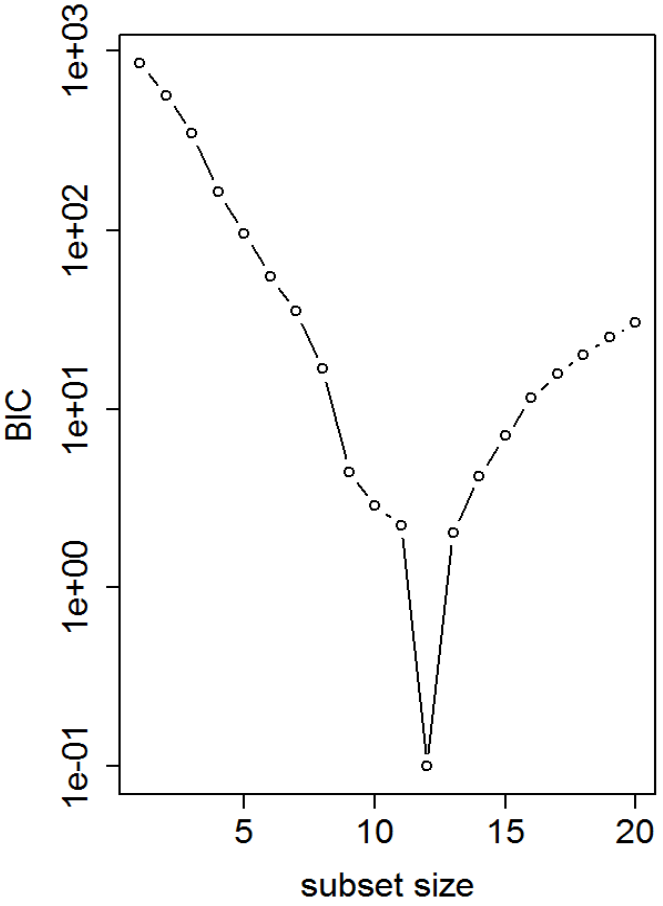
```
require(leaps)
AllSubsets <- regsubsets(Weight ~ ., nvmax
= 20, data = BodyMeasurements)
AllSubsets.summary <-
summary(AllSubsets)
```

	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				*				*			*	*		*	*		*		*			*	*	

All subsets: body weight







Summary: all subsets for body weight

- Sometimes, there are several models that yield roughly the same performance as assessed by adjusted-R^2, AIC, BIC, C_p
- For the body weight data set, it appears that models with 12–15 variables, each of which has the smallest RSS of all models with 12, 13, 14, and 15 variables, fit equally well and provide a useful compromise between 'goodness-of-fit' and complexity as measured by number of variables in the model
- For now, will keep only the 12-variable model suggested by BIC, but keep in mind that in practice, we might want to carry forward more models

All subsets: body weight (12 variables)

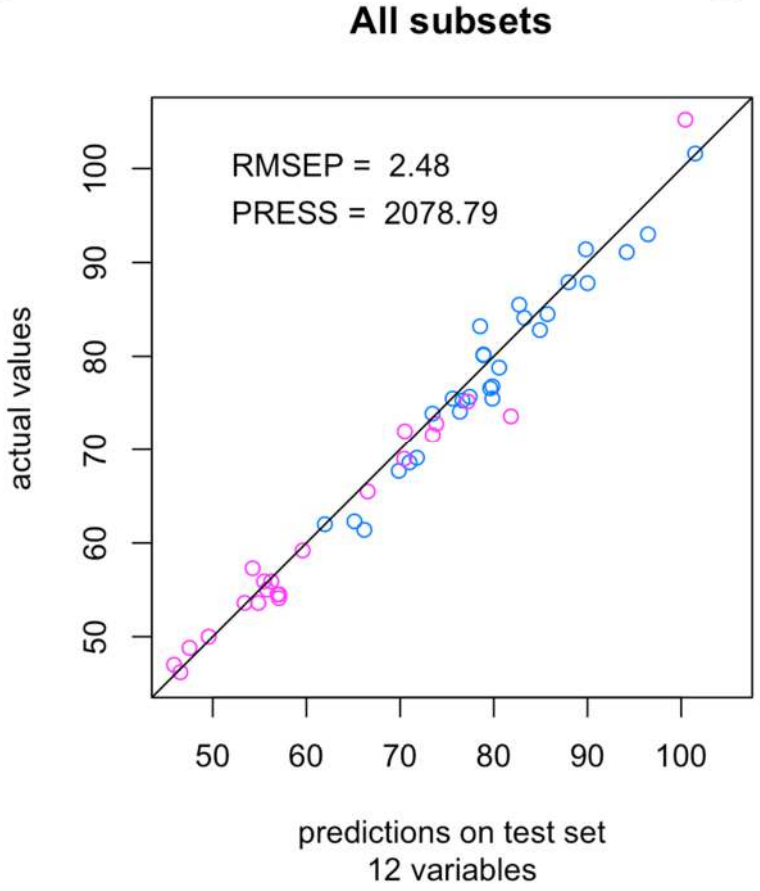
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -122.26113    2.52608 -48.400  < 2e-16 ***
CheDep         0.26584    0.06879   3.864 0.000126 ***
KneDia         0.64053    0.11727   5.462 7.47e-08 ***
ShoGir         0.08655    0.02825   3.063 0.002308 **
CheGir         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
```

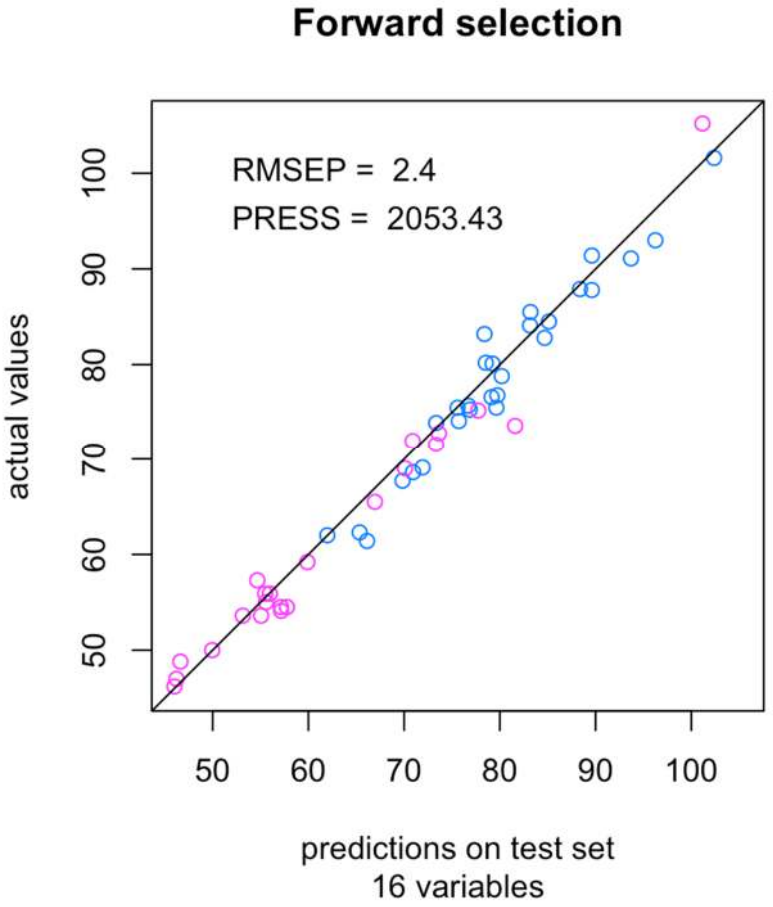
Body weight: forward selection

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -121.11380    2.53409 -47.794  < 2e-16
WaiGir         0.37465    0.02571  14.572  < 2e-16
KneGir         0.18622    0.07389   2.520 0.012044
Height         0.29486    0.01728  17.064  < 2e-16
ThiGir         0.26130    0.04914   5.317 1.60e-07
ForGir         0.52963    0.10140   5.223 2.60e-07
CheGir         0.14703    0.03560   4.130 4.26e-05
CalGir         0.34008    0.06149   5.531 5.19e-08
HipGir         0.20624    0.03910   5.275 2.00e-07
KneDia         0.45002    0.12726   3.536 0.000444
Age           -0.05606    0.01183  -4.739 2.82e-06
CheDep         0.28004    0.06922   4.046 6.06e-05
ShoGir         0.07936    0.02920   2.717 0.006813
Gender        -1.42332    0.48909  -2.910 0.003777
BiiDia         0.09451    0.05663   1.669 0.095752
CheDia         0.12415    0.07673   1.618 0.106283
ElbDia         0.26400    0.16630   1.588 0.113040
```

- Model selected by forward selection contains 16 (out of 25) variables, more than that suggested by all subsets regression

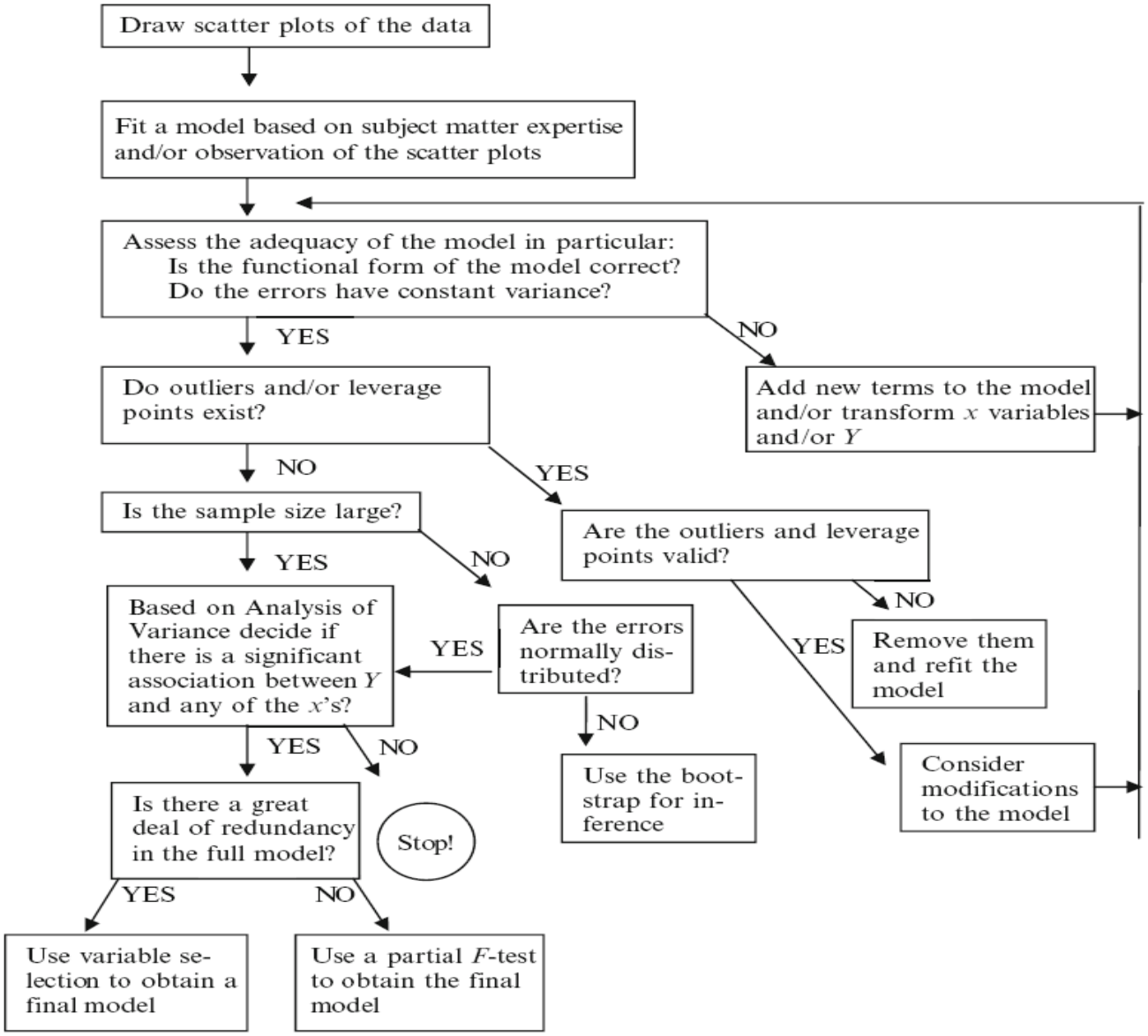
Body weight: evaluation of predictive capacity





Summary

- Variable subset selection methods can provide us with a smaller set of models to carry forward for predictive evaluation
- Two kinds of predictive evaluation:
  - Internal: PRESS
  - External: RMSEP on test set



Aim 2 Strategies for Modelling

Sheather (2009), p. 252

Empirical model building – cautions

- In most studies, we don't know the underlying 'true' model, or even if one exists – a linear model might be 'good enough'
- Any model that we come up with will be a useful approximation, i.e., we might use it purely for prediction purposes and not try to interpret it physically
- Always incorporate outside knowledge and don't forget diagnostic checks
- And remember,

"**All models are wrong, some are useful**"

G.E.P. Box (1920 – 2013)

2.1 Bias-variance trade-off

- In general, we can express an additive model as

$$y = f(X) + \epsilon, \qquad X = (X_1, X_2, \ldots, X_p)$$

- Our best estimate of the response is $\hat{Y} = \hat{f}(X)$, and for any new data (x_0, y_0), our prediction is $\hat{y}_0 = \hat{f}(x_0)$
- If we had a large number of test observations, we could compute the average squared prediction error

$$\text{Ave}\left(y_0 - \hat{f}(x_0)\right)^2$$

- We want to select a model for which this quantity is as small as possible

Bias-variance trade-off

- It is possible to show that the mean squared error averaged over a large number test sets can be decomposed as

$$MSE = E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}\left(\hat{f}(x_0)\right) + \left[\mathrm{Bias}\left(\hat{f}(x_0)\right)\right]^2 + \mathrm{Var}(\epsilon)$$

- *Variance* refers to the amount by which $\hat{f}$ would change if we estimated it using different training sets
  - In general, more flexible methods (or more complex models) can have higher variance
- *Bias* refers to the error that is introduced by approximating a real-life problem, which may be very complicated, by a much simpler model
- $Var(\epsilon)$ is the *irreducible* error – it provides an upper bound on the accuracy of our prediction for the response

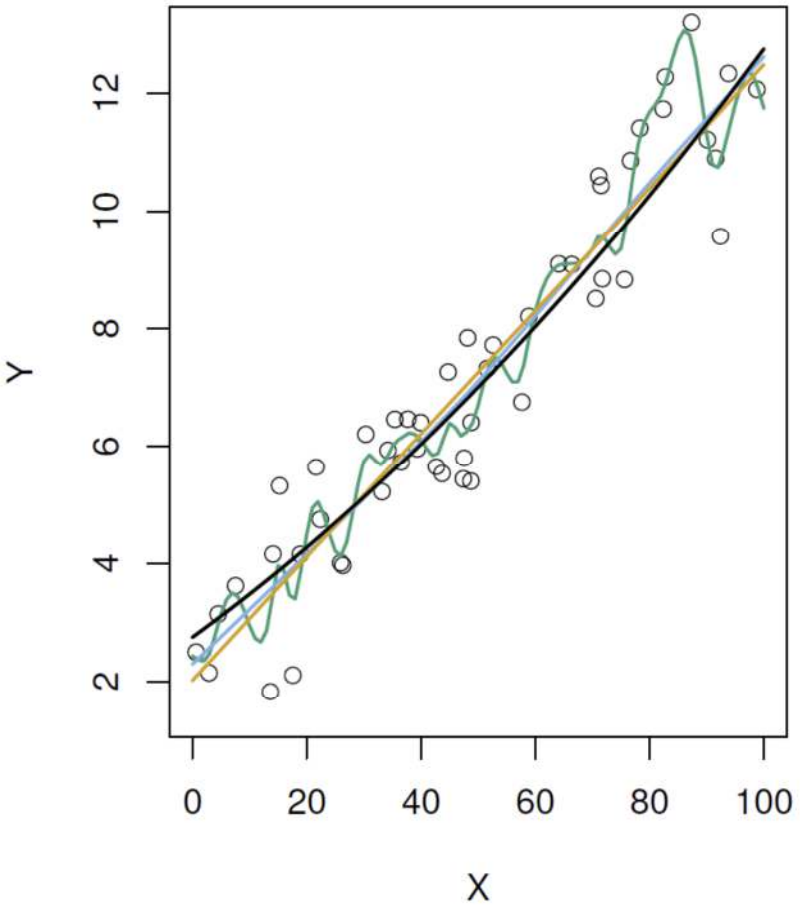
Bias-variance trade-off

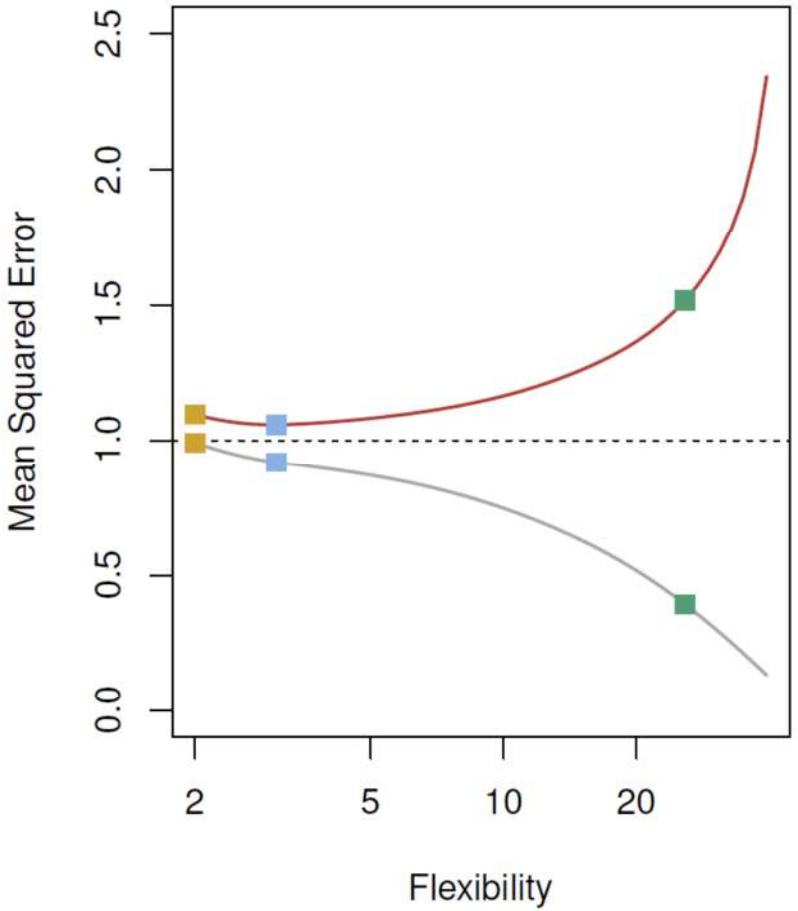
- It is possible to show that the mean squared error averaged over a large number test sets can be decomposed as

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}\left(\hat{f}(x_0)\right) + \left[\mathrm{Bias}\left(\hat{f}(x_0)\right)\right]^2 + \mathrm{Var}(\epsilon)$$

- **As the flexibility of $\hat{f}$ *increases*, its variance increases, and its bias decreases.**
- **As the flexibility of $\hat{f}$ *decreases*, its variance decreases but its bias increases.**
- **Choosing flexibility based on average test error amounts to a bias-variance tradeoff.**

Example 3: Bias-variance tradeoff



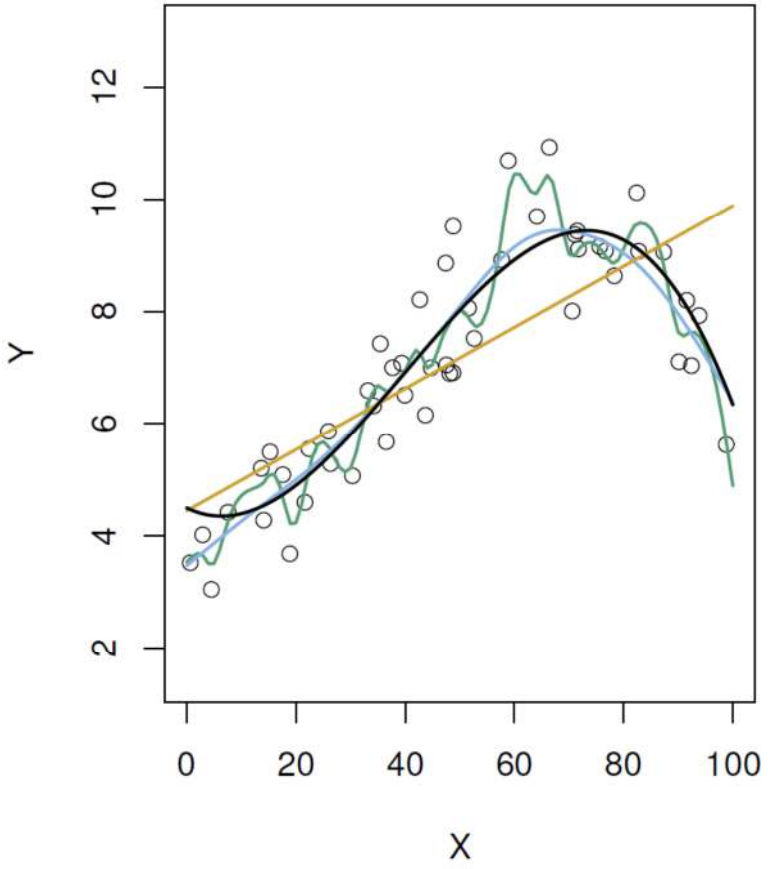


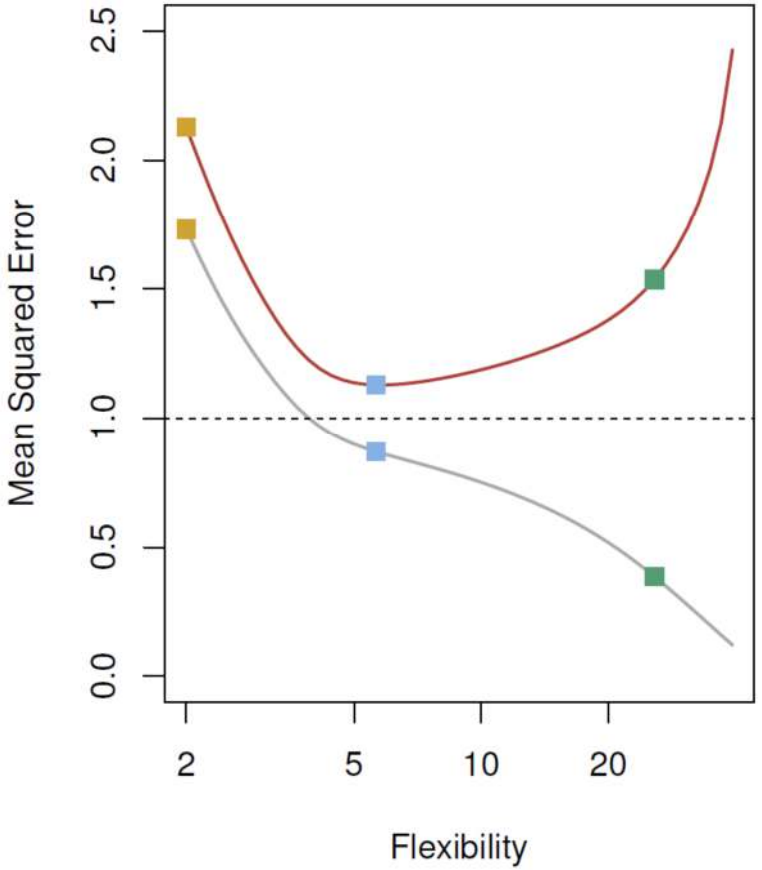
Red Curve Variance

Grey curve - Bias

FIGURE 2.10 (James et al 2023 Ch 2.2) Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel

Example 4: Training -Testing MSE





Red Curve Test MSE

Grey curve – Training MSE

FIGURE 2.9. (James et al 2023 Ch 2.2) Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the three fits shown in the left-hand panel.

2.2 Issues when there are many explanatory variables

- Although we may be seeking the most important set of “predictors” to answer our question of interest at the end of the day when we have many explanatory variables to choose from the actual chosen set is often one of possibly many “good” sets.
- So we should choose a set the best suits our objective.

- A “most” important point should be made when deciding the best subset whether we are using
  - “drop/partial in F-test” or
  - Cp Mallow’s criteria or
  - BIC/AIC via backwards selection, forward selection, or brute force

is that any variable selection process should be sensitive to the objectives the study

Possible Objectives

- Often with observational studies the main purpose is prediction and we don’t need to over interpret our explanatory variables.
- The various selection techniques may produce a few models so the clinician may select the set that best can be used for future prediction.

Selecting from many Variables

- Know your (or clinician's/researcher's/scientist's) objectives.
- Look at the list of "predictors" and exclude those that are not useful for the objectives.
- Do the exploratory data analysis looking for correlations and obvious assumption violations
- The former should show us if any transformations are needed
- Fit a full model and hence do our residual analysis seeking outliers and other influential observations.
- Now use our selection methods and choose "the best set"
- Write up our findings.

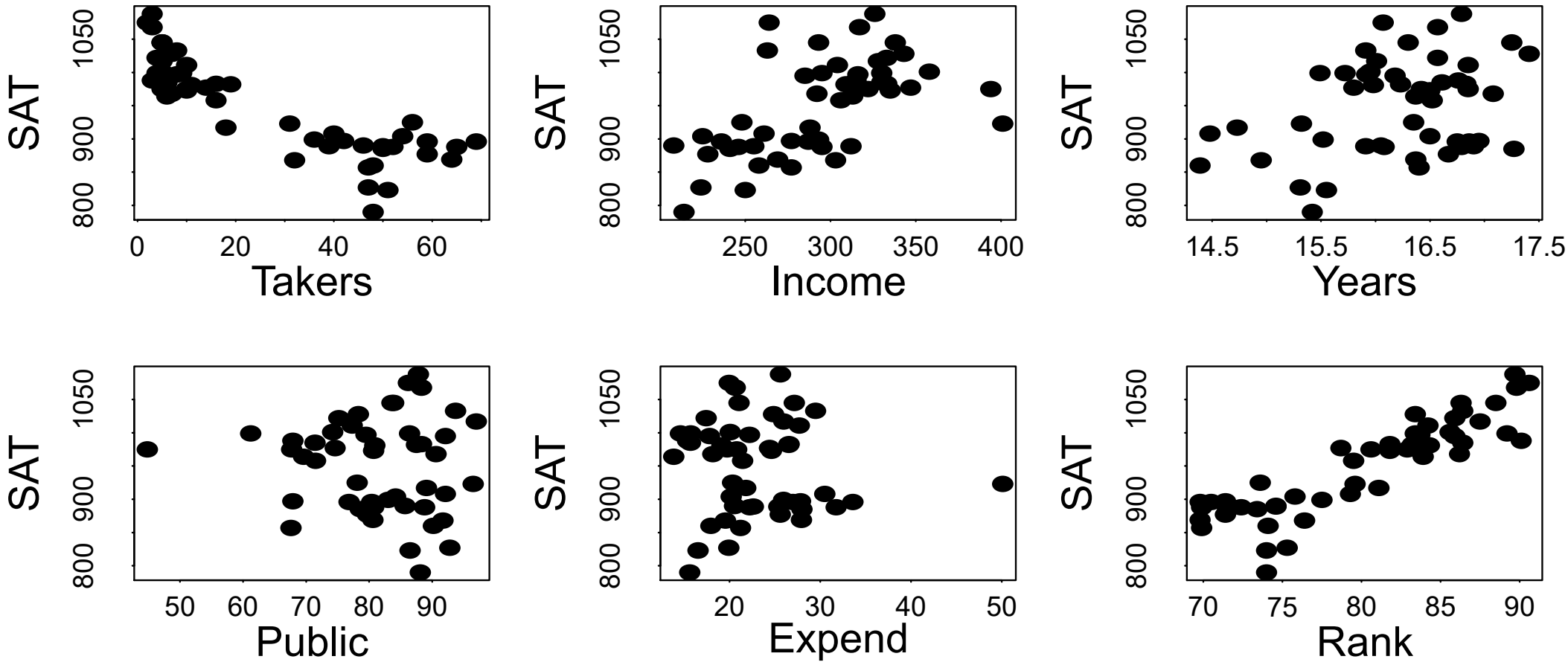
Example 5 Case study: SAT (Scholastic Aptitude Test) dataset

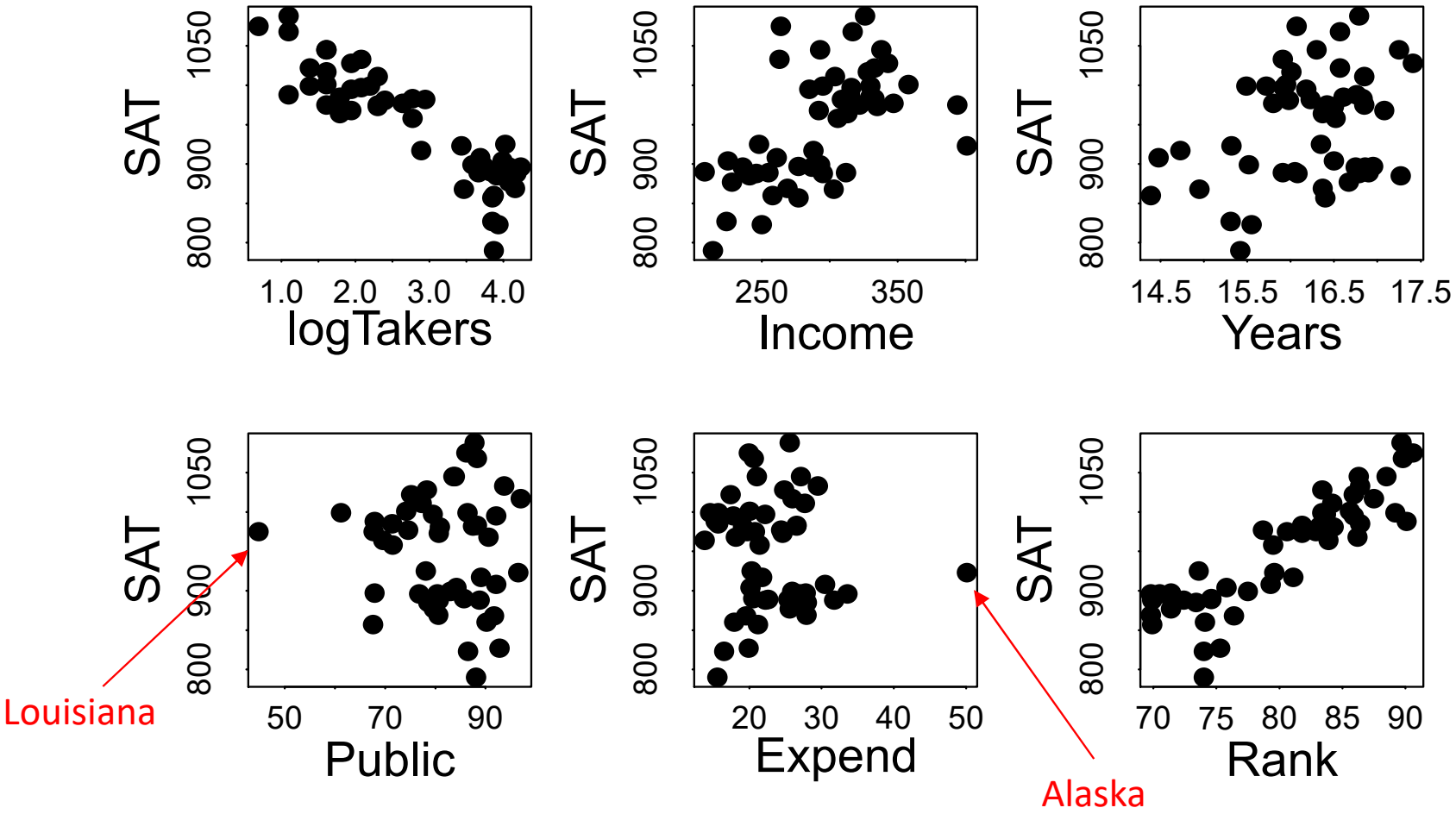
The Statistical Sleuth: A Course in Methods of Data Analysis, 3rd Edition
Fred Ramsey; Daniel Schafer

– observational study of US state-by-state SAT scores discussed in Statistical Sleuth
– researchers set out to assess the extent to which the compositional /demographic and school-structured characteristics are implicated in SAT differences

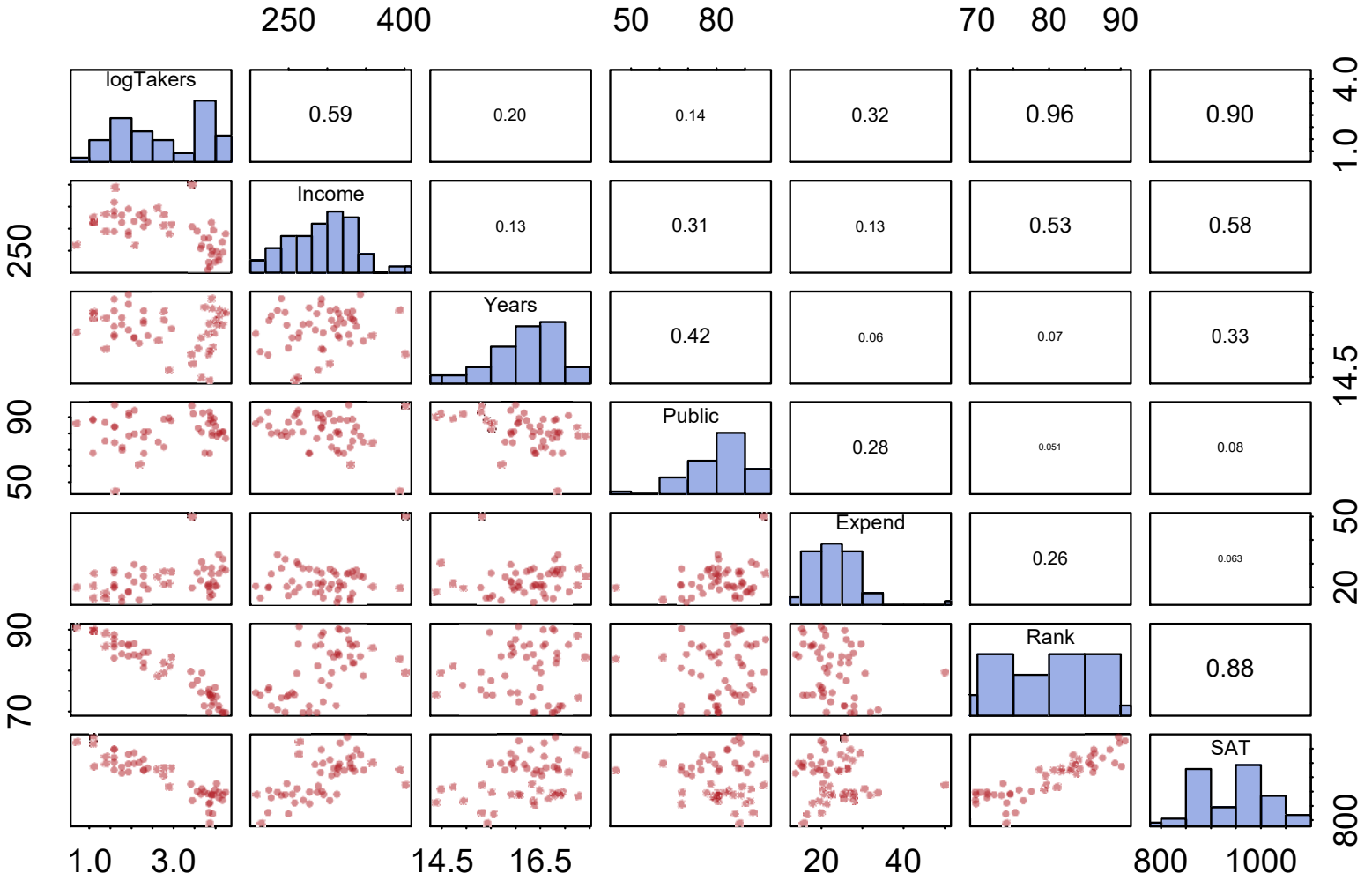
Takers : *% of eligible students in the state who took the College Entrance Exam*
Income: *median income of families of test-takers*
Years: *average number of years that test-takers had formal studies in social sciences, natural sciences and humanities*
Public: *% of test-takers who attended public secondary schools*
Expend: *total state expenditure on secondary schools (in $100s per student)*
Rank: *median ranking of test-taker within their secondary school classes*

Scatter plot of dependent variable against explanatory variables to check if transformations are needed



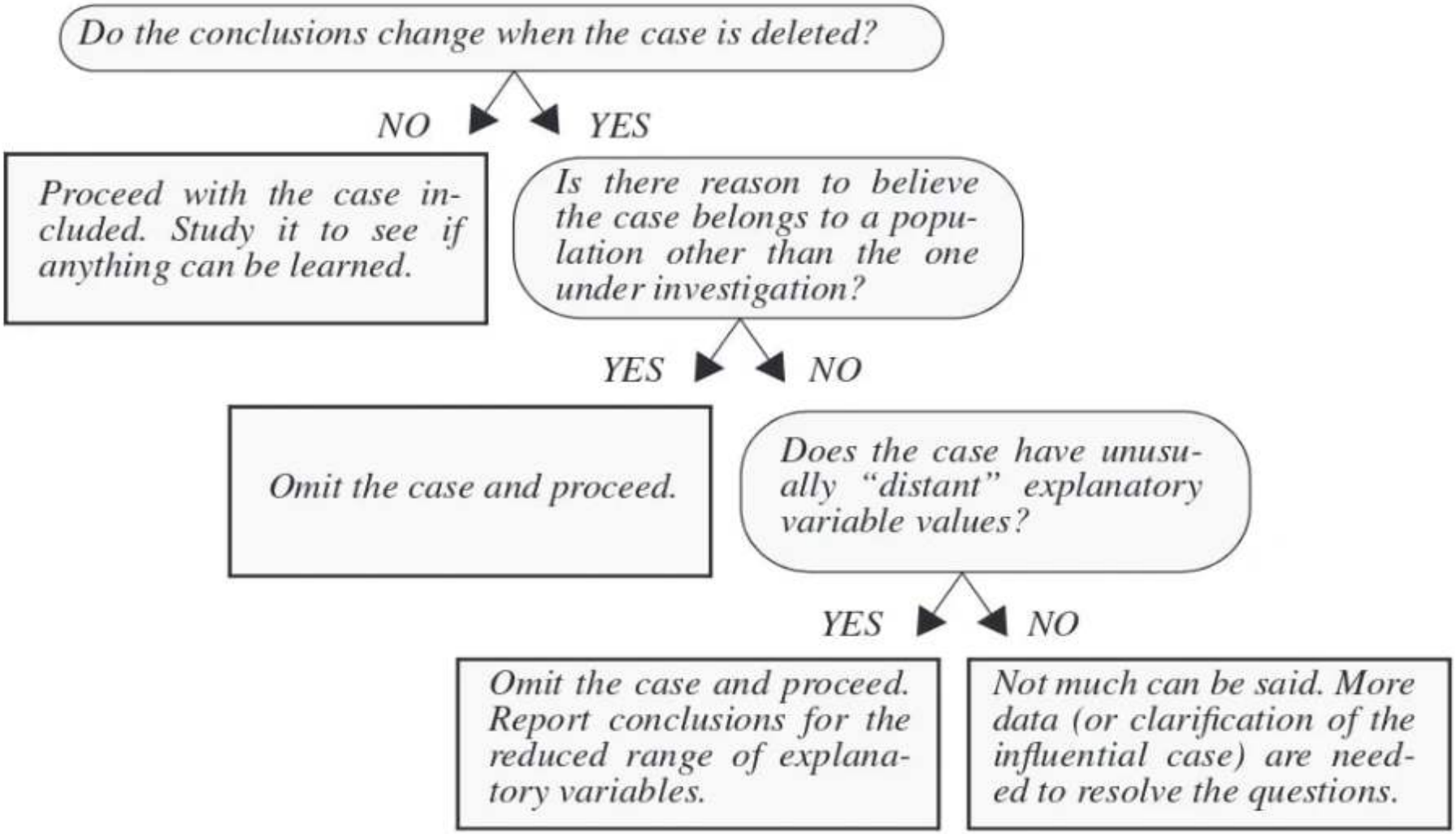


- Re-assessed with transformed Takers
- 2 potential outliers

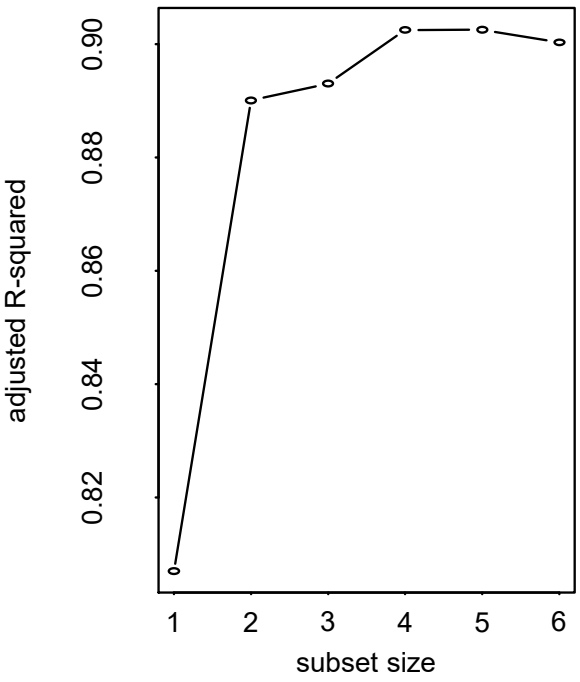


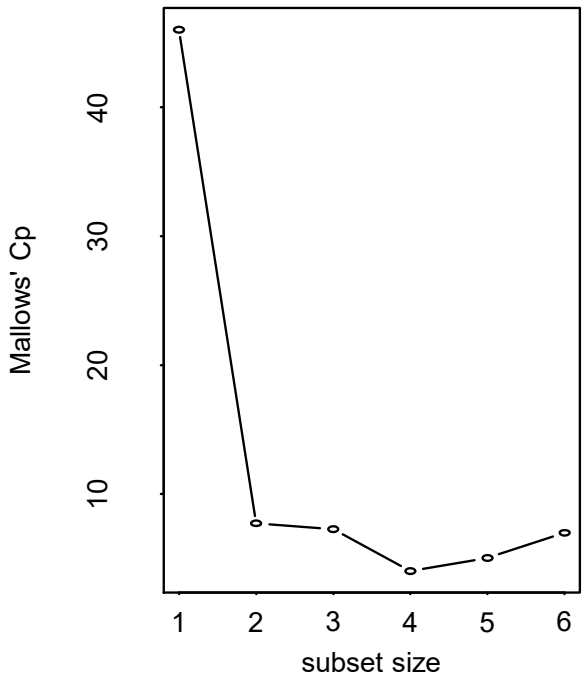
“pairs” plot – useful for also checking correlation between the explanatory variables (see the potential outliers still)

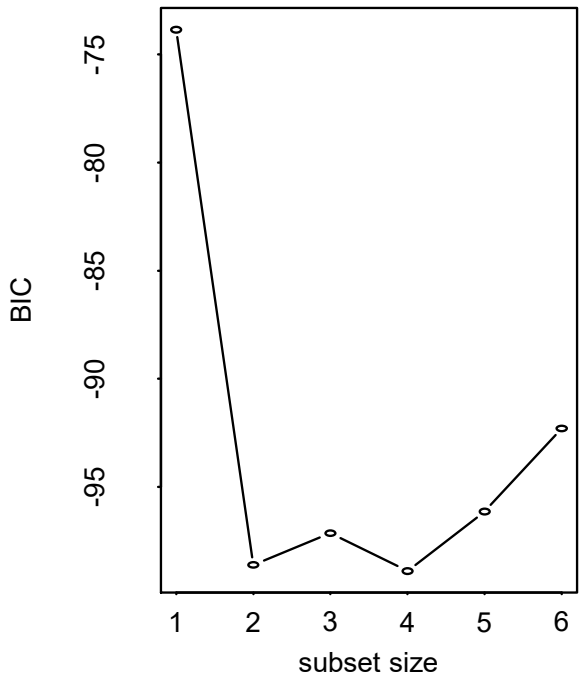
DISPLAY 11.8 A strategy for dealing with suspected influential cases



All subset selection (with Alaska removed)







Appears the models with either the 2 best or the 4 best of the explanatory variables are worth carrying forward to explore further

Best 2 variables:

```
lm(formula = SAT ~ logTakers + Expend, data = SATdat)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1028.5818    16.7253  61.499  < 2e-16 ***
logTakers    -66.1695     3.3505 -19.749  < 2e-16 ***
Expend         4.6049     0.7619   6.044 2.49e-07 ***
---
Residual standard error: 23.7 on 46 degrees of freedom
Multiple R-squared:  0.8947,   Adjusted R-squared:  0.8901
F-statistic: 195.4 on 2 and 46 DF,  p-value: < 2.2e-16
```

Best 4 variables:

```
lm(formula = SAT ~ logTakers + Expend + Years + Rank, data = SATdat)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 399.1145   232.3716   1.718  0.09291 .
logTakers   -38.1005    11.9152  -3.198  0.00257 **
Expend        3.9957     0.7642   5.228 4.52e-06 ***
Years        13.1473     5.4778   2.400  0.02069 *
Rank          4.4003     1.8989   2.317  0.02520 *
---
Residual standard error: 22.32 on 44 degrees of freedom
Multiple R-squared:  0.9107,   Adjusted R-squared:  0.9025
F-statistic: 112.1 on 4 and 44 DF,  p-value: < 2.2e-16
```

Forward selection:

```
step(lm0, scope = formula(lm.all), direction = "forward", trace = 0)

Call:
lm(formula = SAT ~ logTakers + Expend + Years + Rank, data = SATdat)

Coefficients:
(Intercept)     logTakers        Expend         Years          Rank
    399.115       -38.100         3.996        13.147         4.400
```

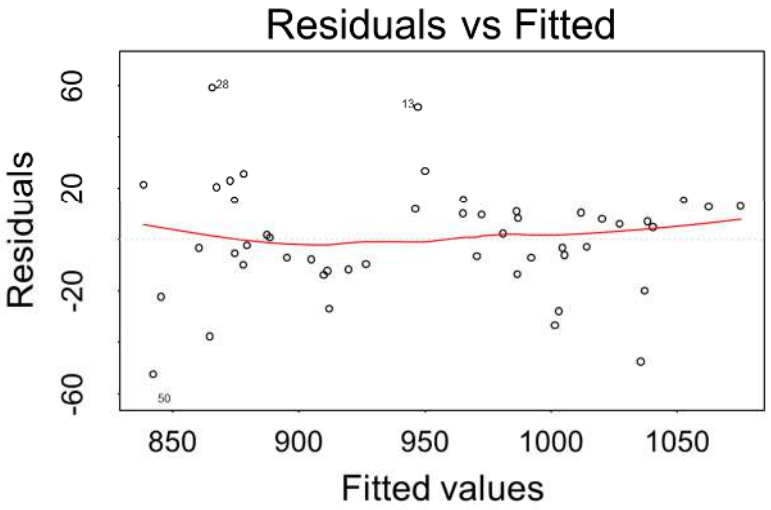
Backward selection:

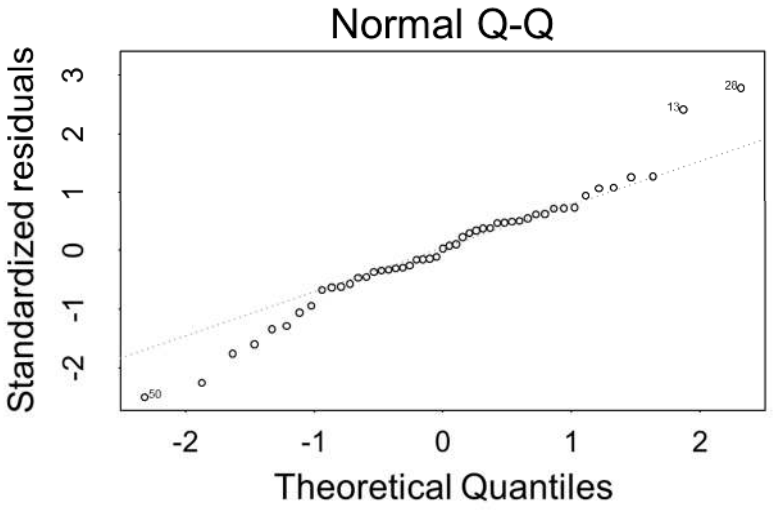
```
step(lm.all,  direction = "backward", trace = 0)

Call:
lm(formula = SAT ~ logTakers + Years + Expend + Rank, data = SATdat)

Coefficients:
(Intercept)     logTakers         Years        Expend          Rank
    399.115       -38.100        13.147         3.996         4.400
```

Same conclusion from backward and forward

Diagnostic plots: *plot(fit4)*

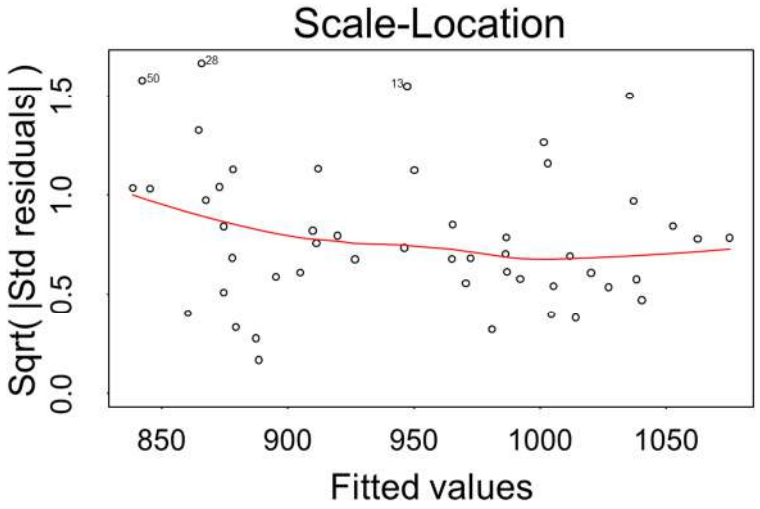
Any non-linear pattern? *Reasonably flat red line so ok*

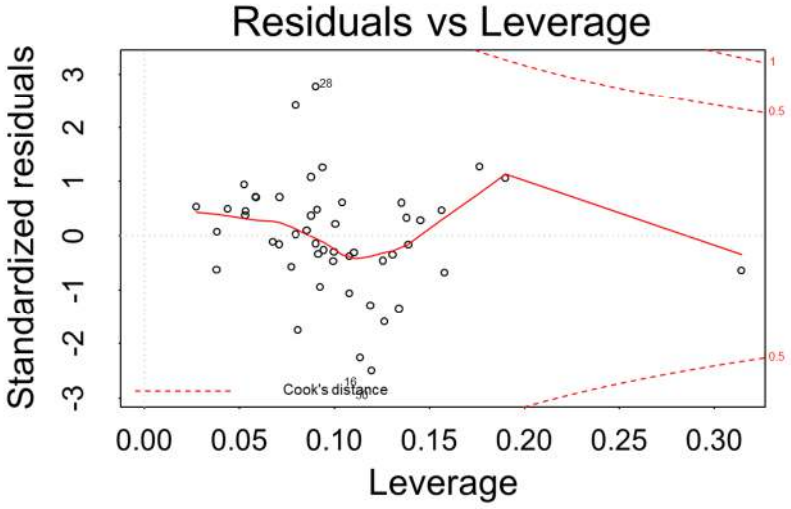




Is normality of residuals feasible? *Deviation in the tails but bulk on quite straight line*

Is the assumption of homoscedasticity (constant variance) reasonable? *Reasonably flat red line so ok*





Is there evidence of any particularly influential points?

No points beyond the Cook's distance of 0.5 (the dashed red line) so ok

Using linear regression analysis to address key question of interest.

- The data examined here had been used to rank the states on their success in secondary education, but this is problematic due to selection bias. For example, in some states only the best students may have sat the test so there is a self-selected sample from contributing schools.

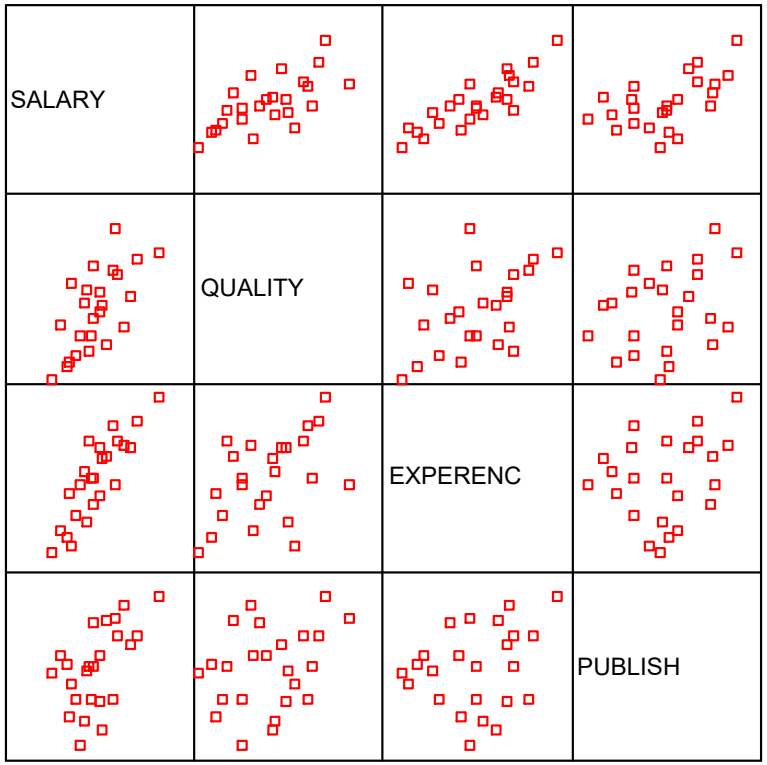
- Ranking states after subtracting out the effects of different proportions of students sitting the test, and their different median class rankings goes some way to addressing this issue.

Exploratory analysis

- It would be reasonable to ask are there any other variables associated with SAT scores? The variable selection techniques illustrated in this class come into play to answer this. But it really only makes sense to build on models that accommodate the obvious selection bias.

Aim 4 Multi-collinearity

In most real data sets with multiple explanatory variables, these variables will be correlated with each other to some degree



Example 6 This graph shows n=24 salaries of mathematics lecturers, plus a quality of work score, years of experience and publication record.

The variables all seem related to salary, but also to each other.

Correlations

		SALARY	QUALITY	EXPERENC	PUBLISH
SALARY	Pearson Correlation	1.000	.667**	.859**	.558**
	Sig. (2-tailed)	.	.000	.000	.005
	N	24	24	24	24
QUALITY	Pearson Correlation	.667**	1.000	.467*	.323
	Sig. (2-tailed)	.000	.	.021	.124
	N	24	24	24	24
EXPERENC	Pearson Correlation	.859**	.467*	1.000	.254
	Sig. (2-tailed)	.000	.021	.	.232
	N	24	24	24	24
PUBLISH	Pearson Correlation	.558**	.323	.254	1.000
	Sig. (2-tailed)	.005	.124	.232	.
	N	24	24	24	24

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

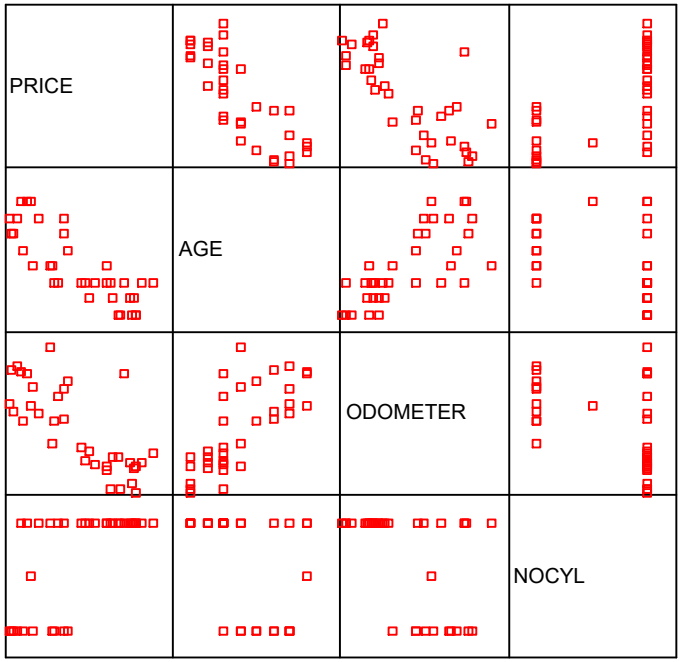
- These variables are not strongly related to each other, but seem strongly related to salary and so can be put in the same regression model.

- When explanatory variables in a data set are highly correlated with each other, problems can arise in fitting a regression model.

Correlations

		PRICE	AGE	ODOMETER	NOCYL
PRICE	Pearson Correlation	1.000	-.821**	-.753**	.658**
	Sig. (2-tailed)	.	.000	.000	.000
	N	35	35	35	35
AGE	Pearson Correlation	-.821**	1.000	.729**	-.471**
	Sig. (2-tailed)	.000	.	.000	.004
	N	35	35	35	35
ODOMETER	Pearson Correlation	-.753**	.729**	1.000	-.481**
	Sig. (2-tailed)	.000	.000	.	.003
	N	35	35	35	35
NOCYL	Pearson Correlation	.658**	-.471**	-.481**	1.000
	Sig. (2-tailed)	.000	.004	.003	.
	N	35	35	35	35

**. Correlation is significant at the 0.01 level (2-tailed).



For the *used car* example, Age and Odometer reading are highly correlated, also Age and NoCyl and Odometer and NoCyl are correlated.

Given this, we shouldn't put Age and Odometer in the same model. They are measuring essentially the same quantity

When explanatory variables in a data set are highly correlated with each other, problems can arise in fitting a regression model. Why?

Consider the 2 variable linear model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

We can show that

$$E(\hat{\beta}_2) = \beta_2 + r_{12}\beta_1; Var(\hat{\beta}_2) = \sigma^2 \left(\frac{1}{1 - r_{12}^{\ 2}} \right) \left(\frac{1}{\sum (X_{2i} - \overline{X}_2)^2} \right)$$

where the correlation *r* is the correlation between X_1 and X_2

If this correlation is large then $\left(\frac{1}{1 - r_{12}^2} \right)$ will also be large.

When explanatory variables in a data set are highly correlated with each other, problems can arise in fitting a regression model. Why?

- We can show that $E(\hat{\beta}_2) = \beta_2 + r_{12}\beta_1; Var(\hat{\beta}_2) = \sigma^2 \left(\frac{1}{1 - r_{12}^2} \right) \left(\frac{1}{\sum (X_{2i} - \overline{X}_2)^2} \right)$

where the correlation *r* is the correlation between X_1 and X_2

- Compare this with the case $r_{12} = 0$ (unbiased and the variance is a standard variance)

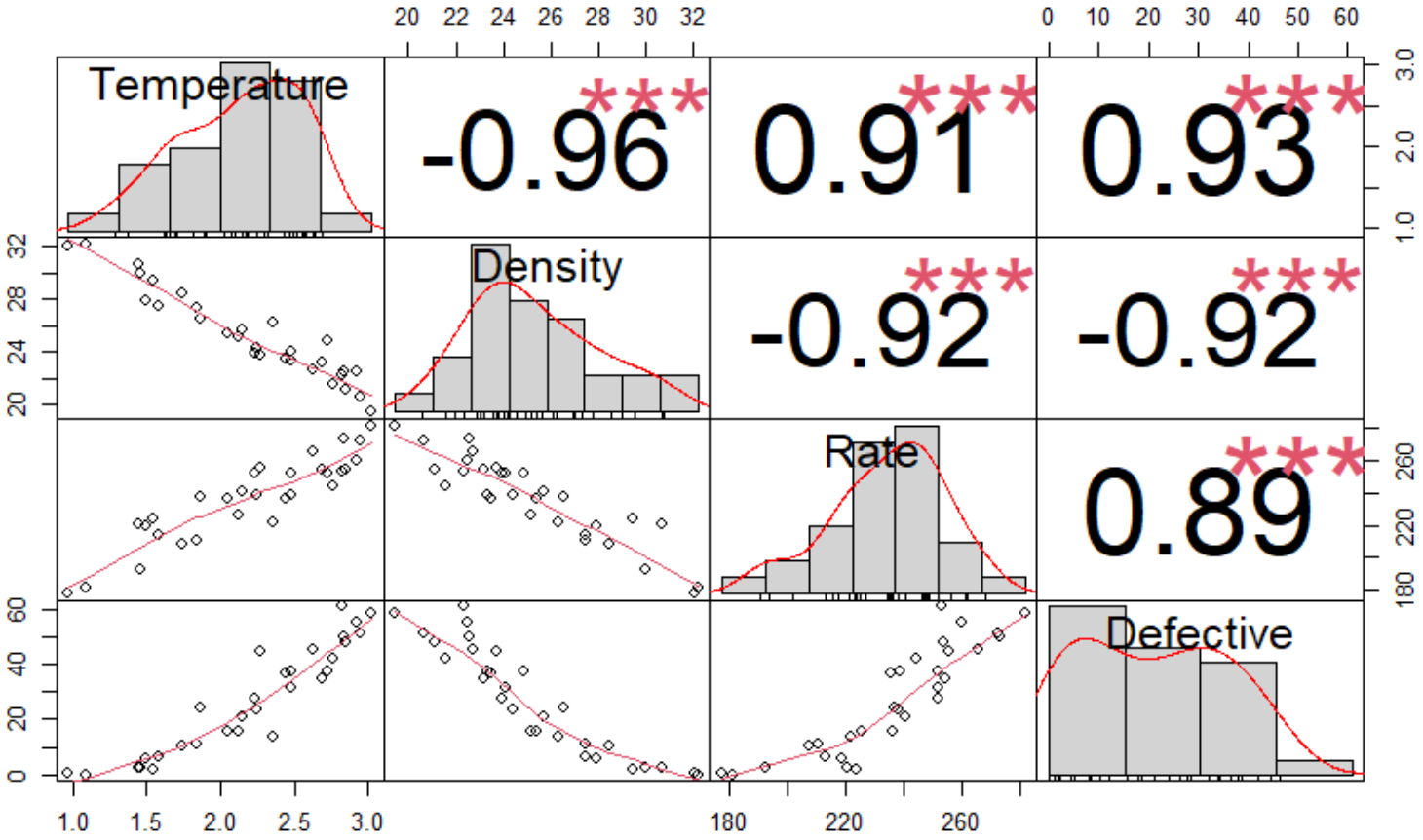
- Clearly, **when explanatory variables are highly correlated**:
  - Their variances become inflated and
  - parameter estimates may be biased.

Example 7 Defect data

- The data frame `Defects` provides data on the average number of defects per 1000 parts (`Defective`) produced in an industrial process along with the values of other variables (`Temperature`, `Density`, and `Rate`).
- The production engineer wishes to construct a linear model relating `Defective` to the potential predictors.

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

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
```

The scatterplot matrix



Defective has:

- positive, linear relationships with Temperature and Rate as an explanatory.
- negative, linear relationships with Density as an explanatory.
- some relationships among the explanatories that related to multicollinearity

VIF (variance inflation factor)

- $var(\hat{\beta}_j) = \frac{1}{1-r_{12}^2}\frac{\hat{\sigma}^2}{s_{x_j}^2}$ VIF > 5, multicollinearity

where r_{12} denote the correlation between x1 and x2 and denote $s_{x_j}^2$ the variance of xj. The term $\frac{1}{1-r_{12}^2}$ is called a variance inflation factor (VIF).

- Using R

```
library(car)
vif(def.lm)
Temperature     Density       Rate
  13.431614   14.508872   6.642619
```

- The associated regression coefficients are poorly estimated due to multicollinearity.