Experimental Design and Statistical Methods

Workshop

MULTIPLE LINEAR REGRESSION

Jesús Piedrafita Arilla

jesus.piedrafita@uab.cat

Departament de Ciència Animal i dels Aliments

UAB

Items

- Multiple linear regression
  - Model and matrix notation
  - General protocol
  - Model fit
  - Multicollinearity
  - Analysis of residuals
  - Influence
  - Choice of the best model
  - Variable selection procedures
- Basic commands
  - seq
  - vif, tol
  - cor (multiple)
  - step
- Libraries
  - car (scatterplot)
  - leaps
  - faraway (cp plot)

Why multiple linear regression?

The general purpose of multiple regression (the term was first used by Pearson, 1908) is to learn more about the relationship between several independent or predictor variables (regressors) and a dependent variable.

Multiple linear regression attempts to model this relationship by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y.

In the multivariate case, when there is more than one independent variable, the regression line cannot be visualized in the two dimensional space, but can be computed just as easily.

The general **model** is now:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

where the betas are **partial regression** coefficients that represent the expected change (positive or negative) in response (dependent variable), per unit of change in a x_i with the other x's held constant. One or more x_i can be indicator variables, i.e., taking 0 or 1 values.

Matrix notation

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & x_{31} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & x_{32} & \dots & x_{k2} \\ 1 & x_{13} & x_{23} & x_{33} & \dots & x_{k3} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} & x_{3n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \dots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

This can be solved by least squares (as before) giving the normal equations:

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Multiple linear regression - General protocol -

1. Plot y against each x variable in turn. Curvature indicates the need for transformations. Linear plots indicate important predictors.
2. Check for evenness in variables using boxplots and histograms.
3. Fit the regression equation, after transformation if necessary. Look at studentized residuals, leverages, Dffits, Dfbetas and Cook's D.
4. Perform diagnostic checks.
5. Check for multicollinearity, i.e., two or more predictor variables in a multiple regression model are highly correlated.
6. Investigate the possibility of reducing the number of predictor variables (variable selection procedures).
7. When an optimum regression equation is obtained, assess its value for the purposes for which it was intended. If it is used for prediction, calculate the confidence intervals or prediction intervals.

(Adapted from Fry, 1993)

Carcass and meat data (AM beef breed)

	ADG	LMAREA	CONF	FAT	INTRAMFAT	WHC	DRYMAT	COOKLOSS	TEXTWB	TENDERNESS
1	1.03	65.65	7	9	1.60	20.34	23.69	11.74	3.43	6.56
2	1.00	29.57	2	7	3.20	15.93	24.75	11.13	5.62	6.80
3	1.06	42.50	8	11	1.75	21.75	24.52	15.26	5.85	4.70
4	1.13	51.13	8	10	2.40	23.73	24.59	15.52	4.64	5.31
5	1.21	44.50	8	9	1.25	21.33	23.50	12.80	8.19	5.75
6	1.11	36.10	8	8	1.30	16.85	24.11	17.57	3.46	6.72
7	1.11	49.32	8	9	0.75	22.80	24.09	8.30	5.25	6.00
8	1.05	44.73	8	7	2.50	21.95	24.24	19.61	7.51	4.90
9	1.08	41.22	7	11	1.15	18.96	24.19	13.06	4.96	6.18
10	1.26	39.65	7	10	1.80	21.45	24.04	8.50	5.24	6.92
11	1.19	44.67	6	11	2.90	22.91	25.39	6.44	3.64	6.92
12	1.20	51.40	9	8	3.95	23.68	25.39	13.21	2.76	5.93
13	1.12	41.70	8	10	1.50	23.20	24.81	11.80	3.90	6.38
14	1.13	53.50	8	6	4.25	19.96	25.22	16.32	3.84	6.64
15	0.96	37.45	8	8	2.50	19.98	26.04	16.06	4.12	6.08
16	1.01	39.50	7	7	3.15	20.33	24.48	16.99	3.99	6.27
17	1.10	45.50	8	8	1.90	20.20	24.47	13.65	3.58	6.98
18	0.96	42.35	7	7	1.05	22.86	25.37	13.38	4.47	6.96
19	1.06	44.65	7	7	1.45	24.25	24.39	9.73	4.66	6.85
20	1.06	45.50	8	6	2.20	25.24	25.03	16.75	4.59	6.54
21	1.08	41.55	8	8	3.20	24.72	25.01	11.72	3.62	7.03
22	1.20	49.55	8	9	0.80	26.51	24.40	10.98	4.71	6.66
23	1.07	52.45	9	8	2.10	20.33	24.21	14.08	6.02	3.99
24	1.18	46.40	8	8	1.00	22.28	23.33	12.55	4.52	6.33
25	1.14	48.65	7	5	1.55	23.65	25.09	19.37	6.94	4.97
26	1.10	39.65	6	9	2.85	19.90	24.71	14.73	5.23	6.08
27	1.10	42.65	8	8	3.00	22.73	25.32	14.74	2.74	7.23
28	1.13	51.53	9	6	4.45	22.58	25.70	14.42	3.91	6.29
29	1.15	38.82	8	8	1.95	23.45	24.52	9.83	5.61	6.18
30	0.97	43.73	7	7	2.05	24.97	24.61	10.43	3.40	6.42
31	1.28	36.70	8	8	4.55	21.95	27.05	12.39	4.35	6.09
32	1.01	40.50	9	9	1.95	21.77	24.38	12.83	4.08	5.95
33	1.13	40.97	9	6	4.20	24.22	24.52	9.94	4.38	7.08
34	1.05	34.70	8	6	2.40	21.61	24.48	13.71	5.62	6.67
35	1.13	38.90	8	7	2.30	20.12	24.55	14.91	2.33	8.01

ADG: average daily gain

LMAREA: loin muscle area

CONF: conformation score

FAT: fat score

INTRAMFAT: intramuscular fat

WHC: water holding capacity

DRYMAT: dry matter

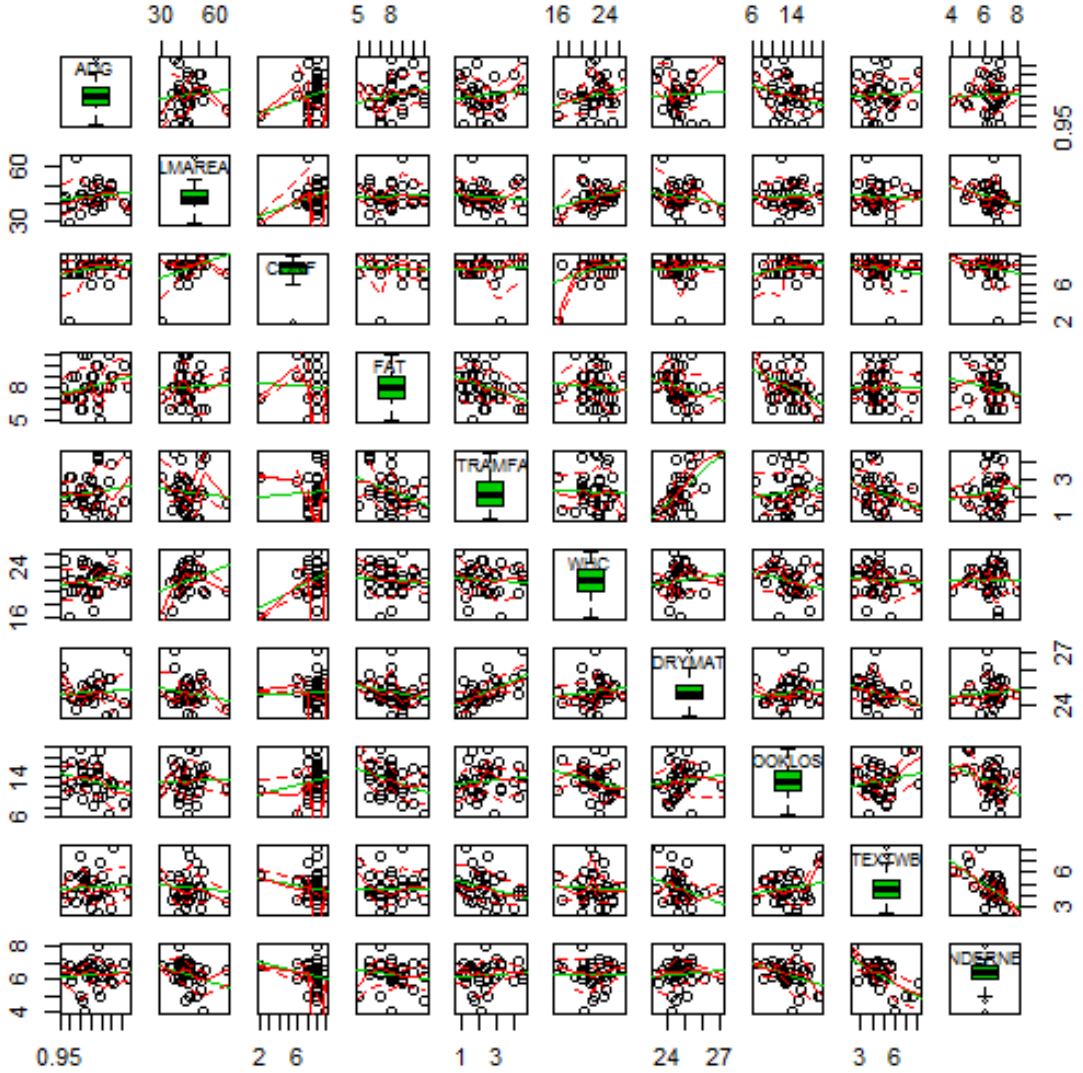
COOKLOSS: cooking losses

TEXTWB: Warner Blatzer shear force

TENDERNESS: meat tenderness

Assumptions of linearity, normality, homog. var.

```
> scatterplot.matrix(~ADG+LMAREA+CONF+FAT+INTRAMFAT+WHC+DRYMAT+COOKLOSS+
+ +TEXTWB+TENDERNESS, diag="boxplot")
```



In general, good agreement to normality: symmetrical boxplots (CONF with a outlier?).

No evidence of non-homogeneity of variance (even spread of points around each trend) or non-linearity.

Observe the relationship between tenderness and texture WB.

Multiple linear regression – model -

We assume that *k* (=9) independent variables can be associated to the response variable TENDERNESS in the Asturiana de la Montaña beef breed:

```
> TEND.LM<-lm(TENDERNESS~ADG+LMAREA+CONF+FAT+INTRAMFAT+WHC+DRYMAT+
> +       COOKLOSS+TEXTWB)
> summary(TEND.LM)
```

Observe that the sentence is similar to that of simple linear regression but adding variables with the + sign. All regressors can be represented by a dot (.).

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.08589    4.04187   3.980 0.000522 ***
ADG          2.41867 ÷  1.20541 =  2.007 0.055732 .
LMAREA      -0.03725    0.01400  -2.660 0.013451 *
CONF        -0.15123    0.08507  -1.778 0.087613 .
FAT         -0.20507    0.06618  -3.099 0.004756 **
INTRAMFAT   -0.11912    0.11372  -1.047 0.304925
WHC          0.01089    0.04950   0.220 0.827721
DRYMAT      -0.20156    0.16608  -1.214 0.236226
COOKLOSS    -0.06310    0.03644  -1.732 0.095650 .
TEXTWB      -0.47059    0.07360  -6.394 1.08e-06 ***
```

The significance of the beta coefficients are tested through a *t*-test. Three regressors are significant at the 5% level (fat, lmarea, textwb) and three more at the 10% level.

Multiple linear regression – model (cont) -

Some more information about the model is:

```
Residual standard error: 0.4716 on 25 degrees of freedom
Multiple R-squared: 0.7388,     Adjusted R-squared: 0.6447
F-statistic: 7.855 on 9 and 25 DF,  p-value: 2.065e-05
```

Observe that model d.f. equal the number of regressors (or number total of β minus 1).

The adjustment by the model is highly significant.

Around 64% of the variance of the dependent variable tenderness is explained by the regressors included in the model.

Multiple linear regression – residuals and influence -

```
> data.frame(Tender.=TENDERNESS, Predicted=fitted(TEND.LM),
+ Residuals=resid(TEND.LM), RIstudent=rstandard(TEND.LM),
+ REstudent=rstudent(TEND.LM))
```

	Tender.	Predicted	Residuals	RIstudent	REstudent
1	6.56	6.128296	0.43170385	1.36230868	1.38726955
2	6.80	7.121661	-0.32166052	-1.30326969	-1.32265959
3	4.70	4.971050	-0.27105042	-0.68139636	-0.67391669
4	5.31	5.506986	-0.19698606	-0.48423168	-0.47668897
5	5.75	4.984138	0.76586155	2.00130156	2.13974671
6	6.72	7.007448	-0.28744788	-0.80911589	-0.80335686
7	6.00	6.186869	-0.18686856	-0.45498913	-0.44765376
8	4.90	4.597686	0.30231394	0.78729301	0.78113061
9	6.18	5.883602	0.29639842	0.70062775	0.69331260
10	6.92	6.718424	0.20157627	0.49181816	0.48422966
11	6.92	6.803969	0.11603116	0.30250069	0.29693286
12	5.93	6.609211	-0.67921075	-1.61922520	-1.67687688
13	6.38	6.474147	-0.09414654	-0.21384015	-0.20971158
14	6.64	6.176600	0.46340008	1.16463863	1.17338317
15	6.08	5.881178	0.19882243	0.53167874	0.52390706

Remember that Residuals are equal to Observed minus Predicted values.

Multiple linear regression – residuals and influence (cont) -

(continues from the previous slide)

	Tender.	Predicted	Residuals	RIstudent	REstudent
16	6.27	6.525356	-0.25535577	-0.60854716	-0.60071788
17	6.98	6.716444	0.26355607	0.59066083	0.58280795
18	6.96	6.398485	0.56151522	1.44373999	1.47750463
19	6.85	6.860609	-0.01060922	-0.02469703	-0.02419835
20	6.54	6.265179	0.27482092	0.65351301	0.64584971
21	7.03	6.703682	0.32631809	0.75574628	0.74908336
22	6.66	6.452923	0.20707660	0.51319759	0.50549862
23	3.99	5.088394	-1.09839417	-2.74627603	-3.21997842
24	6.33	6.863100	-0.53310029	-1.28837693	-1.30646369
25	4.97	5.474445	-0.50444474	-1.54714249	-1.59411889
26	6.08	6.022310	0.05769010	0.13353734	0.13088603
27	7.23	6.874301	0.35569941	0.81014858	0.80440983
28	6.29	6.093660	0.19634009	0.47989518	0.47238014
29	6.18	6.391316	-0.21131576	-0.48497646	-0.47742910
30	6.42	7.117991	-0.69799051	-1.72789599	-1.80413916
31	6.09	6.380125	-0.29012536	-0.89980553	-0.89625832
32	5.95	6.174438	-0.22443759	-0.53476399	-0.52698229
33	7.08	6.834025	0.24597534	0.67705812	0.66954567
34	6.67	6.397946	0.27205438	0.64622696	0.63852607
35	8.01	7.684010	0.32599021	0.81473550	0.80908811

Rstudent values higher than 2 could suggest a weak outlier, and higher than 3 an outlier. This could be the condition for observations 5 and 23.

Multiple linear regression – Influence 1-

```
> influence.measures(TEND.LM)
```

	dffit	cov.r	cook.d	hat	inf
1	1.5289	1.5402	2.25e-01	0.548	*
2	-2.1534	2.7171	4.50e-01	0.726	*
3	-0.4291	1.7526	1.88e-02	0.288	
4	-0.2795	1.8395	8.06e-03	0.256	
5	1.5409	0.3985	2.08e-01	0.341	*
6	-0.7013	2.0326	4.99e-02	0.432	
7	-0.2526	1.8248	6.59e-03	0.241	
8	0.5569	1.7649	3.15e-02	0.337	
9	0.3415	1.5328	1.19e-02	0.195	
10	0.2756	1.8068	7.83e-03	0.245	
11	0.2124	2.1917	4.68e-03	0.338	
12	-0.8614	0.6278	6.92e-02	0.209	
13	-0.0805	1.6944	6.74e-04	0.128	
14	0.7464	1.2095	5.49e-02	0.288	
15	0.4025	2.1349	1.67e-02	0.371	
16	-0.3081	1.6364	9.74e-03	0.208	
17	0.1993	1.4598	4.08e-03	0.105	
18	1.0131	0.9259	9.80e-02	0.320	
19	-0.0110	1.8122	1.25e-05	0.170	
20	0.3278	1.5921	1.10e-02	0.205	
21	0.3289	1.4240	1.10e-02	0.162	
22	0.3058	1.8480	9.64e-03	0.268	
23	-2.0114	0.0577	2.94e-01	0.281	*
24	-0.7143	0.9829	4.96e-02	0.230	
25	-1.6658	1.1501	2.61e-01	0.522	

Values inside a green rectangle are bigger than their respective critical values: 0.57 (= 2*10/35) for leverage, 0.114 (= 4/35) for Cook's D, and 1.069 (= $2*\sqrt{(10/35)}$) for DFFITS.

Five observations (1, 2, 5, 23, 25) have two or more values bigger than the critical values for Cook's D, leverage and DFFITS, and thus are potentially influential.

Multiple linear regression – Influence 2 -

Obs	dfb.1_	dfb.ADG	dfb.LMAR	dfb.CONF	dfb.FAT	dfb.INTR	dfb.WHC	dfb.DRYM	dfb.COOK	dfb.TEXT
1	0.21370	-0.29938	1.260121	-0.25096	0.00883	-0.01270	-0.40919	-0.101710	-0.24109	-2.18e-01
2	-0.53031	0.13806	0.147696	1.24744	0.28990	-0.36256	0.32347	0.136544	0.26077	-2.08e-01
5	0.07995	0.38585	-0.001381	0.31323	0.00865	0.03917	-0.25881	-0.241745	-0.19858	1.04e+00
23	0.10130	0.66973	-0.987220	-1.22423	0.20285	-0.02432	1.19739	-0.333793	0.85039	-1.16e+00
25	0.71288	-0.64668	-0.398246	0.61047	0.68517	0.71097	-0.24452	-0.507008	-0.61434	-3.37e-01

This table presents DFBETAS only for those observations with two or more values indicating potential influence (see previous slide).

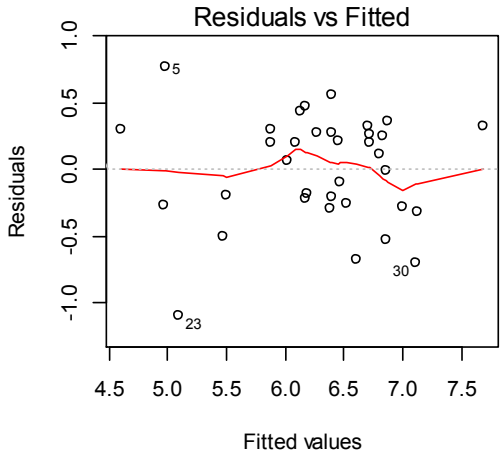
There are some DFBETAS with an absolute value greater than the critical value 0.3381 ($=2/\sqrt{35}$) suggested before, but none of the values reaches a more general critical value of 2 (Myers). This is consistent with what can be observed in the regression plots.

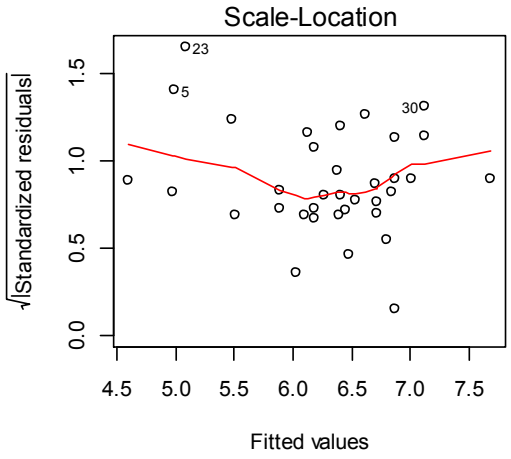
No remedial measures are needed to overcome the effects of potential influential observations in this analysis.

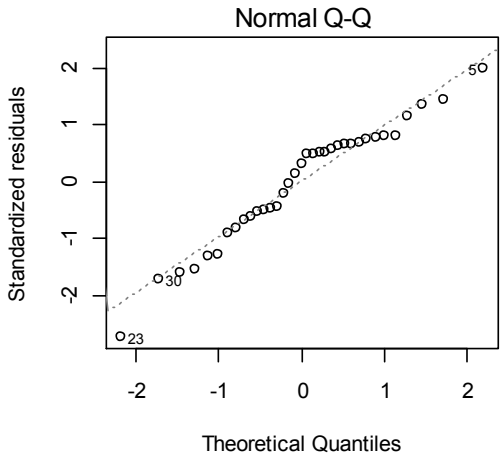
When some observations are clearly influential, the researcher must analyze carefully and decide what to do with these observations.

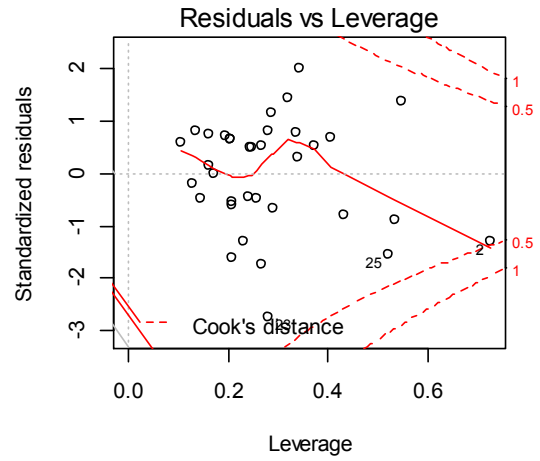
Multiple linear regression – Graphic diagnostics -

```
> layout(matrix(c(1,2,3,4),2,2))
> plot(TEND.LM)
```









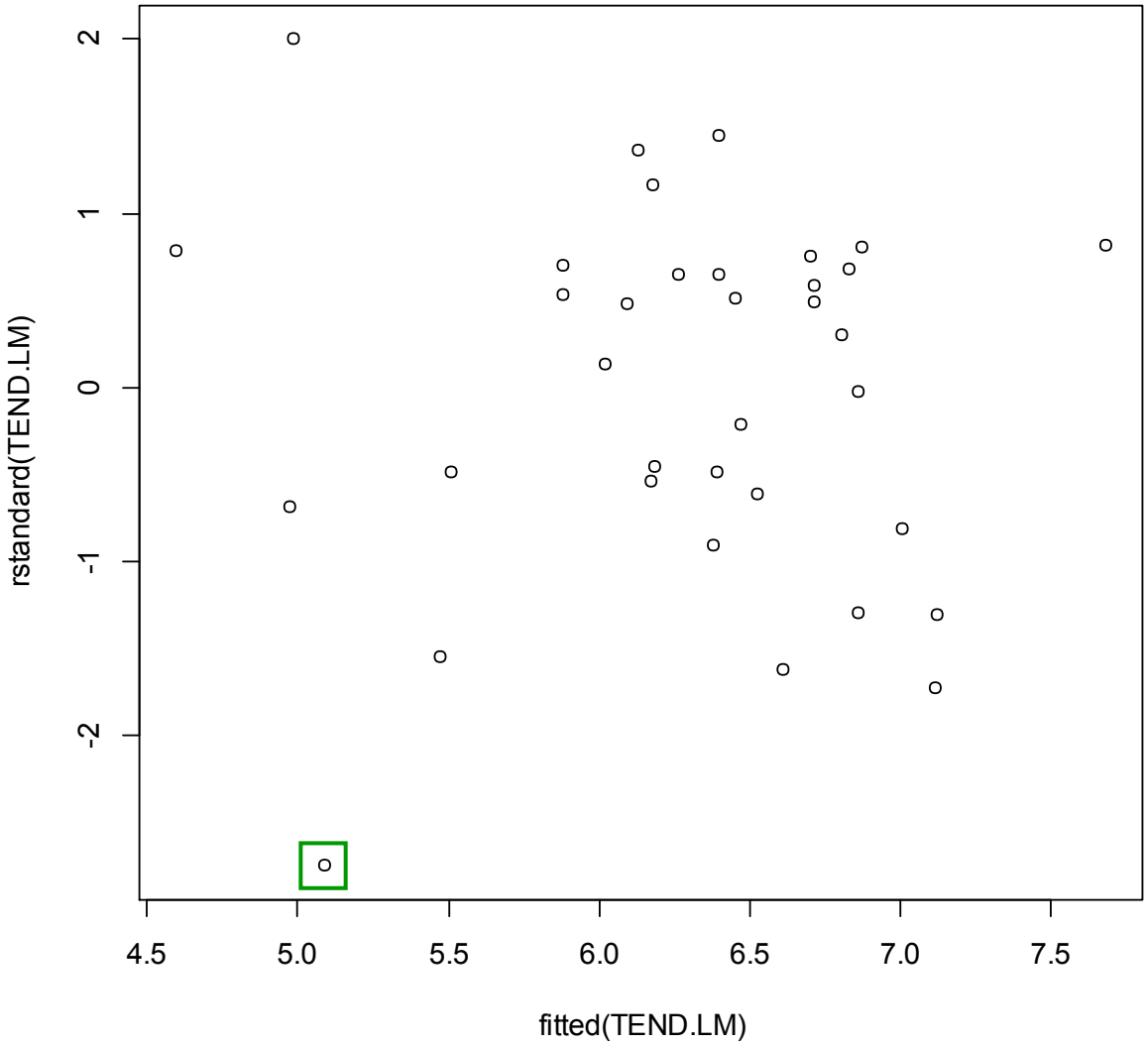
The distribution of residuals is about random (homogeneity of variance).

The Q-Q plot does not deviate greatly from normality.

Only one observation approaches Cook's contours, suggesting that we must analyze if it is potentially influential.

Multiple linear regression – another graphic diagnostic -

```
> plot(fitted(TEND.LM), rstandard(TEND.LM))
```



Observe that one of the observations has a |studentized residual|>2, but lower than 3: weak outlier.

Multicollinearity

A very desirable condition in a set of regression data is to have regressors (covariates) that are not "moving with each other" (redundant) in the data set, i.e., not correlated. Near linear dependencies render it more difficult to sort out the impact of each regressor on the response.

The variances of the coefficients are inflated due to collinearity. The variance inflation factor (VIF) is the factor of multiplication of that variance and could led to erroneous *p*-values. For the *i*-th regression coefficient, VIF can be written as

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}; \qquad \left(\mathrm{VIF} = \frac{1}{Tol}; \quad \mathrm{Tol} = \frac{1}{VIF} \right)$$

where Tol stands for Tolerance, and R_i^2 is the coefficient of multiple determination of the regression produced by regressing the variable x_i against the other regression variables:

$$x_{ii} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{(i-1)} x_{(i-1)i} + \beta_{(i+1)} x_{(i+1)i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

Another effect of multicollinearity is the instability of regression coefficients, i.e., coefficients that are very much dependent of the data set.

Multiple correlation

```
> cor(TEND.AM[,1:10])
```

	ADG	LMAREA	CONF	FAT	INTRAMFAT
ADG	1.00000000	0.11308494	0.23086676	0.233538931	0.14648715
LMAREA	0.11308494	1.00000000	0.34498818	0.039055283	-0.08071684
CONF	0.23086676	0.34498818	1.00000000	-0.055828172	0.04923206
FAT	0.23353893	0.03905528	-0.05582817	1.000000000	-0.35340524
INTRAMFAT	0.14648715	-0.08071684	0.04923206	-0.353405242	1.00000000
WHC	0.20277949	0.30990410	0.42039164	-0.086593004	-0.04564285
DRYMAT	0.05713545	-0.18620606	0.02066455	-0.220360438	0.63879943
COOKLOSS	-0.25960403	0.03042370	0.19294086	-0.392611852	0.10446587
TEXTWB	0.06894135	-0.07750252	-0.13892865	-0.004208519	-0.29135912
TENDERNESS	0.04072221	-0.27374126	-0.20522130	-0.150228271	0.11045691

	WHC	DRYMAT	COOKLOSS	TEXTWB	TENDERNESS
ADG	0.202779490	0.05713545	-0.2596040	0.068941354	0.040722211
LMAREA	0.309904098	-0.18620606	0.0304237	-0.077502521	-0.273741264
CONF	0.420391640	0.02066455	0.1929409	-0.138928648	-0.205221304
FAT	-0.086593004	-0.22036044	-0.3926119	-0.004208519	-0.150228271
INTRAMFAT	-0.045642850	0.63879943	0.1044659	-0.291359124	0.110456912
WHC	1.000000000	0.13346766	-0.2514503	-0.039393442	-0.007659911
DRYMAT	0.133467663	1.00000000	0.1148036	-0.316161438	0.089036413
COOKLOSS	-0.251450275	0.11480358	1.0000000	0.156279740	-0.368306227
TEXTWB	-0.039393442	-0.31616144	0.1562797	1.000000000	-0.625808183
TENDERNESS	-0.007659911	0.08903641	-0.3683062	-0.625808183	1.000000000

Note that the correlation matrix is symmetrical, with duplicated information.

Given the sample size (35) only correlations bigger than 0.32 are significant (7 in total). Probably we will not find multicollinearity.

Multiple linear regression – VIF and Tol -

```
> VIF<-vif(TEND.LM)
> TOL<-1/vif(TEND.LM)
> cbind(VIF=VIF, Tol=TOL)
```

	VIF	Tol
ADG	1.387940	0.7204921
LMAREA	1.320040	0.7575525
CONF	1.697632	0.5890558
FAT	1.594628	0.6271053
INTRAMFAT	2.223073	0.4498277
WHC	1.912595	0.5228499
DRYMAT	2.173664	0.4600527
COOKLOSS	1.871217	0.5344115
TEXTWB	1.375392	0.7270654

All values of Variance Inflation are below 10 ⇒ suggest **absence of multicollinearity**.

VIF and Tol define the same. Only one of them (VIF) would be needed.

Diagnosis of multicollinearity

Simple correlations among regressor variables
The simple correlations not always underscore the extent of the problem.

Variance Inflation Factor (VIF)
It involves the notion of multiple association. If R_i^2 is near unity, VIF_i will be quite large. This will occur if the regressor variable has a linear strong association with the other variables. Myers says that any VIF exceeds 10, there is reason of some concern.

Eigenvalues of X'X
The eigenvalues would all be 1 if the variables define an orthogonal system (variables are independent). Very small eigenvalues indicate multicollinearity, but there is not rule of thumb.

Condition index
Square root of the quotient between the first and the *i*th eigenvalues. Values exceeding 30 indicate multicollinearity.

$$\text{Cond. Index} = \sqrt{\lambda_1 / \lambda_i}$$

Variance proportions
The appearance of a small eigenvalue implies that any or all regression coefficients may be adversely affected. It is of interest to determine what proportion of the variance of each coefficient is attributed to each dependency. A small eigenvalue (or a high condition index), accompanied by a subset of regressors (at least 2) with high variance proportions (greater than 0.5), represents a dependency involving the regressors in that subset, damaging the precision of estimation of the coefficients in the subset.

Criteria to choose the best model

The researcher faces the question of what terms to include in the regression model, as not all are significant. This can be complicated by the existence of multicollinearity and scientist's prior views regarding the importance of individual variables.

The successful model builder must understand that several models can be fit that would be nearly equal in effectiveness: **Prior** to the analysis, the question should be: **What will be done with the model**?

1. **Learn something about the system** from which data were taken: a slope, a sign, optimum operating conditions, …
2. **Learn which regressors are important** and which are not: conduct a variable selection. Usually a prelude to a more elaborate search for a model. 1 and 2 are related.
3. **Prediction**: selection of one model that best predicts from a pool of candidate models. Often a difficult task, not covered in this course.

Standard criteria for comparing models

1. Coefficient of Determination, R^2. It is a measure of the model's capability to fit the present data. The insertion of a new regressor into a model can not bring about a decrease in R^2. Not conceptually prediction oriented.
2. Estimate of Error Variance, often called Residual Mean Square (RMS). A reasonable plan is to choose the candidate model with the smallest value of RMS.
3. Adjusted R^2. This is a R^2-like statistic that guards against the practice of overfitting. This statistic punishes the user who includes marginally important model terms at the expenses of error degrees of freedom.
4. PRESS (Prediction Sum of Squares) statistic. It is computed from the PRESS residuals. For choice of the best model one might favour the model with the smallest PRESS.

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

5. Conceptually predictive criteria (C_p) Statistic. Prediction oriented criterion. A reasonable choice is the model with the smallest value and $C_p = p$, the number of parameters. This can be done through a plot of p against C_p.

Sequential variable selection procedures

Sequential algorithms are an important heritage of least squares regression and are used by a large number of analysts (probably because they are implemented in many statistical packages). Myers, however, says that coupled with the reality that multicollinearity often clouds the picture, sequential algorithms might not be of practical use except in rare cases. There are three general types:

1. Forward selection.
2. Backward elimination.
3. Stepwise regression.

In addition we will consider Mallows *Cp*.

FORWARD selection

The initial model contains only a constant term. The procedure selects for entry the variable that produces the largest R^2 of any single regressor. Lets call this regressor x_1. The second regressor (x_2) is chosen which produces the largest increase in R^2 in the presence of x_1. This is equivalent to choosing the regressor with the largest partial F. Thus, at stage 2 we have

$$F = \frac{MSReg(x_2 \mid x_1)}{MSError(x_2, x_1)}$$

The above process continues until the candidate regressor for entry does not exceed a preselected F or its significance level.

BACKWARD elimination

This procedure begins with all regressors in the model and eliminates one at a time. The criterion is to eliminate the regressor with the highest p-value. The procedure is continued until the candidate regressor for removal experiences a partial t value which exceeds the preselected level.

It is recommended not to impose a too strict significance level, for example 0.10, to maintain a variable in the model.

It is basically a manual process from the console. Consist of working in the console and removing from the full model the variable that has a larger p-value for the t-statistic. This process is done repeatedly until all variables have a p-value higher than the preselected level.

BACKWARD elimination – results 1 step 0 -

```
> summary(TEND.LM)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.08589    4.04187   3.980 0.000522 ***
ADG          2.41867    1.20541   2.007 0.055732 .
LMAREA      -0.03725    0.01400  -2.660 0.013451 *
CONF        -0.15123    0.08507  -1.778 0.087613 .
FAT         -0.20507    0.06618  -3.099 0.004756 **
INTRAMFAT   -0.11912    0.11372  -1.047 0.304925
WHC          0.01089    0.04950   0.220 0.827721
DRYMAT      -0.20156    0.16608  -1.214 0.236226
COOKLOSS    -0.06310    0.03644  -1.732 0.095650 .
TEXTWB      -0.47059    0.07360  -6.394 1.08e-06 ***

Residual standard error: 0.4716 on 25 degrees of freedom
Multiple R-squared: 0.7388,     Adjusted R-squared: 0.6447
F-statistic: 7.855 on 9 and 25 DF,  p-value: 2.065e-05
```

Full model

From here, the process is fully manual.

BACKWARD elimination – results 2, step 1 -

```
Call:
lm(formula = TENDERNESS ~ ADG + LMAREA + CONF + FAT +
INTRAMFAT + DRYMAT + COOKLOSS + TEXTWB)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.95165    3.92171   4.068 0.000392 ***
ADG          2.42825    1.18237   2.054 0.050189 .
LMAREA      -0.03634    0.01313  -2.767 0.010278 *
CONF        -0.14260    0.07409  -1.925 0.065281 .
FAT         -0.20975    0.06151  -3.410 0.002130 **
INTRAMFAT   -0.12669    0.10638  -1.191 0.244448
DRYMAT      -0.18733    0.15014  -1.248 0.223240
COOKLOSS    -0.06705    0.03114  -2.153 0.040739 *
TEXTWB      -0.46775    0.07112  -6.577 5.65e-07 ***

Residual standard error: 0.4629 on 26 degrees of freedom
Multiple R-squared: 0.7383,     Adjusted R-squared: 0.6577
F-statistic: 9.167 on 8 and 26 DF,  p-value: 6.583e-06
```

WHC removed

BACKWARD elimination – results 3, step2 -

```
Call:
lm(formula = TENDERNESS ~ ADG + LMAREA + CONF + FAT +
DRYMAT + COOKLOSS + TEXTWB)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.13804    3.49218   5.194 1.81e-05 ***
ADG          2.06800    1.15184   1.795  0.08379 .
LMAREA      -0.03652    0.01323  -2.759  0.01027 *
CONF        -0.13661    0.07449  -1.834  0.07771 .
FAT         -0.18500    0.05834  -3.171  0.00376 **
DRYMAT      -0.28372    0.12743  -2.226  0.03452 *
COOKLOSS    -0.06808    0.03136  -2.171  0.03892 *
TEXTWB      -0.45160    0.07035  -6.419 7.07e-07 ***

Residual standard error: 0.4664 on 27 degrees of freedom
Multiple R-squared: 0.724,      Adjusted R-squared: 0.6524
F-statistic: 10.12 on 7 and 27 DF,  p-value: 3.683e-06
```

INTRAMFAT removed

No additional regressors need to be removed with a t-$_{\text{stay}}$ of 0.10.

STEPWISE selection

This is a combination of the two previous types.

In R, the process starts from the full model and is a Backward elimination as described before, but assessing the possibility of re-introducing a variable that was removed before.

Thus, at each stage a regressor can be entered, and another can be eliminated. This is because multicollinearity can render a regressor of little value even though it was an important candidate at an early stage of the procedure.

The procedure ends when no additional regressors can be eliminated on the basis of t_{stay} and no additional regressors must be introduced in the model. Typical values for entry and stay are 0.15.

STEPWISE selection – results 1 -

```
> step(TEND.LM)

Start:  AIC=-44.39
TENDERNESS ~ ADG + LMAREA + CONF + FAT + INTRAMFAT +
WHC + DRYMAT + COOKLOSS + TEXTWB

             Df Sum of Sq     RSS     AIC
- WHC         1    0.0108  5.5704 -46.326
- INTRAMFAT   1    0.2440  5.8037 -44.890
<none>                     5.5597 -44.393
- DRYMAT      1    0.3276  5.8872 -44.390
- COOKLOSS    1    0.6669  6.2266 -42.428
- CONF        1    0.7028  6.2625 -42.227
- ADG         1    0.8954  6.4550 -41.167
- LMAREA      1    1.5732  7.1329 -37.672
- FAT         1    2.1355  7.6952 -35.016
- TEXTWB      1    9.0918 14.6515 -12.478
```

Full model

STEPWISE selection – results 2 -

```
Step:  AIC=-46.33
TENDERNESS ~ ADG + LMAREA + CONF + FAT + INTRAMFAT +
DRYMAT + COOKLOSS + TEXTWB

             Df Sum of Sq     RSS     AIC
- INTRAMFAT   1    0.3039  5.8743 -46.467
<none>                     5.5704 -46.326
- DRYMAT      1    0.3336  5.9040 -46.290
- CONF        1    0.7936  6.3641 -43.664
- ADG         1    0.9036  6.4741 -43.064
- COOKLOSS    1    0.9935  6.5639 -42.582
- LMAREA      1    1.6404  7.2109 -39.292
- FAT         1    2.4914  8.0618 -35.387
- TEXTWB      1    9.2679 14.8383 -14.035
```

WHC eliminated

STEPWISE selection – results 3 -

```
Step:  AIC=-46.47
TENDERNESS ~ ADG + LMAREA + CONF + FAT + DRYMAT +
COOKLOSS + TEXTWB

           Df Sum of Sq     RSS     AIC
<none>                   5.8743 -46.467
- ADG       1    0.7013  6.5756 -44.519
- CONF      1    0.7317  6.6060 -44.358
- COOKLOSS  1    1.0251  6.8994 -42.837
- DRYMAT    1    1.0785  6.9528 -42.567
- LMAREA    1    1.6565  7.5308 -39.772
- FAT       1    2.1879  8.0622 -37.386
- TEXTWB    1    8.9649 14.8392 -16.033
```

INTRAMFAT removed

STEPWISE selection – results 4 -

```
Call:
lm(formula = TENDERNESS ~ ADG + LMAREA + CONF + FAT +
DRYMAT + COOKLOSS + TEXTWB)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.13804    3.49218   5.194 1.81e-05 ***
ADG          2.06800    1.15184   1.795  0.08379 .
LMAREA      -0.03652    0.01323  -2.759  0.01027 *
CONF        -0.13661    0.07449  -1.834  0.07771 .
FAT         -0.18500    0.05834  -3.171  0.00376 **
DRYMAT      -0.28372    0.12743  -2.226  0.03452 *
COOKLOSS    -0.06808    0.03136  -2.171  0.03892 *
TEXTWB      -0.45160    0.07035  -6.419 7.07e-07 ***

Residual standard error: 0.4664 on 27 degrees of freedom
Multiple R-squared: 0.724,      Adjusted R-squared: 0.6524
F-statistic: 10.12 on 7 and 27 DF,  p-value: 3.683e-06
```

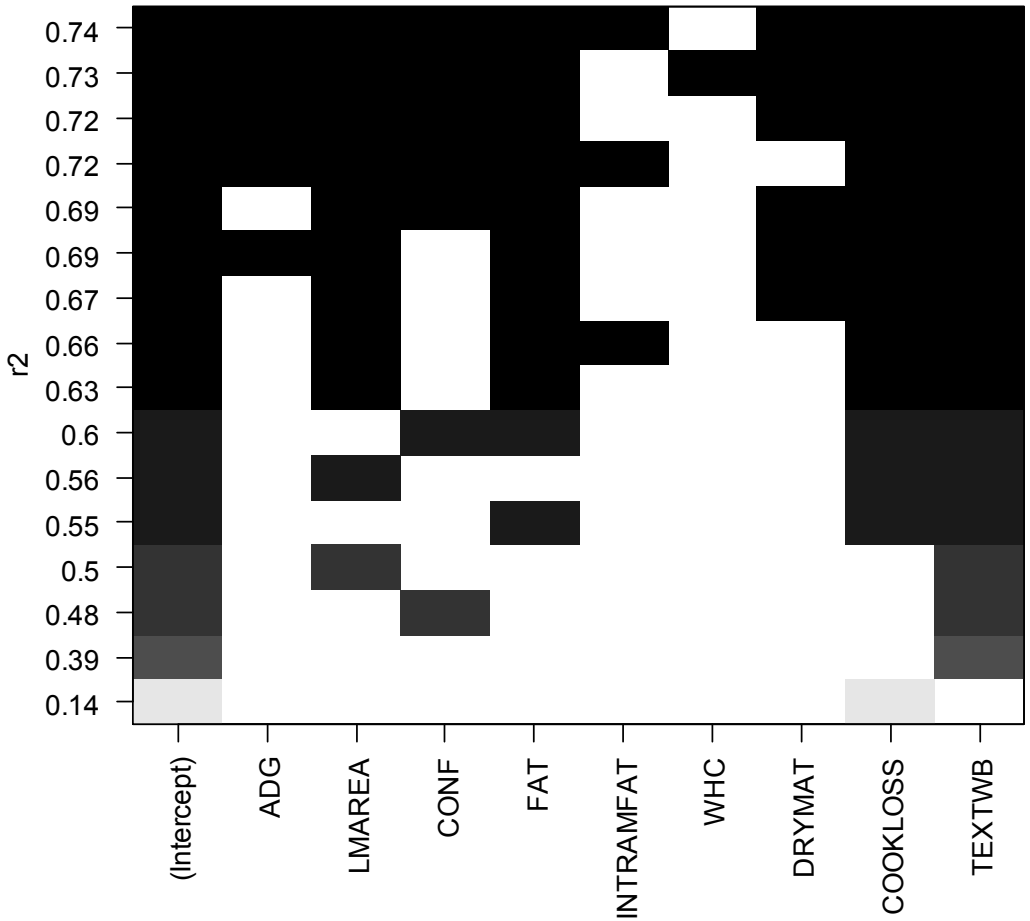
Those are the variables retained for the final model and their coefficients.

Some comments (from Myers)

1. The data analyst should view the sequential model-building algorithms not as a black box, which produces one final model, but rather as an exercise that allows the user to see several models perform: **exploratory procedure**.
2. The sequential procedures can be partially ineffective with data sets involving collinearity among a large number of regressors.
3. Bear in mind that there is no assurance that any specific sequential model-building strategy will result in the best variable subset. **The final result is very much dependent on the sequential method chosen and the t_{entry} and t_{stay} values**.
4. There are procedures that rather than producing a final single model, several models are suggested with final choice left to the analyst: maximum R^2 procedure with a *Cp* statistic.

Graphical summary of different models

```
library(leaps)
> TEND.LEAPS<-regsubsets(TENDERNESS~., data=TEND.AM, nbest=2)
> # plot a table of models showing variables in each model
> plot(TEND.LEAPS, scale="r2")
```



This figure shows the dependent variables that most appear in the different models. In this case: TEXTWB, COOKLOSS, FAT, LMAREA.

R-Square and Mallows *Cp*

The C_p statistic is a measure of total squared error for a subset of models containing k independent variables.

The total squared error is a measure of the error variance plus the bias introduced by not including important variables in the model.

$$C_p = (\text{SSE}(k)/\text{MSE}) - (n - 2k) + 1$$

MSE, error mean square for the full model.

SSE(k), error sum of squares for the subset model containing k independent variables (intercept not included).

n, total sample size.

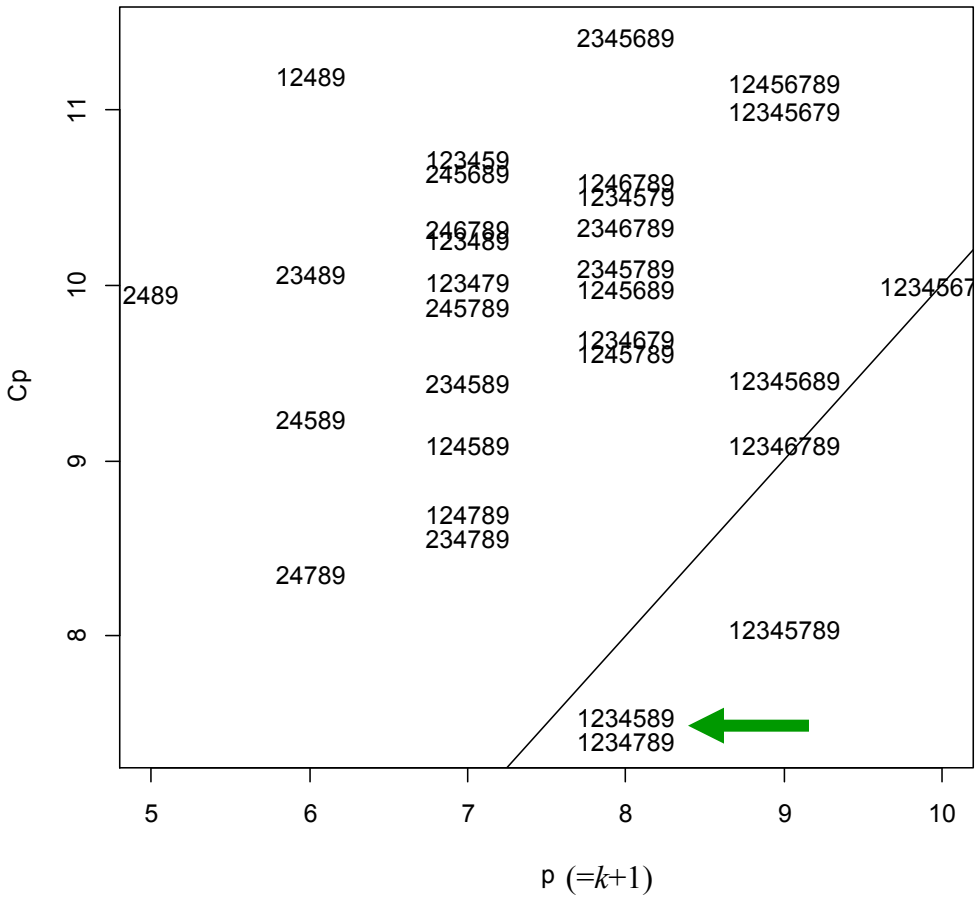
$C_p > (k+1) \Rightarrow$ bias due to an incompletely specified model.

$C_p < (k+1) \Rightarrow$ model overspecified, with too many variables.

Mallows recommends that C_p be plotted against k, and further recommends selecting that subset size where the minimum C_p first approaches $k+1$, starting from the full model: i.e., models with small $k+1$ and C_p around or less than $k+1$.

Cp plot

```
> x <- model.matrix(TEND.LM)[,-1]
> y <- TEND.AM$TENDERNESS
> g <- leaps(x,y)
> library(faraway); Cpplot(g)
```



The numbers indicate the independent variables in the order of the original (full) model.

This plot clearly favours models with a high number of regressors; 7 would be the best option.

In this case, 1234589 and 1234789 would be very similar from a statistical point of view (see next slide).

Note that the model 1234789 correspond to the one selected by the Stepwise procedure.

Getting R^2 of selected models

With the libraries **leaps** and **faraway** activated:

```
> adjr <- leaps(x,y, method="adjr2")
> maxadjr(adjr,4)

  1,2,3,4,5,7,8,9     1,2,3,4,7,8,9     1,2,3,4,5,8,9
            0.658             0.652             0.651

1,2,3,4,5,6,7,8,9
            0.645
```

The two models of 7 variables selected have adjusted R^2 of 0.652 and 0.651, respectively. The researcher must decide the best model depending upon statistical criteria and also the objective of the analysis: a) study of the relationships among variables or b) prediction.

Final comment

Statistics is rarely a substitute for sound scientific knowledge and reasoning

We cannot ignore input from experts in the scientific discipline involved. Statistical procedures are vehicles that lead us to conclusions, but scientific logic paves the road along the way.

However, a good scientist must remember that to arrive to an adequate prediction equation, balance must be achieved that takes into account what the data can support.

There are times when inadequacies in the data and random noise may not allow the true structure to come through.

For these reasons, a proper marriage must exist between the experienced statistician and the learned expert in the discipline involved.

(Myers, 1990)

References

Fry J.C. 1993. *Biological Data Analysis*. IRL Press, Oxford.

Myers R.H. 1990. *Classical and Modern Regression with Applications*. Duxbury Press, Belmont, CA.