Chapter 6

Variable Screening Methods

Based on slides from Linear Regression Analysis 5E
Montgomery, Peck & Vining

Model-Building Problem

Two "conflicting" goals in regression model building:

1. Want as many regressors as possible so that the "information content" in the variables will influence $\hat{y}$
2. Want as few regressors as necessary because the variance of $\hat{y}$ will increase as the number of regressors increases. (Also, more regressors can cost more money in data collection/model maintenance)

Principle of parsimony

A compromise between the two hopefully leads to the *best* regression equation.

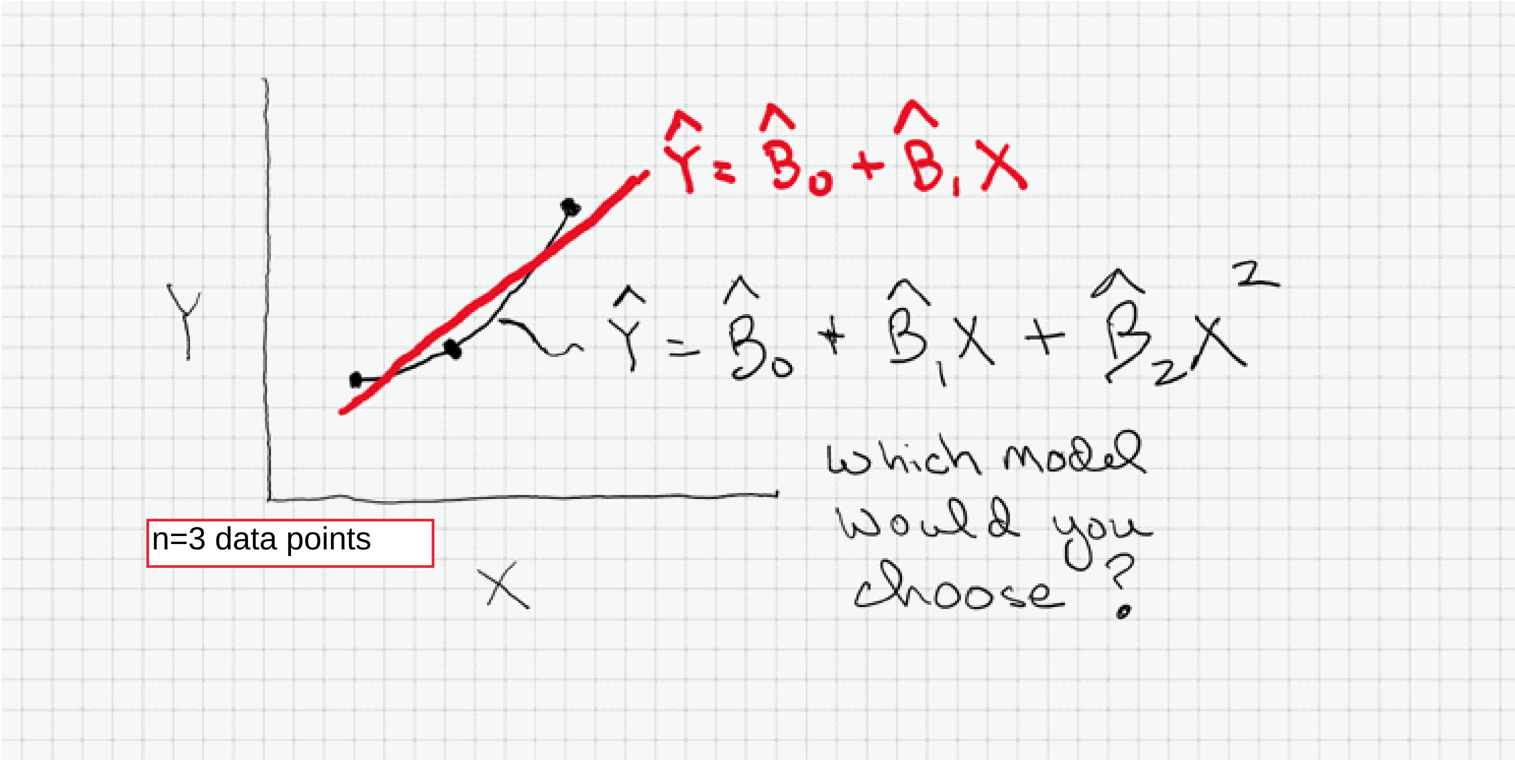
All models are wrong, but some are useful. George Box (1919 - 2013)

We will cover some variable selection techniques. Keep in mind the following:

1. None of the variable selection techniques can guarantee the best regression equation for the dataset of interest.
2. The techniques may very well give different results.
3. Complete reliance on the algorithm for results is to be avoided. Other valuable information such as experience with and knowledge of the data and problem.

Consequences of Model Misspecification

- Deleting variables improves the precision of the parameter estimates of retained variables.
- Deleting variables improves the precision of the variance of the predicted response.
- Deleting variables can induce bias into the estimates of coefficients and variance of predicted response. (But, if the deleted variables are "insignificant" the MSE of the biased estimates will be less than the variance of the unbiased estimates).
- Retaining insignificant variables can increase the variance of the parameter estimates and variance of the predicted response.



The straight-line model follows the principle of parsimony. The quadratic model fits the data perfectly, but could be fitting noise - could be “overfitting”.

Two approaches to model building to be discussed:

1. All possible regressions
2. Stepwise regression

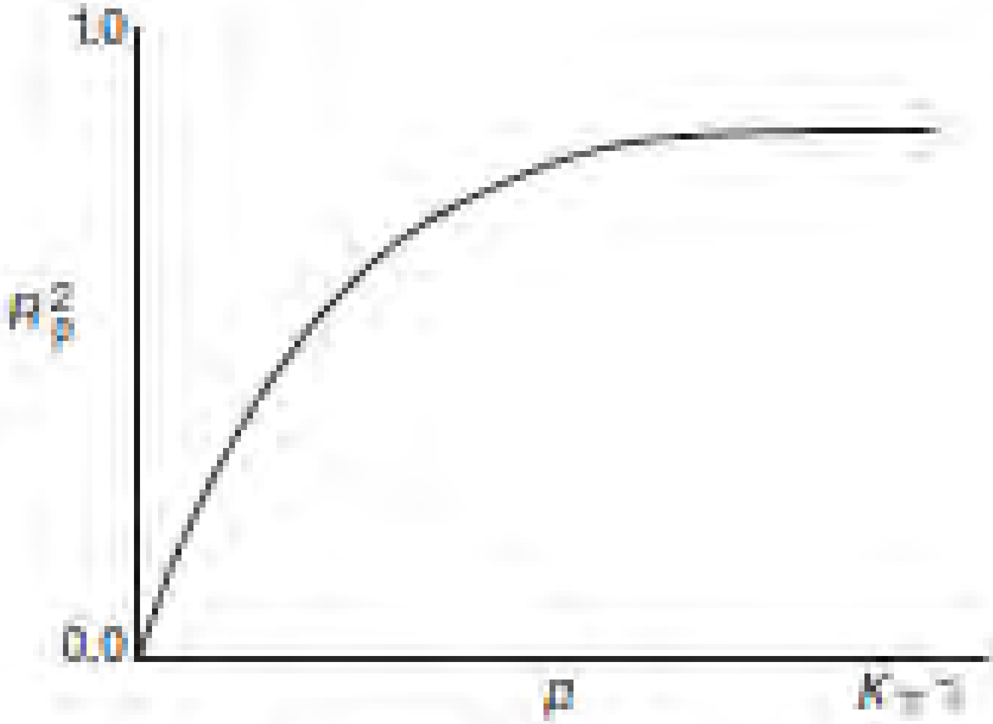
Criteria for Evaluating Subset Regression Models

Coefficient of Multiple Determination

- Say we are investigating a model with *p* terms,

$$R_p^2 = \frac{SS_R(p)}{SS_T} = 1 - \frac{SS_{\mathrm{Re}s}(p)}{SS_T}$$

- Models with large values of R_p^2 are preferred, but adding terms will increase this value.



Plot of R_p^2 versus p.

Adjusted R^2

- Say we are investigating a model with *p* terms,

$$R^2_{adj,p} = 1 - \left(\frac{n-1}{n-p}\right)(1 - R_p^2)$$

- This value will not necessarily increase as additional terms are introduced into the model. We want a model with the maximum adjusted R^2.

Residual Mean Square

- The MS_{res} for a subset regression model is

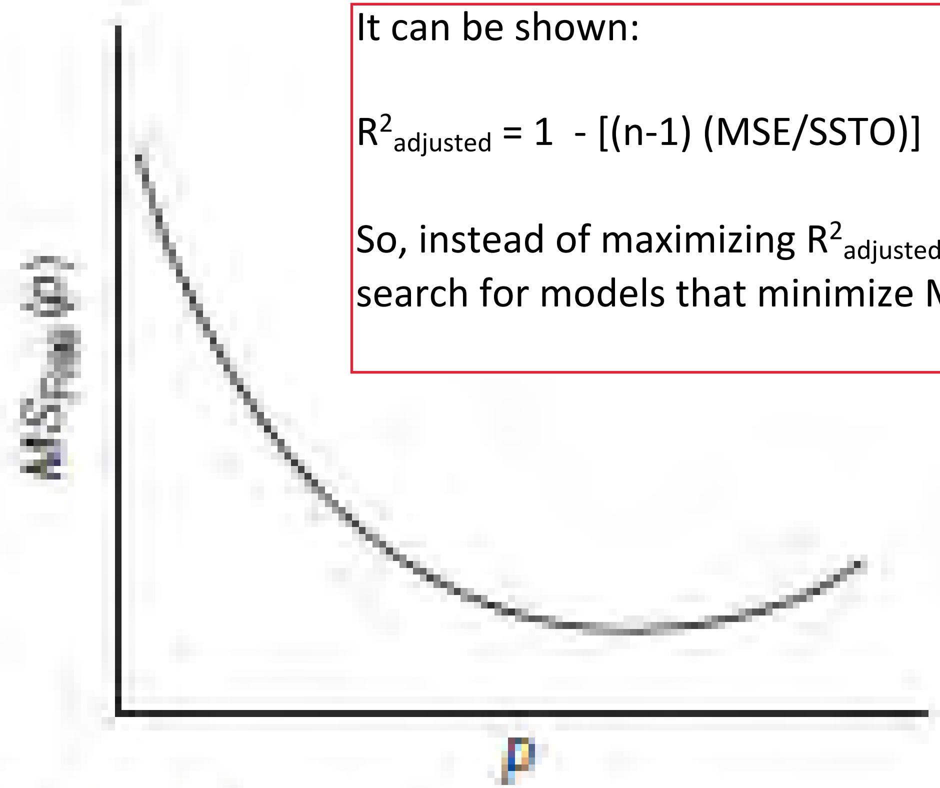
$$MS_{\mathrm{Res}}(p) = \frac{SS_{\mathrm{Res}}(p)}{n-p}$$

- $MS_{\mathrm{Res}}(p)$ $\boxed{\text{decreases}}$ ~~increases~~ as p increases, in general. The increase in $MS_{\mathrm{Res}}(p)$ occurs when the reduction in $SS_{\mathrm{Res}}(p)$ from adding a regressor to the model is not sufficient to compensate for the loss of <u>one</u> degree of freedom. We want a model with a minimum $MS_{\mathrm{Res}}(p)$.

It can be shown:

$R^2_{adjusted} = 1 - [(n-1)\ (MSE/SSTO)]$

So, instead of maximizing $R^2_{adjusted}$, equivalently search for models that minimize MSE.



Plot of $MS_{Res}(p)$ versus p.

E(yihat) = mean response for the fitted subset model, E(yi) is the mean response for the true model.

Mallow's C_p Statistic

- This criterion is related to the MSE of the fitted value, that is

$$E\left[\hat{y}_i - E(y_i)\right]^2 = \left[E(y_i) - E(\hat{y}_i)\right]^2 + Var(\hat{y}_i)$$

- where $\left[E(y_i) - E(\hat{y}_i)\right]^2$ is the *squared bias*. The total squared bias for a *p*-term model is

$$SS_B(p) = \sum_{i=1}^{n}\left[E(y_i) - E(\hat{y}_i)\right]^2$$

The objective is to compare the total mean square error of the subset regression model with sigma^2(the variance of the random error for the true model using the ratio (TMSE/sigma^2). Cp is a good estimator for the ratio (TMSE/sigma^2).

Mallow's C_p Statistic

author says Total Mean Squared Error (TMSE)

- The standardized total squared error is

$$\Gamma_p = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}\left[E(y_i)-E(\hat{y}_i)\right]^2+\sum_{i=1}^{n}Var(\hat{y}_i)\right)$$

B=bias

$$= \frac{SS_B(p)}{\sigma^2}+\frac{1}{\sigma^2}\sum_{i=1}^{n}Var(\hat{y}_i)$$

Recall

$$\sum_{i=1}^{n}Var(\hat{y}_i) = p\sigma^2$$

- Making some appropriate substitutions, we can find the estimate of Γ_p, denoted C_p:

See page 345 of text

$$C_p = \frac{SS_{\mathrm{Res}}(p)}{\hat{\sigma}^2}-n+2p$$

Mallow’s C_p Statistic

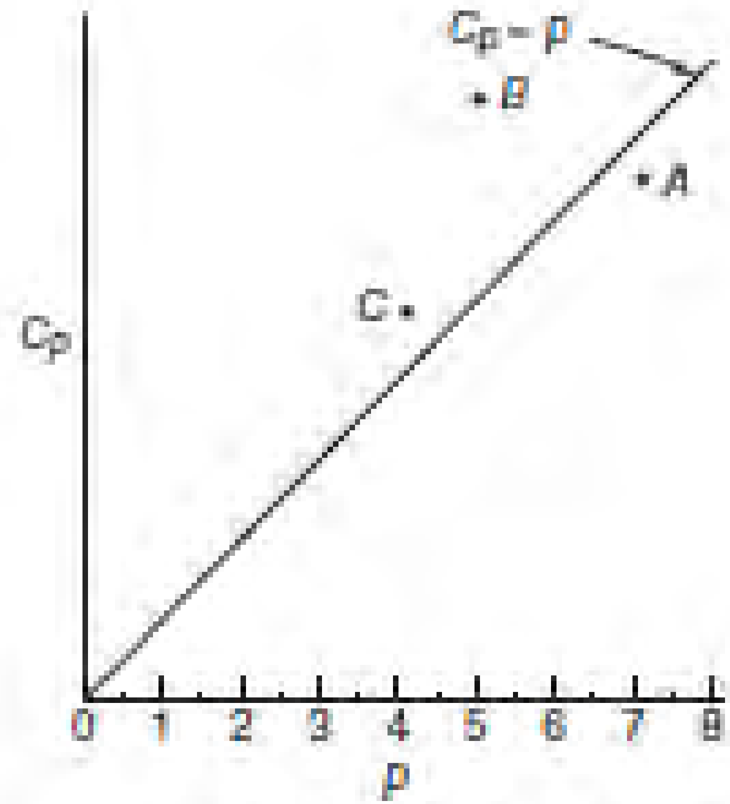
- It can be shown that if Bias = 0, the expected value of C_p is

$$E\left[C_p \mid Bias = 0\right] = \frac{(n-p)\hat{\sigma}^2}{\hat{\sigma}^2} - n + 2p = p$$

Mallow's C_p Statistic

Notes:

1. C_p is a measure of variance in the fitted values and $(\text{bias})^2$. (Large bias can be a result of important variables being left out of the model).
2. $C_p >> p$, then significant bias.
3. Small C_p values are desirable.
4. Beware of negative values of C_p. These could result because the MSE for the full model overestimates the true σ^2.



A C_p Plot

see AICBICSBC.pptx and Measuresoffit.pdf

The Akaike Information Criterion and Bayesian Analogues (BICs) Akaike proposed an information criterion, AIC, based on maximizing the expected *entropy* of the model. Entropy is simply a measure of the expected information, in this case the Kullback-Leibler information measure. Essentially, the AIC is a penalized log-likelihood measure. Let L be the likelihood function for a specific model. The AIC is

$$\mathrm{AIC} = -2\ln(L) + 2p,$$

where p is the number of parameters in the model. In the case of ordinary least squares regression,

$$\mathrm{AIC} = n\ln\left(\frac{SS_{\mathrm{Res}}}{n}\right) + 2p.$$

The key insight to the AIC is similar to R^2_{Adj} and Mallows C_p. As we add regressors to the model, SS_{Res}, cannot increase. The issue becomes whether the decrease in SS_{Res} justifies the inclusion of the extra terms.

There are several Bayesian extensions of the AIC. Schwartz (1978) and Sawa (1978) are two of the more popular ones. Both are called BIC for Bayesian information criterion. As a result, it is important to check the fine print on the statistical software that one uses! The Schwartz criterion ($\mathrm{BIC}_{\mathrm{Sch}}$) is

$$\mathrm{BIC}_{\mathrm{Sch}} = -2\ln(L) + p\ln(n).$$

This criterion places a greater penalty on adding regressors as the sample size increases. For ordinary least squares regression, this criterion is

$$\mathrm{BIC}_{\mathrm{Sch}} = n\ln\left(\frac{SS_{\mathrm{Res}}}{n}\right) + p\ln(n).$$

R uses this criterion as its BIC. SAS uses the Sawa criterion, which involves a more complicated penalty term. This penalty term involves σ^2 and σ^4, which SAS estimates by MS_{Res} from the full model.

The AIC and BIC criteria are gaining popularity. They are much more commonly used in the model selection procedures involving more complicated modeling situations than ordinary least squares, for example, the mixed model situation outlined in Section 5.6. These criteria are very commonly used with generalized linear models (Chapter 13).

Uses of Regression and Model Evaluation Criteria

- Regression equations may be used to make predictions. So, minimizing the MSE for prediction may be an important criterion. The PRESS statistic can be used for comparisons of candidate models.

$$PRESS_p = \sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2$$
$$= \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

- We want models with small values of PRESS.

Variable Selection/Screening Methods

All Possible Regressions

- Assume the intercept term is in all equations considered. Then, if there are K regressors, we would investigate 2^K possible regression equations. Use the criteria above to determine some candidate models and complete regression analysis on them.

Hald Cement Data

The response variable y is the heat evolved in a cement mix. The four explanatory variables are ingredients of the mix, i.e., x1: tricalcium aluminate, x2: tricalcium silicate, x3: tetracalcium alumino ferrite, x4: dicalcium silicate. An important feature of these data is that the variables x1 and x3 are highly correlated (corr(x1,x3)=-0.824), as well as the variables x2 and x4 (with corr(x2,x4)=-0.975). Thus we should expect any subset of (x1,x2,x3,x4) that includes one variable from highly correlated pair to perform as any subset that also includes the other member.

Hald Cement Data

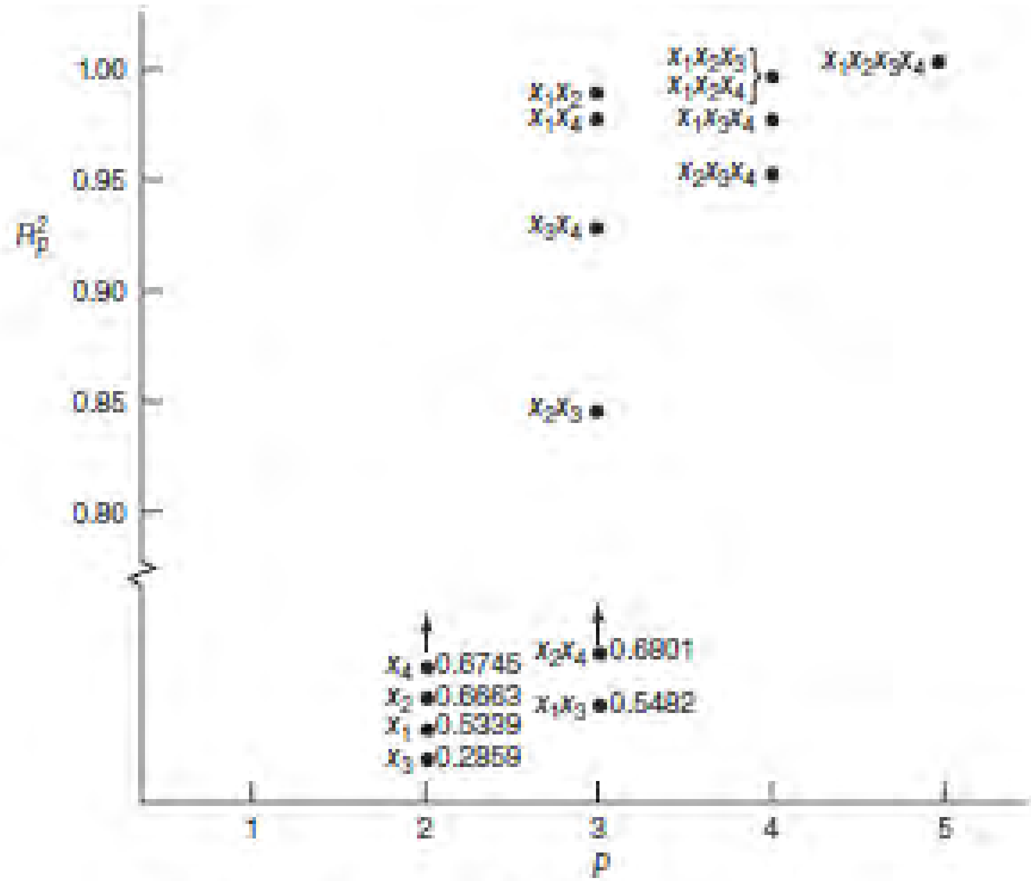
Observation i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Summary of All Possible Regressions for the Hald Cement Data

Number of Regressors in Model	p	Regressors in Model	$SS_{Res}(p)$	R_p^2	$R_{Adj,p}^2$	$MS_{Res}(p)$	C_p
None	1	None	2715.7635	0	0	226.3136	442.92
1	2	x_1	1265.6867	0.53395	0.49158	115.0624	202.55
1	2	x_2	906.3363	0.66627	0.63593	82.3942	142.49
1	2	x_3	1939.4005	0.28587	0.22095	176.3092	315.16
1	2	x_4	883.8669	0.67459	0.64495	80.3515	138.73
2	3	x_1x_2	57.9045	0.97868	0.97441	5.7904	2.68
2	3	x_1x_3	1227.0721	0.54817	0.45780	122.7073	198.10
2	3	x_1x_4	74.7621	0.97247	0.96697	7.4762	5.50
2	3	x_2x_3	415.4427	0.84703	0.81644	41.5443	62.44
2	3	x_2x_4	868.8801	0.68006	0.61607	86.8880	138.23
2	3	x_3x_4	175.7380	0.93529	0.92235	17.5738	22.37
3	4	$x_1x_2x_3$	48.1106	0.98228	0.97638	5.3456	3.04
3	4	$x_1x_2x_4$	47.9727	0.98234	0.97645	5.3303	3.02
3	4	$x_1x_3x_4$	50.8361	0.98128	0.97504	5.6485	3.50
3	4	$x_2x_3x_4$	73.8145	0.97282	0.96376	8.2017	7.34
4	5	$x_1x_2x_3x_4$	47.8636	0.98238	0.97356	5.9829	5.00

Least-Squares Estimates for All Possible Regressions (Hald Cement Data)

Variables in Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
x_1	81.479	1.869			
x_2	57.424		0.789		
x_3	110.203			−1.256	
x_4	117.568				−0.738
x_1x_2	52.577	1.468	0.662		
x_1x_3	72.349	2.312		0.494	
x_1x_4	103.097	1.440			−0.614
x_2x_3	72.075		0.731	−1.008	
x_2x_4	94.160		0.311		−0.457
x_3x_4	131.282			−1.200	−0.724
$x_1x_2x_3$	48.194	1.696	0.657	0.250	
$x_1x_2x_4$	71.648	1.452	0.416		−0.237
$x_2x_3x_4$	203.642		−0.923	−1.448	−1.557
$x_1x_3x_4$	111.684	1.052		−0.410	−0.643
$x_1x_2x_3x_4$	62.405	1.551	0.510	0.102	−0.144



Plot of R_p^2 versus p

Matrix of Simple Correlations for Hald's Data

	x_1	x_2	x_3	x_4	y
x_1	1.0				
x_2	0.229	1.0			
x_3	−0.824	−0.139	1.0		
x_4	−0.245	−0.973	0.030	1.0	
y	0.731	0.816	−0.535	−0.821	1.0

Comparisons of Two Models for Hald's Cement Data

Observation i	$\hat{y} = 52.58 + 1.468x_1 + 0.662x_2$[a] e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$	$\hat{y} = 71.65 + 1.452x_1 + 0.416x_2 - 0.237x_4$[b] e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$
1	−1.5740	0.25119	4.4184	0.0617	0.52058	0.0166
2	−1.0491	0.26189	2.0202	1.4327	0.27670	3.9235
3	−1.5147	0.11890	2.9553	−1.8910	0.13315	4.7588
4	−1.6585	0.24225	4.7905	−1.8016	0.24431	5.6837
5	−1.3925	0.08362	2.3091	0.2562	0.35733	0.1589
6	4.0475	0.11512	20.9221	3.8982	0.11737	19.5061
7	−1.3031	0.36180	4.1627	−1.4287	0.36341	5.0369
8	−2.0754	0.24119	7.4806	−3.0919	0.34522	22.2977
9	1.8245	0.17195	4.9404	1.2818	0.20881	2.6247
10	1.3625	0.55002	9.1683	0.3539	0.65244	1.0368
11	3.2643	0.18402	16.0037	2.0977	0.32105	9.5458
12	0.8628	0.19666	1.1535	1.0556	0.20040	1.7428
13	−2.8934	0.21420	13.5579	−2.2247	0.25923	9.0194
		PRESS $x_1, x_2 =$	93.8827		PRESS $x_1, x_2, x_4 =$	85.3516

[a] $R^2_{\text{Prediction}} = 0.9654$, $\text{VIF}_1 = 1.05$, $\text{VIF}_2 = 1.06$.

[b] $R^2_{\text{Prediction}} = 0.9684$, $\text{VIF}_1 = 1.07$, $\text{VIF}_2 = 18.78$, $\text{VIF}_4 = 18.94$.

Ignore VIF and h_{ii} for now. We will cover these topics later.

Best Subsets Regression: y versus x1, x2, x3, x4

Response is y

Vars	R-Sq	R-Sq(adj)	C-p	S	x 1	x 2	x 3	x 4
1	67.5	64.5	138.7	8.9639				X
1	66.6	63.6	142.5	9.0771		X		
1	53.4	49.2	202.5	10.727	X			
1	28.6	22.1	315.2	13.278			X	
2	97.9	97.4	2.7	2.4063	X	X		
2	97.2	96.7	5.5	2.7343	X			X
2	93.5	92.2	22.4	4.1921			X	X
2	84.7	81.6	62.4	6.4455		X	X	
2	68.0	61.6	138.2	9.3214		X		X
3	98.2	97.6	3.0	2.3087	X	X		X
3	98.2	97.6	3.0	2.3121	X	X	X	
3	98.1	97.5	3.5	2.3766	X		X	X
3	97.3	96.4	7.3	2.8638		X	X	X
4	98.2	97.4	5.0	2.4460	X	X	X	X

Computer output (Minitab) for Furnival and Wilson all-possible-regression algorithm.

All Possible Regressions Notes

- Once some candidate models have been identified, run regression analysis on each one individually and make comparisons (include the PRESS statistic).
- A caution about the regression coefficients. If the estimates of a particular coefficient tends to “jump around”, this could be an indication of multicollinearity. Jumping around is a technical term – example: if some estimates are positive and then negative.

Stepwise Regression Methods

Three types of stepwise regression methods

1. forward selection
2. backward elimination
3. stepwise regression (combination of forward and backward)

Stepwise Regression Methods

Forward Selection

- Procedure is based on the idea that no variables are in the model originally, but are added one at a time. The selection procedure is:

1. The first regressor selected to be entered into the model is the one with the highest correlation with the response. If the F statistic corresponding to the model containing this variable is significant (larger than some predetermined value, F_{in}), then that regressor is left in the model.
2. The second regressor examined is the one with the largest partial correlation with the response. If the *F*-statistic corresponding to the addition of this variable is significant, the regressor is retained.
3. This process continues until all regressors are examined.

Stepwise Regression: y versus x1, x2, x3, x4

Forward selection. Alpha-to-enter: 0.25

Response is y on 4 predictors, with N = 13

Step	1	2	3
Constant	117.57	103.10	71.65
x4	-0.738	-0.614	-0.237
T- Value	-4.77	-12.62	-1.37
P- Value	0.001	0.000	0.205
x1		1.44	1.45
T- Value		10.40	12.41
P- Value		0.000	0.000
x2			0.42
T- Value			2.24
R- Value			0.052
S	8.96	2.73	2.31
R- Sq	67.45	97.25	98.23
R- Sq(adj)	64.50	96.70	97.64
Mallows C- p	138.7	5.5	3.0

Forward selection results from Minitab for the Hald cement data.

Stepwise Regression Methods

Backward Elimination

Procedure is based on the idea that all variables are in the model originally, examined one at a time and removed if not significant.

1. The partial F statistic is calculated for each variable *as if it were the last one added to the model*. The regressor with the smallest F statistic is examined first and will be removed if this value is less than some predetermined value F_{out}.
2. If this regressor is removed, then the model is refit with the remaining regressor variables and the partial F statistics calculated again. The regressor with the smallest partial F statistic will be removed if that value is less than F_{out}.
3. The process continues until all regressors are examined.

Stepwise Regression: y versus x1, x2, x3, x4

Backward elimination. Alpha-to-Remove: 0.1

Response is y on 4 predictors, with N = 13

Step	1	2	3
Constant	62.41	71.65	52.58
x1	1.55	1.45	1.47
T-Value	2.08	12.41	12.10
P-Value	0.071	0.000	0.000
x2	0.510	0.416	0.662
T-Value	0.70	2.24	14.44
P-Value	0.501	0.052	0.000
x3	0.10		
T-Value	0.14		
P-Value	0.896		
x4	-0.14	-0.24	
T-Value	-0.20	-1.37	
P-Value	0.844	0.205	
S	2.45	2.31	2.41
R-Sq	98.24	98.23	97.87
R-Sq(adj)	97.36	97.64	97.44
Mallows C-p	5.0	3.0	2.7

Backward selection results from Minitab for the Hald cement data.

Stepwise Regression Methods

Stepwise Regression

This procedure is a modification of forward selection.

1. The contribution of each regressor variable that is put into the model is reassessed by way of its partial F statistic.
2. A regressor that makes it into the model, may also be removed it if is found to be insignificant with the addition of other variables to the model. If the partial F-statistic is less than F_{out}, the variable will be removed.
3. Stepwise requires both an F_{in} value and F_{out} value.

```
Stepwise Regression: y versus x1, x2, x3, x4

Alpha-to-Enter: 0.15  Alpha-to-Remove: 0.15

Response is y on 4 predictors, with N = 13

Step                1         2         3         4
Constant       117.57    103.10     71.65     52.58

x4             -0.738    -0.614    -0.237
T-Value         -4.77    -12.62     -1.37
P-Value         0.001     0.000     0.205

x1                         1.44      1.45      1.47
T-Value                   10.40     12.41     12.10
P-Value                   0.000     0.000     0.000

x2                                   0.416     0.662
T-Value                               2.24     14.44
P-Value                              0.052     0.000

S                8.96      2.73      2.31      2.41
R-Sq            67.45     97.25     98.23     97.87
R-Sq(adj)       64.50     96.70     97.64     97.44
Mallows C-p     138.7       5.5       3.0       2.7
```

Stepwise selection results from Minitab for the Hald cement data.

Stepwise Regression Methods

Cautions

- No one model may be the “best”
- The three stepwise techniques could result in different models
- Inexperienced analysts may use the final model simply because the procedure spit it out.