San José State University

Math 261A: Regression Theory & Methods

Variable Selection and Model Building

Dr. Guangliang Chen

This lecture is based on the following textbook sections:

- Chapter 10: 10.1 – 10.3

Outline of this presentation:

- Consequences of model misspecification
- Criteria for evaluating subset regression models
- Computational techniques for variable selection

Introduction

In previous chapters when performing regression, we assume that

- we have a very good idea of the basic form of the model, and
- we know all (or nearly all) of the regressors that should be used.

Our focus was on techniques to ensure that

- the functional form of the model was correct, and
- the underlying assumptions were not violated.

Our **basic strategy** is as follows:

1. Fit the full model (with all of the regressors under consideration).
2. Perform a thorough analysis of this model, including a full residual analysis.
3. Determine if transformations of the response or of some of the regressors are necessary.
4. Use the t tests on the individual regressors to edit the model.
5. Perform a thorough analysis of the edited model, especially a residual analysis, to determine the model's adequacy.

However, in most practical problems, we face a rather large pool of candidate regressors, of which only a few are likely to be important.

Additionally, some of the important variables may be correlated with each other, so we don't really need all of them (even though individually they may appear important).

Finding an appropriate subset of regressors for the model is often called the **variable selection** problem.

Building a regression model that includes only a subset of the available regressors involves two **conflicting** objectives:

(1) Use as many regressors as possible for accurate estimation/prediction;

(2) Use as few regressors as possible so that the model is simple, yet still accurate.

The process of finding a model that is a compromise between these two objectives is called selecting the "best" regression equation.

Unfortunately, there is no unique definition of "best", and different variable selection procedures frequently specify different subsets of the candidate regressors as best.

Consequences of model misspecification

Assume a population regression model consisting of $K = k + r$ regressors

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\text{to be retained}} + \underbrace{\beta_{k+1} x_{k+1} + \cdots + \beta_{k+r} x_{k+r}}_{\text{to be deleted}} + \epsilon$$

The sample regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \begin{bmatrix} \mathbf{X}_p & \mathbf{X}_r \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_p \\ \boldsymbol{\beta}_r \end{bmatrix} + \boldsymbol{\epsilon} = \mathbf{X}_p \boldsymbol{\beta}_p + \mathbf{X}_r \boldsymbol{\beta}_r + \boldsymbol{\epsilon}$$

where $p = k + 1$.

For the **full model**, the least squares estimate of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\boldsymbol{\beta}}_p^* \\ \hat{\boldsymbol{\beta}}_r^* \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

In particular, $\hat{\boldsymbol{\beta}}_p^*$ **is an estimator of** $\boldsymbol{\beta}_p$**, and it is unbiased**.

For the subset model,

$$\mathbf{y} = \mathbf{X}_p\boldsymbol{\beta}_p + \boldsymbol{\epsilon}$$

the least squares of estimate of $\boldsymbol{\beta}_p$ is

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

We have thus obtained two estimators of $\boldsymbol{\beta}_p$: $\hat{\boldsymbol{\beta}}_p^*$ and $\hat{\boldsymbol{\beta}}_p$.

As variables are deleted from the model, for the retained variables in $\mathbf{X}_p$,

- we may potentially introduce bias into their coefficient estimates $\hat{\boldsymbol{\beta}}_p$

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\beta}}_p) &= (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'(\mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_r\boldsymbol{\beta}_r) \\ &= \boldsymbol{\beta}_p + (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{X}_r\boldsymbol{\beta}_r \end{aligned}$$

  unless the two sets of variables are orthogonal ($\mathbf{X}_p'\mathbf{X}_r = \mathbf{O}$).

- meanwhile, we may improve the variance (precision) of $\hat{\boldsymbol{\beta}}_p$

Overall, we could reduce the mean square error (MSE) of $\hat{\boldsymbol{\beta}}_p$, if the deleted variables have small effects.

Criteria for Evaluating Subset Regression Models

Two key aspects of the variable selection problem are generating the subset models and deciding if one subset is better than another.

We have the following **evaluation criteria**:

- Coefficient of determination R^2
- Adjusted R^2
- Residual mean square MS_{Res}
- Mallow's C_p statistic
- AIC and BIC

Coefficient of determination

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}$$

R^2 can be used to compare subset regression models that have the same number of predictors.

Generally, R^2 is not used as a criterion for choosing the number of regressors to include in the model.

Adjusted R^2

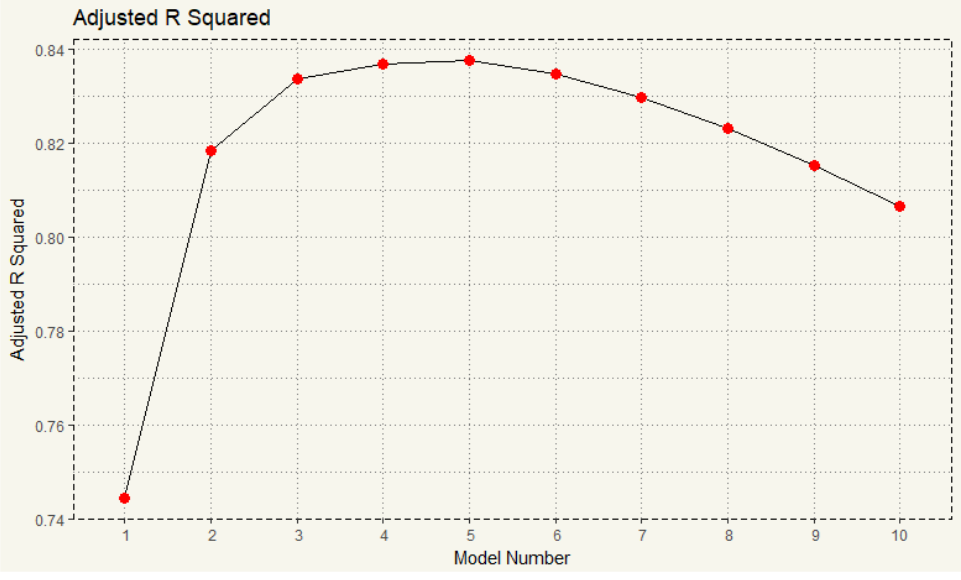
$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-p}(1-R^2) = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)}$$

where $p = k + 1$ and k is the number of regressors (subset size).

This measure can be used to compare subset regression models with different numbers of regressors.

How to use R^2_{adj} for choosing the optimal subset of regressors:

- For each subset size $k = 1, \ldots, K$, find the best k regressors that maximize R^2 (and also R^2_{adj}). Denote the maximum by $R^2_{\text{adj}}(k)$.
- Compare $R^2_{\text{adj}}(k)$ for all k and select k such that $R^2_{\text{adj}}(k)$ is highest.



Residual mean square

$$MS_{Res} = \frac{SS_{Res}}{n-k-1}$$

It can also be used as a model evaluating/selection criterion:

- For each subset size $k = 1, \ldots, K$, find the best subset of k regressors that minimizes MS_{Res}. Denote the minimum by $MS_{Res}(k)$.
- Compare $MS_{Res}(k)$ for different k and select k such that $MS_{Res}(k)$ is smallest, or approximately equal to that of the full model.

This criterion (minimum MS_{Res}) is equivalent to the maximum adjusted R^2 criterion, because $R^2_{\text{adj}} = 1 - \frac{MS_{Res}}{SS_T/(n-1)}$

Mallows' C_p statistic

$$C_p = \frac{1}{\sigma^2}\underbrace{SS_{Res}(p)}_{\text{fitting error}} - n + \underbrace{2p}_{\text{penalty}} \qquad (p = k + 1)$$

It can be shown that

$$\mathrm{E}(SS_{Res}(p)) = \sum_{i=1}^{n}\underbrace{(\mathrm{E}(\hat{y}_i) - \mathrm{E}(y_i))}_{\text{bias}}^2 + (n-p)\sigma^2$$

If the model has zero bias (such as the OLS),

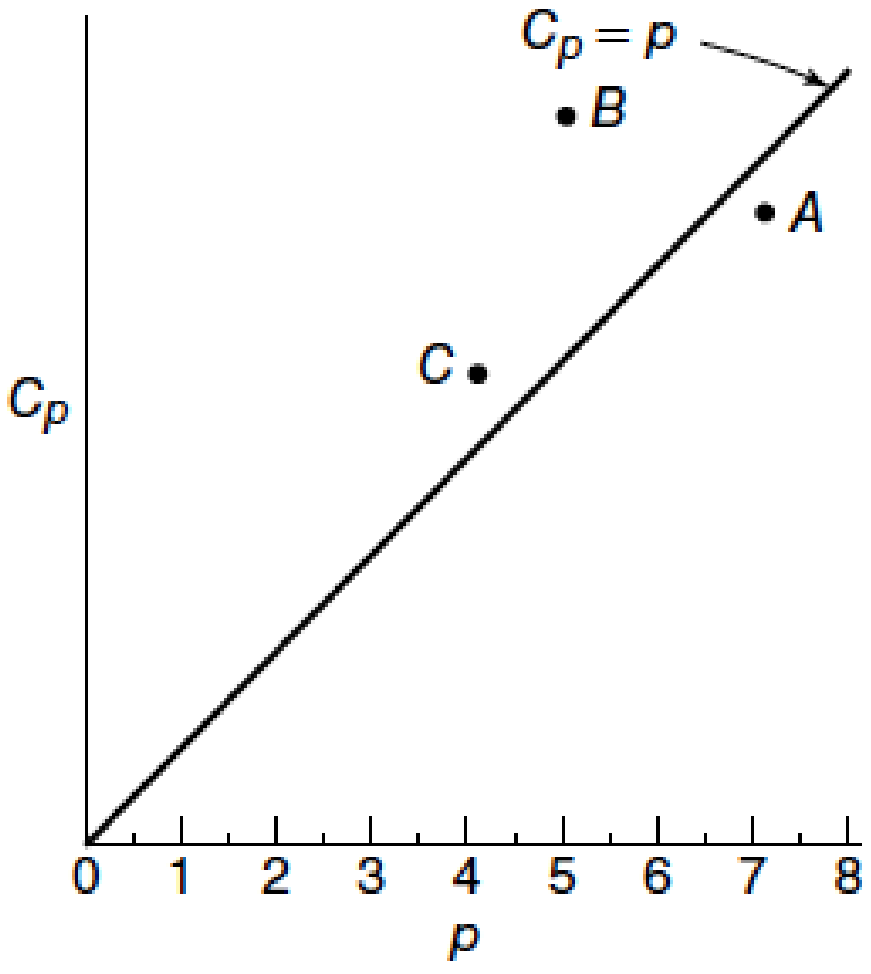
$$\mathrm{E}(C_p) = \frac{(n-p)\sigma^2}{\sigma^2} - n + 2p = p$$

(Otherwise, it will be bigger than p)

Regression equations with little bias will have values of C_p near the line $C_p = p$ while those with substantial bias will fall above this line.

Generally, **small values of C_p are desirable** (in the right plot, Model C should be preferred to A and B).

To calculate C_p, we need an unbiased estimate of σ^2. Frequently, we use the residual mean square of the full model for this purpose.



Two more commonly-used model selection criteria:

- **Akaike Information Criterion**:

$$\text{AIC} = -2\log(L) + 2p = n\log(SS_{Res}/n) + 2p$$

  It is based on maximizing the expected entropy of the model.

- **Bayesian Information Criterion**:

$$\text{BIC} = -2\log(L) + p\log(n) = n\log(SS_{Res}/n) + p\log(n)$$

  This criterion is also based on information theory but set within a Bayesian context. Comparing with AIC, it places a greater penalty on adding regressors as the sample size increases.

Computational Techniques for Variable Selection

- All possible regressions ⟵ brute-force, exhaustive search
- Stepwise regression methods ⟵ smarter, but no guarantee
  - Forward selection
  - Backward elimination
  - Stepwise regression (hybrid scheme)

All possible regressions

This procedure requires that the analyst fit all the regression equations involving 1 candidate regressor, 2 candidate regressors, and so on.

If there are K candidate regressors, there are 2^K total equations to be estimated and examined. $\longleftarrow$ not practical for large K

These equations are evaluated according to some suitable criterion and the "best" regression model selected.

The R function REGSUBSETS() in the *leaps* package can be used to perform all possible regressions.

Example: The Hald Cement Data

Hald [1952] presents data concerning the heat evolved in calories per gram of cement (y) as a function of the amount of each of four ingredients in the mix: tricalcium aluminate (x_1), tricalcium silicate (x_2), tetracalcium alumino ferrite (x_3), and dicalcium silicate (x_4).

The data set is rather small, containing only 13 observations.

Since there are $K = 4$ candidate regressors, there are $2^4 = 16$ possible regression equations.

Variable Selection and Model Building

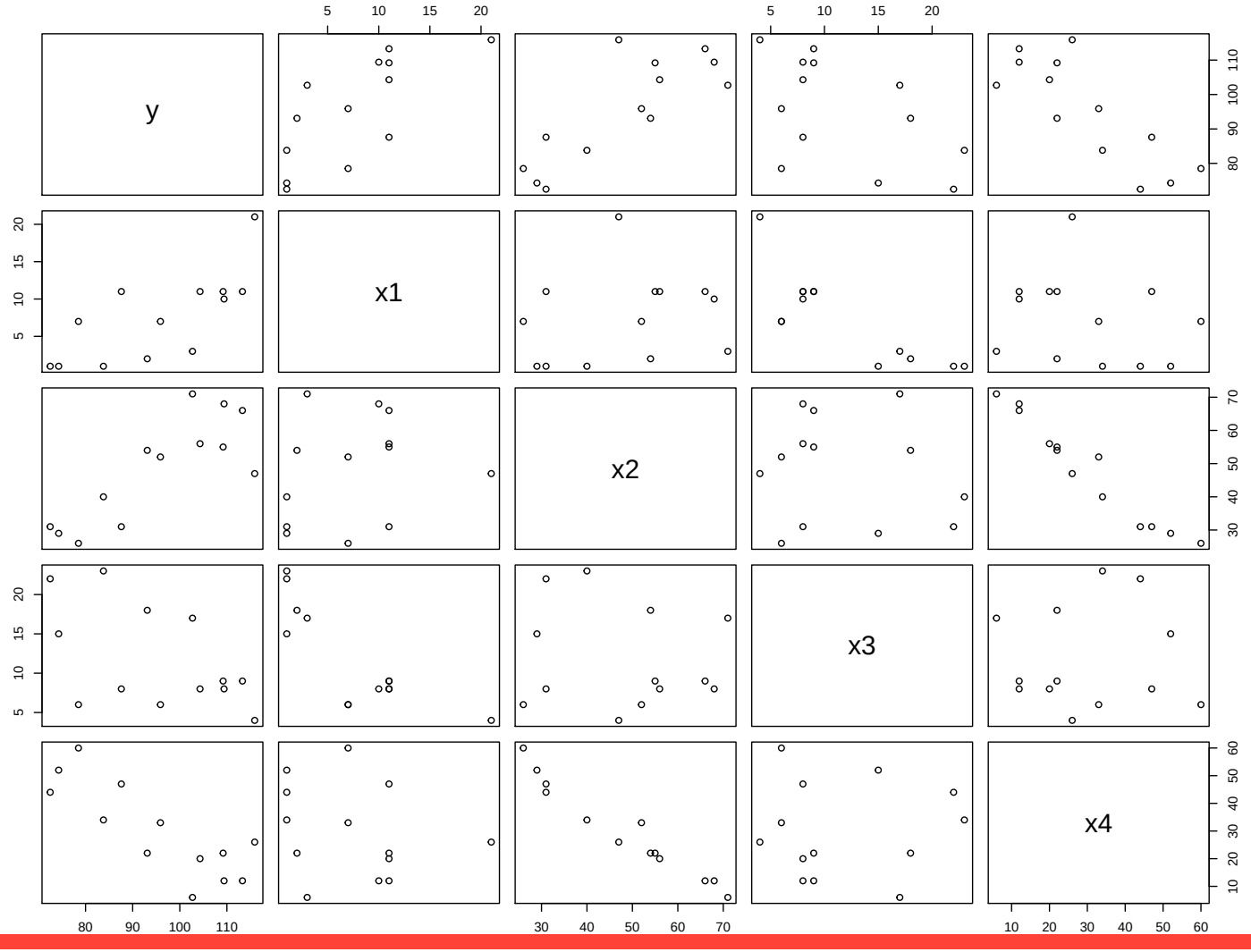


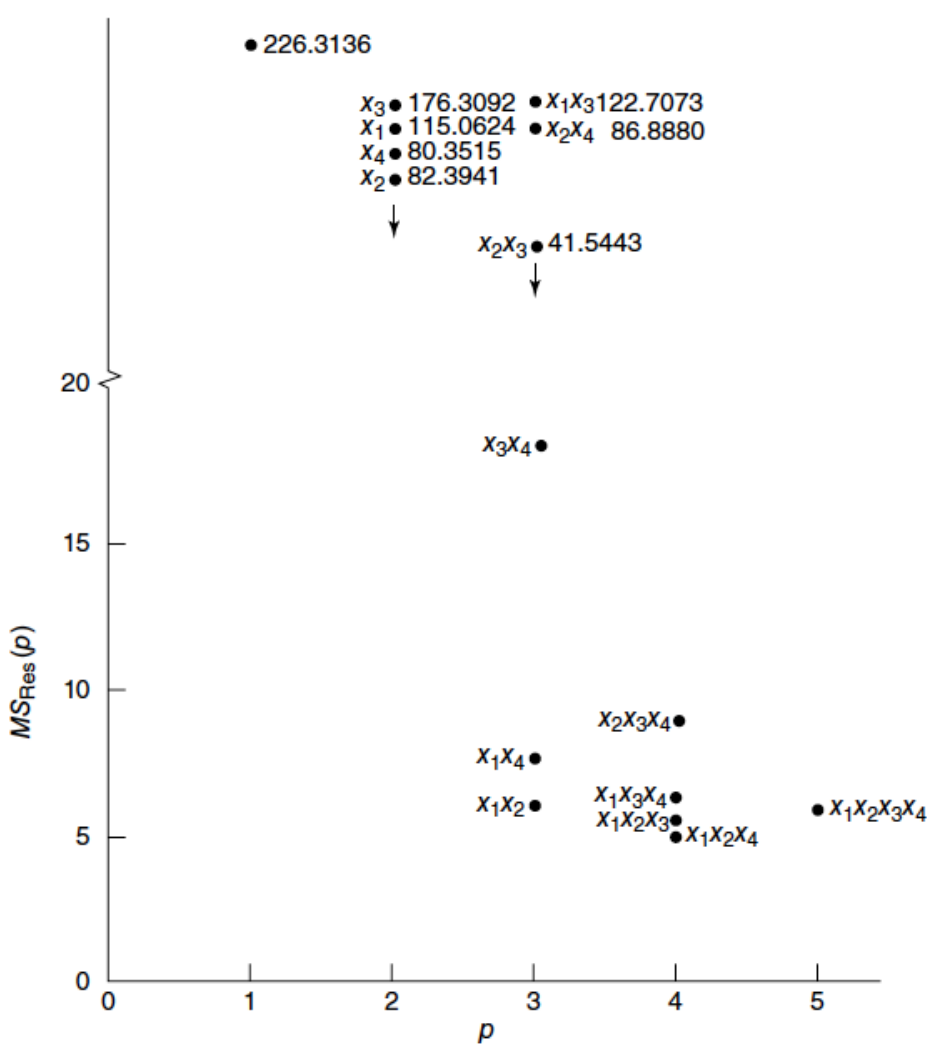
TABLE 10.1 Summary of All Possible Regressions for the Hald Cement Data

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

TABLE 10.2 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

Variable Selection and Model Building



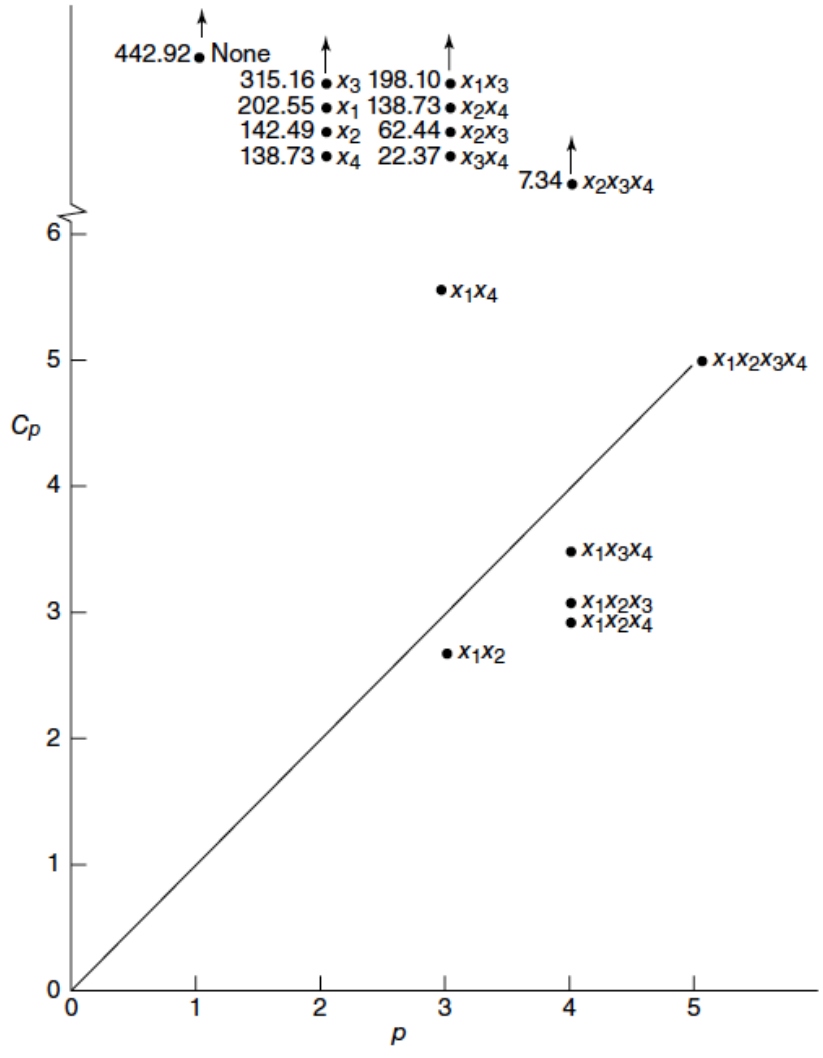


TABLE 10.3 Matrix of Simple Correlations for Hald's Data in Example 10.1

	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

TABLE 10.4 Comparisons of Two Models for Hald's Cement Data

Observation i	$\hat{y} = 52.58 + 1.468x_1 + 0.662x_2$[a]			$\hat{y} = 71.65 + 1.452x_1 + 0.416x_2 - 0.237x_4$[b]		
	e_i	h_{ii}	$[e_i/(1-h_{ii})]^2$	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] $R^2_{\text{Prediction}} = 0.9684$, $\text{VIF}_1 = 1.07$, $\text{VIF}_2 = 18.78$, $\text{VIF}_4 = 18.94$.

All possible regressions with a categorical predictor:

The REGSUBSETS() function can be used in the same way. However, in this scenario, the categorical variable (with ℓ levels) is reduced to $\ell - 1$ indicator variables (treated as new predictors), so that effectively there are a total of $(K - 1) + (\ell - 1)$ predictors for forming subset models.

For certain subset size, it is possible that the best model of that size uses only some but not all of the $\ell - 1$ indicator variables. If such a model turns out to be the best overall, it still implies that the categorical variable is selected by the final model, just that some of the $\ell - 1$ indicator variables have zero coefficients (which means that those levels are no different from the reference level and they will share the same intercept).

Stepwise regression methods

- Forward selection
- Backward elimination
- Stepwise regression (combination of forward and backward actions)

Forward selection: add regressors from a candidate set $\{x_1, \ldots, x_K\}$, one at a time, until certain stopping condition is met.

Cutoff needed: α_{IN}

Step 0: Start without any regressor in the model (only the intercept)

Step 1: Add the most significant regressor to the model if the corresponding F statistic has a p-value $< \alpha_{\text{IN}}$:

$$F = \frac{SS_R(\beta_j \mid \beta_0)}{MS_{Res}(\beta_0, \beta_j)}, \quad j = 1, \ldots, K$$

Suppose x_1 is added to the model.

Step 2: For each remaining regressor $x_j, j = 2, \ldots, K$, add the one with the largest partial F statistic

$$F = \frac{SS_R(\beta_2 \mid \beta_1, \beta_0)}{MS_{Res}(\beta_0, \beta_1, \beta_2)}$$

(if the p-value is less than α_{IN}, otherwise terminate the procedure).

Repeat the procedure with the remaining regressors until no regressor can be added.

Backward elimination: eliminate regressors one at a time.

Cutoff needed: α_{OUT}

Step 0: Fit a model with all regressors

Step 1: Compute the partial F statistic for each regressor in the model (given all other regressors) and remove the regressor with the largest p-value if it exceeds the threshold α_{OUT}

Step 2: Fit a new model with the remaining regressors, and repeat the above procedure until no regressor can be eliminated from the model

Stepwise regression: a combination of forward selection and backward elimination actions

Cutoffs needed: $\alpha_{\text{IN}}, \alpha_{\text{OUT}}$

- $\rightarrow$ Start with no regressors in the model, and add regressors one at a time (using the cutoff α_{IN})
- $\leftarrow$ Each time a new regressor is added, check to see if any of the previously added regressors may be eliminated from the model (using the cutoff α_{OUT})
- $\bigcirc$ Repeat until no regressor can be added to the model

Comments:

- Backward elimination is often a very good variable selection procedure. It is particularly favored by analysts who like to see the effect of including all the candidate regressors.
- Berk [1978] has noted that forward selection tends to agree with all possible regressions for small subset sizes but not for large ones, while backward elimination tends to agree with all possible regressions for large subset sizes but not for small ones.
- The three procedures do not necessarily lead to the same final model.
- None of them guarantees to find the best subset regression model.

Stepwise regression when categorical variables are present

Suppose there are K candidate predictors, among which there is a categorical predictor x_j with ℓ levels.

The R functions for the three methods are used in the same way as for continuous variables (as long as x_j has been converted by as.factor()).

x_j is treated as a single variable (not $\ell - 1$ separate indicator variables) and thus there is only a single partial F statistic

$$F = \frac{SS_R(\beta_j \mid \ldots)/(\ell - 1)}{MS_{Res}(\beta_j, \ldots)}$$

and a single p-value (to be used to make the corresponding decision).