Model Building with Variable Selection – Part III

Johns Hopkins Engineering

625.461 Statistical Models and Regression

Module 9 – Lecture 9D

Computational Techniques for Variable Selection

All Possible Regressors

With K regressors (intercept is always in the model), there are 2^K total equations to consider

The Hald Cement Data (Ex 10.1, Table B.21, p.338)

y: heat evolved in calories per gram of cement
As a function of four ingredients in the mix,
x_1: tricalcium aluminate
x_2: tricalcium silicate
x_3: tetracalcium alumino ferrite
x_4: dicalcium silicate

Table 10.1, 10.2 (p.339): summary of all possible regressions

The Hald Cement Data: Plot of $R^2(p)$ versus p

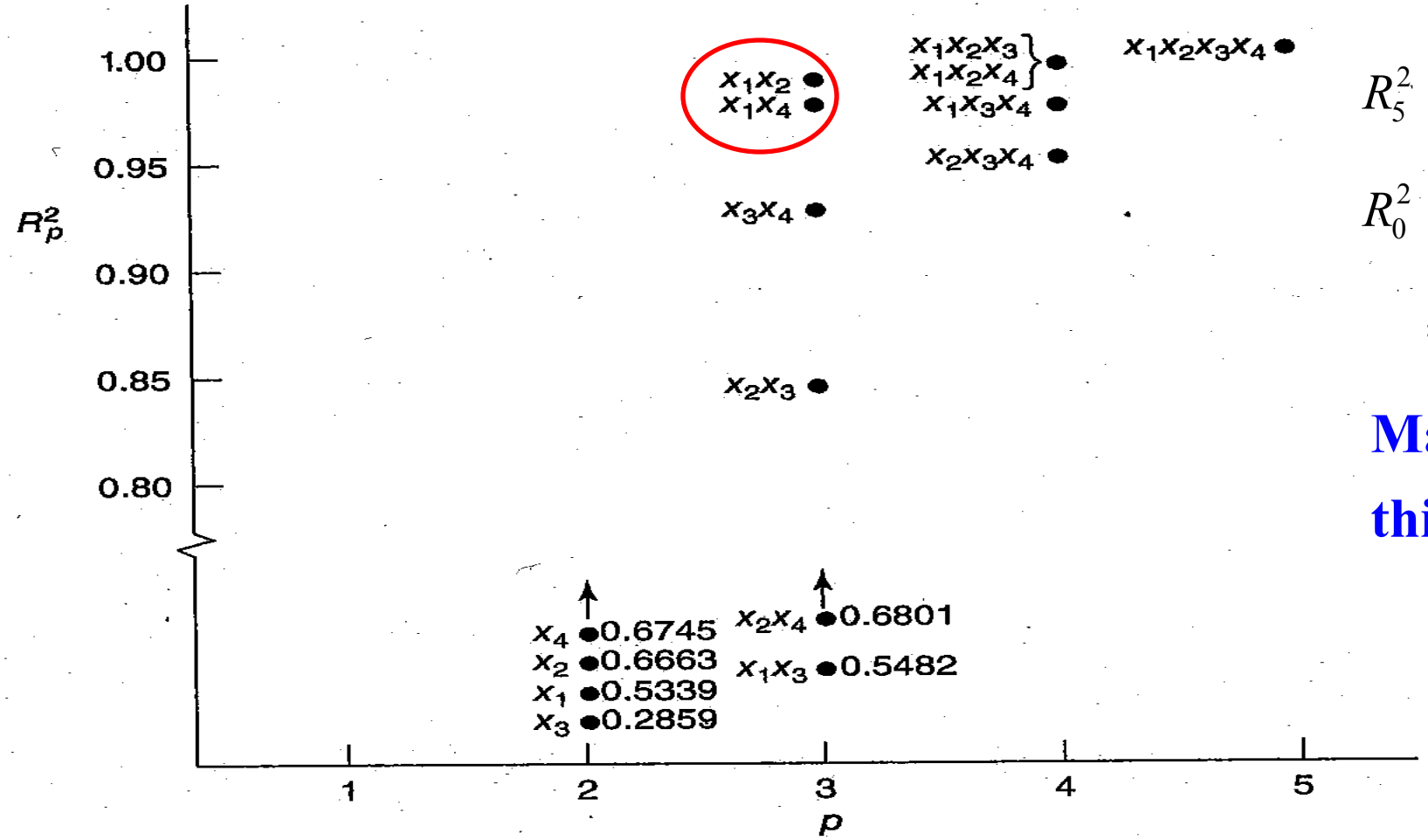


Figure 10.4 Plot of R_p^2 versus p, Example 10.1.

$$R_5^2 = 0.98238$$

$$R_0^2 = 1-(1-R_5^2)\left(1+\frac{4F_{0.05,4,8}}{8}\right)$$

$$= 0.94855$$

Many models satisfy this criterion.

The Hald Cement Data

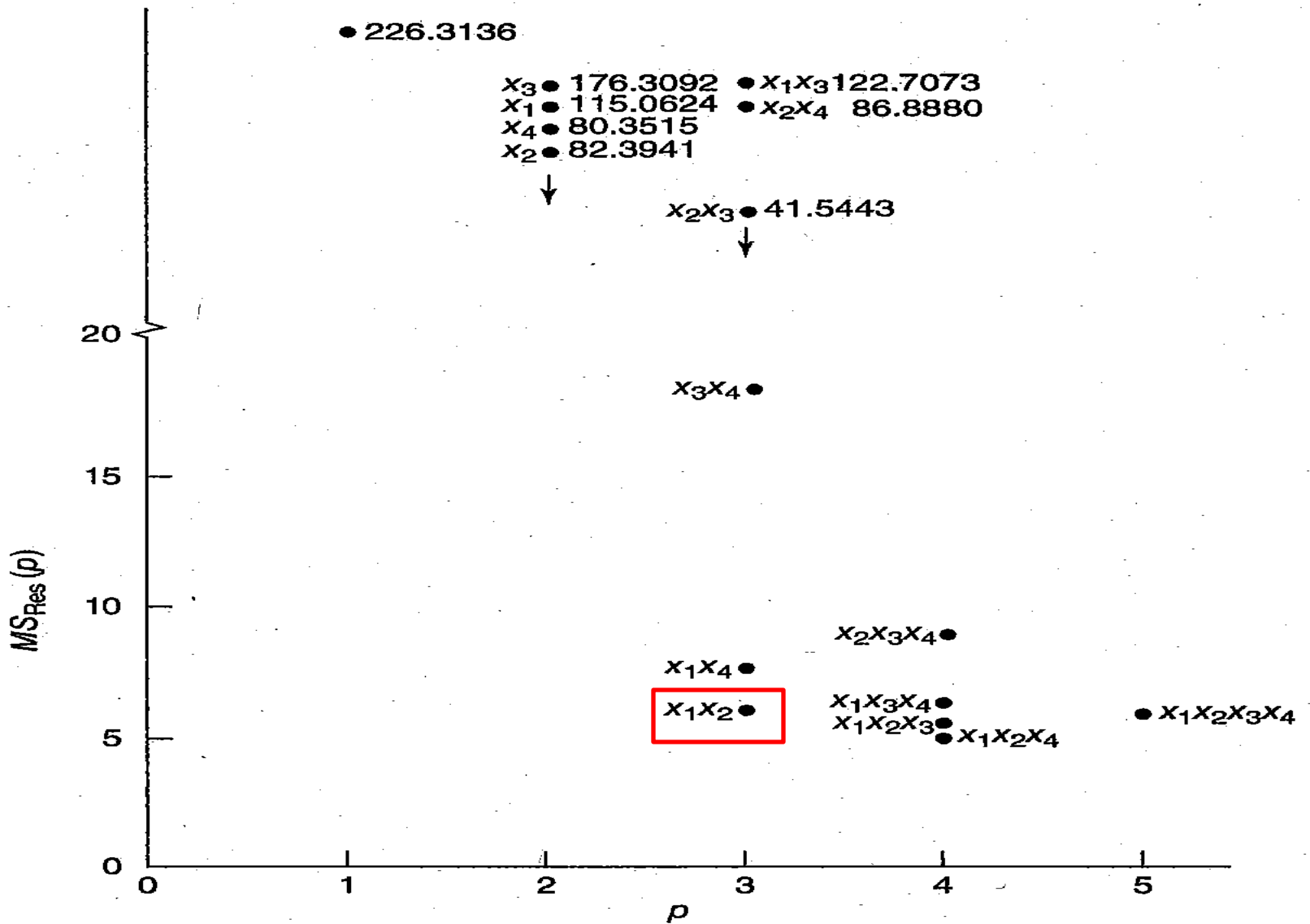
Examine pairwise correlations between the regressors and between y and regressors

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

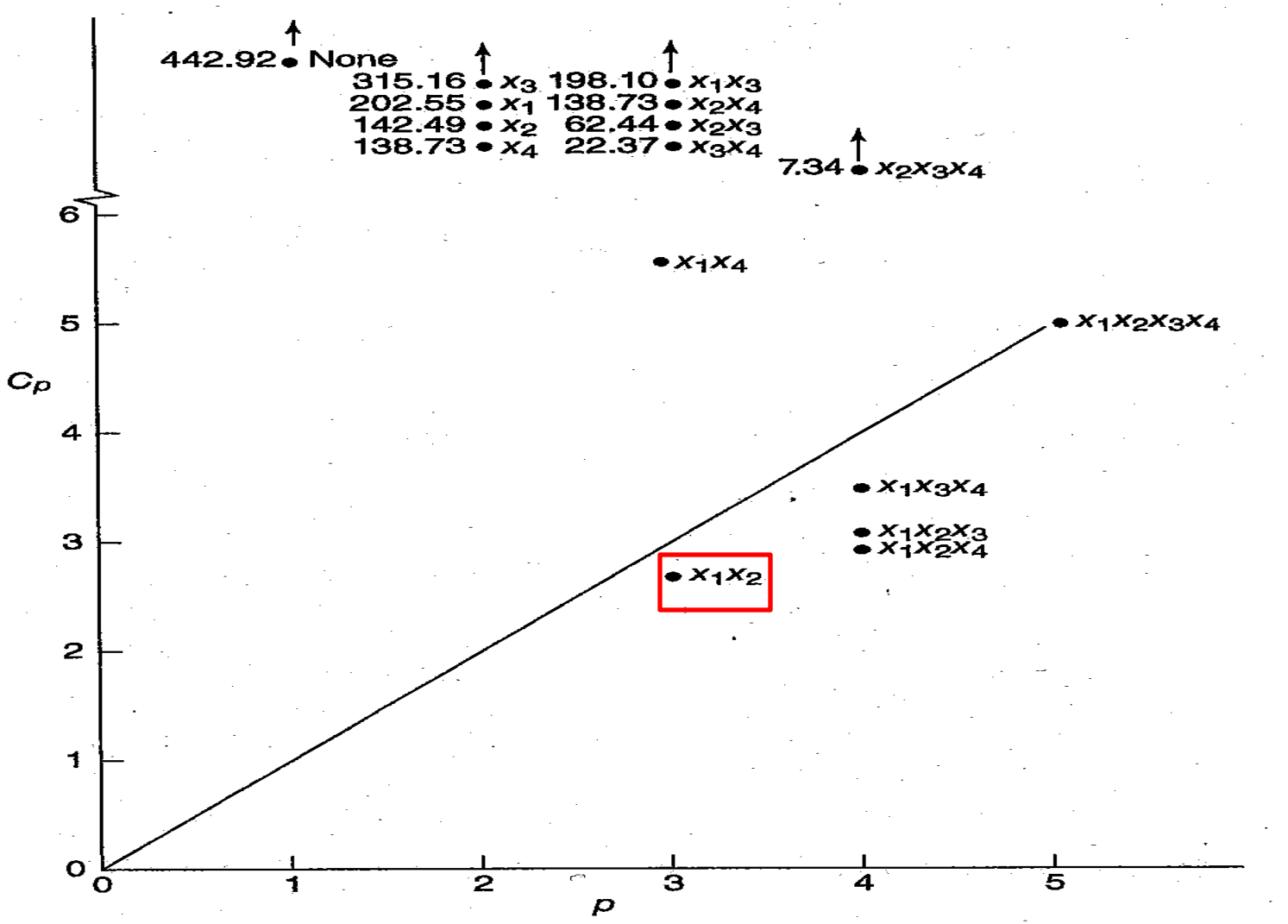
	x_1	x_2	x_3	x_4	y
x_1	1.0				
x_2	0.229	1.0			
x_3	−0.824	−0.139	1.0		
x_4	−0.245	−0.973	0.030	1.0	
y	0.731	0.816	−0.535	−0.821	1.0

When x_1 and x_2 or when x_2 and x_4 are already in the model, adding further regressors will be of little value.

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



Plot C_p versus p



The Hald Cement Data: Comparison of Models

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

⇒ **highly multicollinear**

Some Important Notes

There is no clear-cut choice of the best regression equation. Very often we find that different criteria suggest different equations.

All “final” candidate models should be subjected to the usual tests for adequacy, including investigation of leverage points, influence, and multicollinearity.

JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING