Model Building with Variable Selection – Part II

Johns Hopkins Engineering

625.461 Statistical Models and Regression

Module 9 – Lecture 9C

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Motivation for Variable Selection

By deleting regressors, we may improve the precision of the parameter estimates of the retained variables, and improve the precision of a predicted response

That is, there is a danger in retaining negligible variables – increase the variances of the parameter estimates and a predicted response

but at a cost of "bias"

Criteria for Evaluating Subset Models

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

There are $\binom{K}{p-1}$ values of R_p^2 for each value of p, one for each possible subset model of size p.

Plot maximum R_p^2 for each p versus p.

Criteria for Evaluating Subset Models: Use of R^2

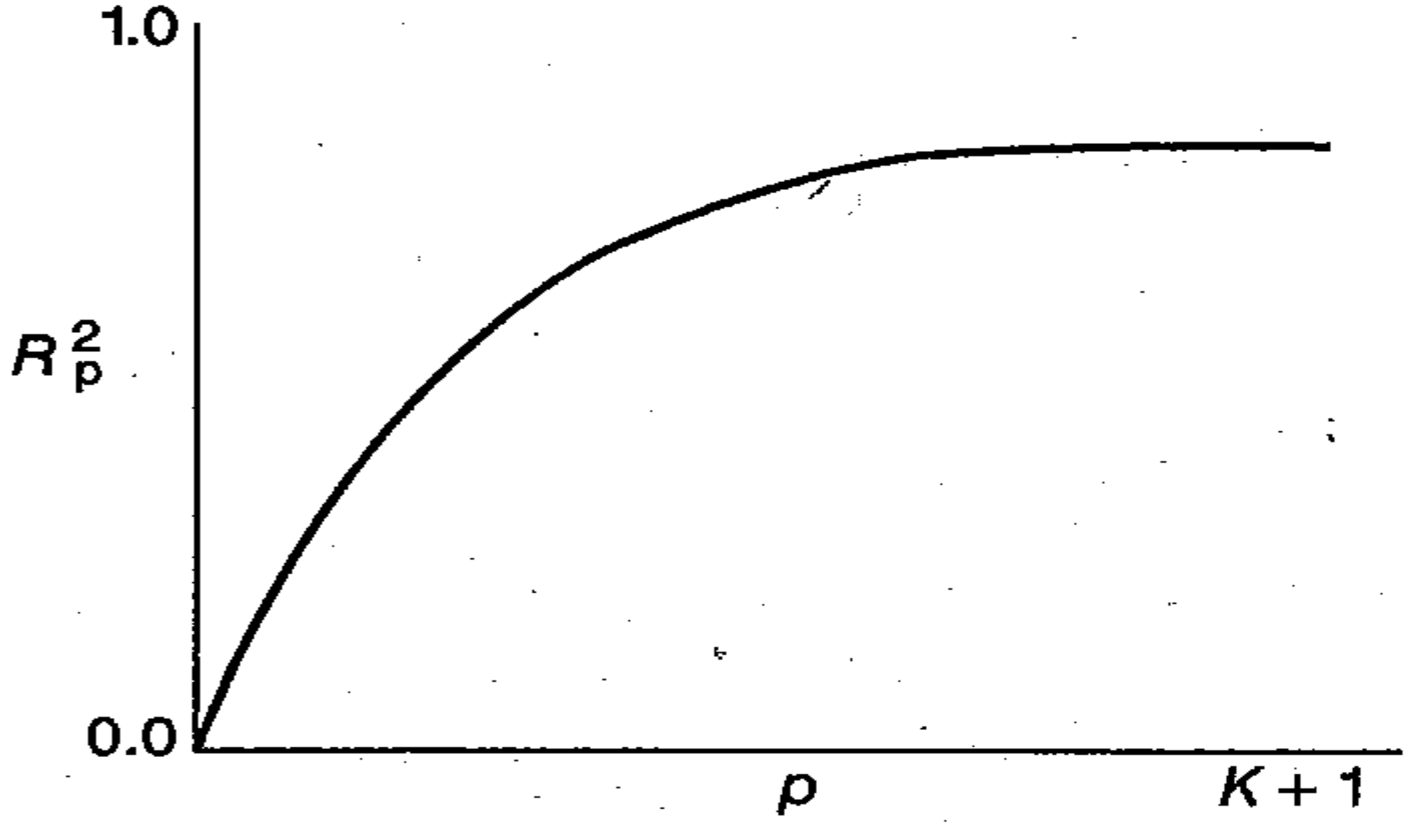

Figure 10.1 Plot of R_p^2 versus p.

Test for Selection based on R^2

Aitkin (1974): Provide a test by which all subset regression models that have an R^2 not significantly different from the R^2 (labeled R^2_{K+1}) for the full model can be identified.

$$R_0^2 = 1 - (1 - R_{K+1}^2)(1 + d_{\alpha,n,k})$$

$$d_{\alpha,n,k} = \frac{KF_{\alpha,K,n-K-1}}{n-K-1}$$

Look for $R^2 > R_0^2$

Residual Mean Squares

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

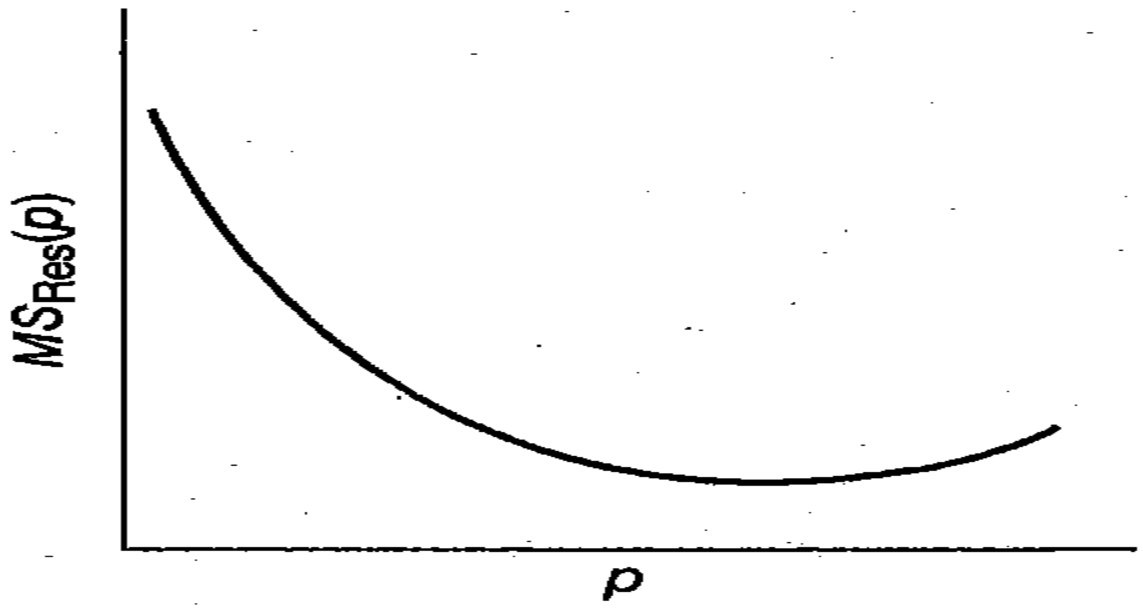

Figure 10.2 Plot of $MS_{\text{Res}}(p)$ versus p.

Mallow's C_p Statistic

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

The total squared bias for a p-term model

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

Mallow's C_p Statistic

The standardized mean square error of fitted values

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

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

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

Mallow's C_p Statistic

$$\Gamma_p = \frac{1}{\sigma^2}\{E[SS_{\text{Res}}(p)] - (n-p)\sigma^2 + p\sigma^2\} = \frac{E[SS_{\text{Res}}(p)]}{\sigma^2} - n + 2p$$

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

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

Mallow's C_p Statistic

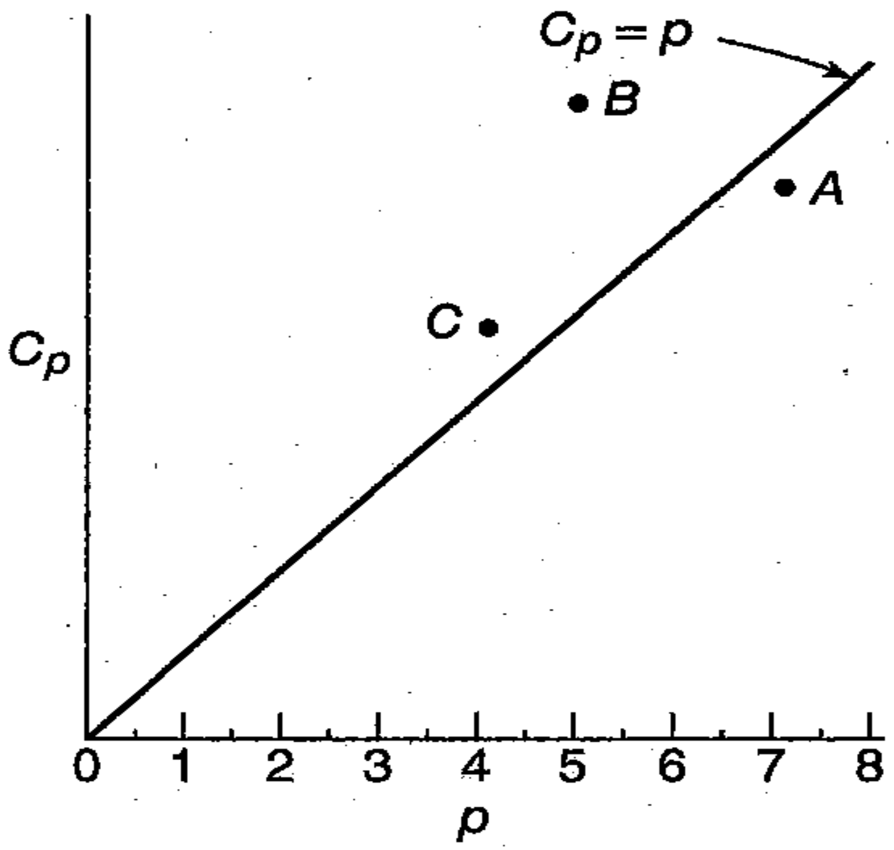

Figure 10.3 A C_p plot.

Smaller values of C_p are desirable

Akaike Information Criterion (AIC) and Bayesian alike (BIC)

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

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

Look for small AIC

Some Notes for Variable Selection

Uses of regressions and Model evaluation criteria

1. Obtain a good description of a process or model a complex system

   Search a regression equation to minimize residual SS

2. Estimate mean response or predict a future observation

   Select a regression model with a small PRESS

Some Notes for Variable Selection

3. Control

Accurate estimates of parameters are important.

The standard errors of the regression coefficients should be small.

JOHNS HOPKINS

WHITING SCHOOL *of* ENGINEERING