1997 Comprehensive Exam: Stat 105 Question

Describe the three quantities R_a^2, PRESS and C_p used as diagnostics for model selection in linear regression. Your account should include the definitions of the three quantities, brief outlines of their derivation, and a discussion of their comparative merits. You may assume the standard assumptions of the linear model.

An experiment results in $2n+1$ pairs of observations (x_i, y_i), where $x_i = 0, \pm 1, \pm 2, \ldots, \pm n$. The errors are uncorrelated with common variance σ^2. The statistician analyzing the data is considering two models, (a) the linear model $\mathrm{E}\{y_i\} = \beta_0 + \beta_1 x_i$, and (b) the quadratic model $\mathrm{E}\{y_i\} = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$. Unknown to the statistician, the quadratic model is in fact the correct one, but it is still possible that if $|\beta_2|$ is small enough, better predictions will be obtained by assuming the linear model, provided "better" is defined in a suitable way.

Derive expressions for the total mean squared prediction error, summed over all the given x_i values, when either of the models (a) or (b) is used to generate the predictions. Show that, if this is the criterion used to compare the models, then it is better to use model (a) whenever

$$\beta_2^2 < \frac{45\sigma^2}{n(n+1)(2n+3)(2n+1)(2n-1)}.$$

[*Hint:* The following formulas may be helpful:

$$\sum_1^n i = \frac{n(n+1)}{2}, \qquad \sum_1^n i^2 = \frac{n(n+1)(2n+1)}{6},$$

$$\sum_1^n i^3 = \frac{n^2(n+1)^2}{4}, \qquad \sum_1^n i^4 = \frac{n(n+1)(2n+1)(3n^2+3n-1)}{30}.]$$

Solution

If the sum of squares of errors and total sum of squares are denotes SSE and SSTO respectively, then R_a^2 is defined by

$$R_a^2 = 1 - \frac{(n-1)\mathrm{SSE}}{(n-p)\mathrm{SSTO}}. \tag{1}$$

One motivation of this is that the mean squared error is $\mathrm{SSE}/(n-p)$ under the model and $\mathrm{SSTO}/(n-1)$ if there is no regression, so the second term in (1) is a ratio of mean squared errors, corrected for degrees of freedom.

The other two criteria, PRESS and C_p, both start from trying to minimize the mean squared prediction error. Suppose y_i is the i'th data point, $\hat{y}_i$ its predicted value under some model with p parameters, and let $b_i = \mathrm{E}\{y_i - \hat{y}_i\}$ be the prediction bias. The sum of squared prediction biases is given by

$$\mathrm{E}\left\{\sum (y_i - \hat{y}_i)^2\right\} = \sum b_i^2 + \sum \mathrm{Var}(\hat{y}_i). \tag{2}$$

The second term in (2) reduces to $\sum h_{ii}\sigma^2 = p\sigma^2$. However the first term in (2) is not so easy to handle because we must estimate the prediction biases.

PRESS deals with this problem by using the deletion residuals, leading to

$$\begin{aligned} \mathrm{PRESS} &= \sum (y_i - \hat{y}_{i(i)})^2 \\ &= \sum \left(\frac{e_i}{1 - h_{ii}}\right)^2 \end{aligned}$$

using a standard formula to relate the deletion residuals $y_i - \hat{y}_{i(i)}$ to the ordinary residuals e_i and the diagonal entries h_{ii} of the hat matrix.

The alternative method due to Mallows begins by defining a standardized form of (2) as

$$\begin{aligned} \Gamma_p &= \frac{1}{\sigma^2}\left\{\sum b_i^2 + \sum \mathrm{Var}(\hat{y}_i)\right\} \\ &= \frac{1}{\sigma^2}\sum b_i^2 + p. \end{aligned}$$

However since we can also show that for a model of order p,

$$\begin{aligned} \mathrm{E}\{\mathrm{SSE}_p\} &= \mathrm{E}\left\{\sum (y_i - \hat{y}_i)^2\right\} \\ &= \sum b_i^2 + \sum \mathrm{Var}(y_i - \hat{y}_i) \\ &= \sum b_i^2 + (n-p)\sigma^2, \end{aligned}$$

it follows that an unbiased estimator of $\sum b_i^2$ is

$$\mathrm{SSE}_p - (n-p)\sigma^2.$$

Hence a suitable estimator of Γ_p is

$$\frac{1}{\sigma^2}\{\mathrm{SSE}_p - (n-p)\sigma^2 + p\sigma^2\} = \frac{\mathrm{SSE}_p}{\sigma^2} - (n-2p).$$

To complete the derivation, we need an estimator of σ^2 that does not depend on the order of the model being considered. For this it is usual to use s_P^2, the mean squared error under the full model with all P regressors included. This leads to

$$C_p = \frac{\mathrm{SSE}_p}{s_P^2} - (n-2p).$$

Comparisons: R_a^2 is a crude measure whose main advantage is that it is easy to compute. It is not especially effective as a model selection device. PRESS and C_p both attempt to estimate the mean squared prediction error in an unbiased way, and are considered equally effective.

All the above is *bookwork* in the sense that the derivations were given in class and in the printed notes to which the students had access. I will accept any reasonable approximation to the above!

Last part: in view of the formula $\sum b_i^2 + p\sigma^2$ for the sum of mean squared prediction errors, this essentially comes down to evaluating the bias terms b_i when the linear model (a) is assumed. There is of course no bias when the quadratic model (b) is used.

In this case the estimators are $\hat{\beta}_0 = \bar{y}$ and $\hat{\beta}_1 = \sum x_i y_i / \sum x_i^2$. We have

$$\begin{aligned} \mathrm{E}\{\hat{\beta}_0\} &= \frac{1}{2n+1}\sum(\beta_0 + \beta_1 x_i + \beta_2 x_i^2) = \beta_0 + \frac{\beta_2}{2n+1}\sum x_i^2, \\ \mathrm{E}\{\hat{\beta}_1\} &= \frac{1}{\sum x_i^2}\sum(\beta_0 x_i + \beta_1 x_i^2 + \beta_2 x_i^3) = \beta_1, \end{aligned}$$

since $\sum x_i = \sum x_i^3 = 0$.

Thus we have

$$\begin{aligned} b_i &= \mathrm{E}\left\{y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right\} \\ &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 - \beta_0 - \frac{\beta_2}{2n+1}\sum x_j^2 - \beta_1 x_i \\ &= \beta_2\left(x_i^2 - \frac{1}{2n+1}\sum x_j^2\right). \end{aligned}$$

Hence

$$\sum b_i^2 = \beta_2^2 \left\{ \sum x_i^4 - \frac{1}{2n+1} \left(\sum x_j^2 \right)^2 \right\} = \beta_2^2 \cdot \frac{n(n+1)(2n+3)(2n+1)(2n-1)}{45}$$

where the last expression is easily derived from the formulas given in the Hint.

The linear model then results in a smaller total mean squared prediction error than the quadratic model whenever $\sum b_i^2 + 2\sigma^2 < 3\sigma^2$, or in other words whenever $\sum b_i^2 < \sigma^2$, and this quickly reduces to the form given in the question.