Other Tests of Model Selection. The *J* test just discussed is only one of a group of tests of model selection. There is the **Cox test,** the **JA test,** the **P test, Mizon–Richard encompassing test,** and variants of these tests. Obviously, we cannot hope to discuss these specialized tests, for which the reader may want to consult the references cited in the various footnotes.[37]

13.9 MODEL SELECTION CRITERIA

In this section we discuss several criteria that have been used to choose among competing models and/or to compare models for forecasting purposes. Here we distinguish between **in-sample** forecasting and **out-of-sample** forecasting. In-sample forecasting essentially tells us how the chosen model fits the data in a given sample. Out-of-sample forecasting is concerned with determining how a fitted model forecasts future values of the regressand, given the values of the regressors.

Several criteria are used for this purpose. In particular, we discuss these criteria: (1) R^2, (2) adjusted $R^2 (= \bar{R}^2)$, (3) Akaike information criterion (AIC), (4) Schwarz Information criterion (SIC), (5) Mallow's C_p criterion, and (6) forecast χ^2 (chi-square). All these criteria aim at minimizing the residual sum of squares (RRS) (or increasing the R^2 value). However, except for the first criterion, criteria (2), (3), (4), and (5) impose a penalty for including an increasingly large number of regressors. Thus there is a tradeoff between goodness of fit of the model and its complexity (as judged by the number of regressors).

The R^2 Criterion

We know that one of the measures of goodness of fit of a regression model is R^2, which, as we know, is defined as:

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \tag{13.9.1}$$

R^2, thus defined, of necessity lies between 0 and 1. The closer it is to 1, the better is the fit. But there are problems with R^2. *First*, it measures *in-sample* goodness of fit in the sense of how close an estimated Y value is to its actual value in the given sample. There is no guarantee that it will forecast well *out-of-sample* observations. *Second,* in comparing two or more R^2's, the dependent variable, or regressand, must be the same. *Third,* and more importantly, an R^2 cannot fall when more variables are added to the model. Therefore, there is every temptation to play the game of "maximizing the R^2" by simply adding more variables to the model. Of course, adding more variables to the model may increase R^2 but it may also increase the variance of forecast error.

[37] See also Badi H. Baltagi, *Econometrics,* Springer, New York, 1998, pp. 209–222.

Adjusted R^2

As a penalty for adding regressors to increase the R^2 value, Henry Theil developed the adjusted R^2, denoted by $\bar{R}^2$, which we studied in Chapter 7. Recall that

$$\bar{R}^2 = 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-k} \tag{13.9.2}$$

As you can see from this formula, $\bar{R}^2 \le R^2$, showing how the adjusted R^2 penalizes for adding more regressors. As we noted in Chapter 8, unlike R^2, the adjusted R^2 will increase only if the absolute t value of the added variable is greater than 1. For comparative purposes, therefore, $\bar{R}^2$ is a better measure than R^2. But again keep in mind that the regressand must be the same for the comparison to be valid.

Akaike Information Criterion (AIC)

The idea of imposing a penalty for adding regressors to the model has been carried further in the AIC criterion, which is defined as:

$$\text{AIC} = e^{2k/n}\frac{\sum \hat{u}_i^2}{n} = e^{2k/n}\frac{\text{RSS}}{\text{n}} \tag{13.9.3}$$

where k is the number of regressors (including the intercept) and n is the number of observations. For mathematical convenience, (13.9.3) is written as

$$\ln \text{AIC} = \left(\frac{2k}{n}\right) + \ln\left(\frac{\text{RSS}}{n}\right) \tag{13.9.4}$$

where $\ln \text{AIC} =$ natural log of AIC and $2k/n =$ penalty factor. Some textbooks and software packages define AIC only in terms of its log transform so there is no need to put ln before AIC. As you see from this formula, AIC imposes a harsher penalty than $\bar{R}^2$ for adding more regressors. In comparing two or more models, the model with the lowest value of AIC is preferred. One advantage of AIC is that it is useful for not only in-sample but also out-of-sample forecasting performance of a regression model. Also, it is useful for both nested and non-nested models. It has been also used to determine the lag length in an AR(p) model.

Schwarz Information Criterion (SIC)

Similar in spirit to the AIC, the SIC criterion is defined as:

$$\text{SIC} = n^{k/n}\frac{\sum \hat{u}^2}{n} = n^{k/n}\frac{\text{RSS}}{n} \tag{13.9.5}$$

or in log-form:

$$\ln \mathrm{SIC} = \frac{k}{n} \ln n + \ln \left(\frac{\mathrm{RSS}}{\mathrm{n}} \right) \tag{13.9.6}$$

where $[(k/n) \ln n]$ is the penalty factor. SIC imposes a harsher penalty than AIC, as is obvious from comparing (13.9.6) to (13.9.4). Like AIC, the lower the value of SIC, the better the model. Again, like AIC, SIC can be used to compare in-sample or out-of-sample forecasting performance of a model.

Mallows's C_p Criterion

Suppose we have a model consisting of k regressors, including the intercept. Let $\hat{\sigma}^2$ as usual be the estimator of the true σ^2. But suppose that we only choose p regressors ($p \leq k$) and obtain the RSS from the regression using these p regressors. Let RSS_p denote the residual sum of squares using the p regressors. Now C. P. Mallows has developed the following criterion for model selection, known as the C_p criterion:

$$C_p = \frac{\mathrm{RSS}_p}{\hat{\sigma}^2} - (n - 2p) \tag{13.9.7}$$

where n is the number of observations.

We know that $E(\hat{\sigma}^2)$ is an unbiased estimator of the true σ^2. Now, if the model with p regressors is adequate in that it does not suffer from lack of fit, it can be shown[38] that $E(\mathrm{RSS}_p) = (n - p)\sigma^2$. In consequence, it is true *approximately* that

$$E(C_p) \approx \frac{(n - p)\sigma^2}{\sigma^2} - (n - 2p) \approx p \tag{13.9.8}$$

In choosing a model according to the C_p criterion, we would look for a model that has a low C_p value, about equal to p. In other words, following the principle of parsimony, we will choose a model with p regressors ($p < k$) that gives a fairly good fit to the data.

In practice, one usually plots C_p computed from (13.9.7) against p. An "adequate" model will show up as a point close to the $C_p = p$ line, as can be seen from Figure 13.3. As this figure shows, Model A may be preferable to Model B, as it is closer to the $C_p = p$ line than Model B.

A Word of Caution about Model Selection Criteria

We have discussed several model selection criteria. But one should look at these criteria as an adjunct to the various specification tests we have

[38] Norman D. Draper and Harry Smith, *Applied Regression Analysis,* 3d ed., John Wiley & Sons, New York, 1998, p. 332. See this book for some worked examples of C_p.

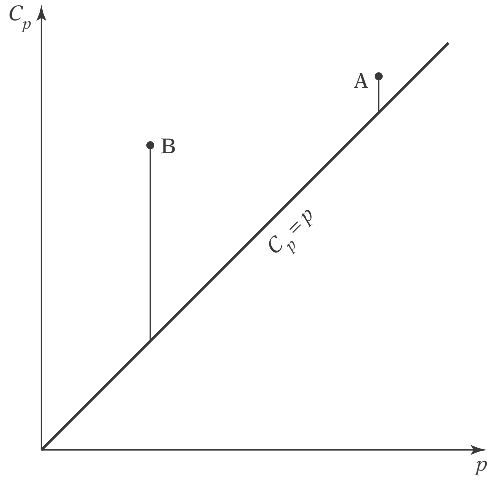

FIGURE 13.3 Mallows's C_p plot.

discussed in this chapter. Some of the criteria discussed above are purely descriptive and may not have strong theoretical properties. Some of them may even be open to the charge of data mining. Nonetheless, they are so frequently used by the practitioner that the reader should be aware of them. No one of these criteria is necessarily superior to the others.[39] Most modern software packages now include R^2, adjusted R^2, AIC, and SIC. Mallows's C_p is not routinely given, although it can be easily computed from its definition.

Forecast Chi-Square (χ^2)

Suppose we have a regression model based on n observations and suppose we want to use it to forecast the (mean) values of the regressand for an additional t observations. As noted elsewhere, it is a good idea to save part of the sample data to see how the estimated model forecasts the observations not included in the sample, the postsample period.

Now the forecast χ^2 test is defined as follows:

$$\text{Forecast, } \chi^2 = \frac{\sum_{n+1}^{n+t} \hat{u}_i^2}{\hat{\sigma}^2} \qquad \textbf{(13.9.9)}$$

where $\hat{u}_i$ is the forecast error made for period $i\ (= n+1, n+2, \ldots, +n+t)$, using the parameters obtained from the fitted regression and the values of the regressors in the postsample period. $\hat{\sigma}^2$ is the usual OLS estimator of σ^2 based on the fitted regression.

[39] For a useful discussion on this topic, see Francis X. Diebold, *Elements of Forecasting*, 2d ed., South Western Publishing, 2001, pp. 83–89. On balance, Diebold recommends the SIC criterion.

If we hypothesize that the parameter values have not changed between the sample and postsample periods, it can be shown that the statistic given in (13.9.9) follows the chi-square distribution with t degrees of freedom, where t is the number of periods for which the forecast is made. As Charemza and Deadman note, the forecast χ^2 test has *weak statistical power*, meaning that the probability that the test will correctly reject a false null hypothesis is low and therefore the test should be used as a signal rather than a definitive test.[40]

13.10 ADDITIONAL TOPICS IN ECONOMETRIC MODELING

As noted in the introduction to this chapter, the topic of econometric modeling and diagnostic testing is so vast and evolving that specialized books are written on this topic. In the previous section we have touched on some major themes in this area. In this section we consider a few additional features that researchers may find useful in practice. In particular, we consider these topics: (1) **outliers, leverage, and influence;** (2) **recursive least squares,** and (3) **Chow's prediction failure test.** Of necessity the discussion of each of these topics will be brief.

Outliers, Leverage, and Influence[41]

Recall that, in minimizing the residual sum of squares (RSS), OLS gives equal weight to every observation in the sample. But every observation may not have equal impact on the regression results because of the presence of three types of special data points called **outliers, leverage points,** and **influence points.** It is important that we know what they are and how they influence regression analysis.

In the regression context, an **outlier** may be defined as an observation with a "large residual." Recall that $\hat{u}_i = (Y_i - \hat{Y}_i)$, that is, the residual represents the difference (positive or negative) between the actual value of the regressand and its value estimated from the regression model. When we say that a residual is large, it is in comparison with the other residuals and very often such a large residual catches our attention immediately because of its rather large vertical distance from the estimated regression line. Note that in a data set there may be more than one outlier. We have already encountered an example of this in exercise 11.22, where you were asked to regress percent change in stock prices (Y) on percent change in consumer prices (X) for a sample of 20 countries. One observation, that relating to Chile, was an outlier.

[40] Wojciech W. Charemza and Derek F. Deadman, *New Directions in Econometric Practice: A General to Specific Modelling, Cointegration and Vector Autoregression,* 2d ed., Edward Elgar Publishers, 1997, p. 30. See also pp. 250–252 for their views on various model selection criteria.

[41] The following discussion is influenced by Chandan Mukherjee, Howard White, and Marc Wyuts, *Econometrics and Data Analysis for Developing Countries,* Routledge, New York, 1998, pp. 137–148.