LONG-RANGE FORECASTING
From Crystal Ball to Computer

Thirteen

TESTING OUTPUTS

Contents

> Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.
>
> Tukey (1962)

Tests on the outputs of a model can be used to gain further insight about the inputs to a model so that improvements can be made. For this purpose, conditional forecasts are useful. The first section in this chapter discusses the use of conditional forecasts.

The major reason for testing outputs, however, is to make comparisons among a set of models. Cost-benefit analyses of the outputs should be made for each model in order to select the best one. These tests should be carried out in situations analogous to the one encountered in the forecasting problem. Much of this chapter is devoted to the choice of an appropriate testing situation.

Accuracy is of major concern. How does one compare the accuracies of alternative models? Measures of statistical and practical significance are examined. The popularity of various accuracy measures is examined and typical forecast errors are presented.

CONDITIONAL VS. UNCONDITIONAL FORECASTS

The unconditional forecast uses only information available at the time of the actual forecast. Thus, if the number of riots in the United States from 1986 to 1991 were being forecast, no information from a time later than 1985 would be used. Of course, this is what we normally mean when we talk about forecasting.

Conditional forecasts can be conditional in different respects. One can have information from the forecast situation that pertains to the *relationships* or to the *variables*. Also, one can have actual ("known") data for the situation, or it can be unknown. The various possibilities are presented in Exhibit 13-1 along with their traditional names.

The *ex ante* forecast provides the best measure of a model's ability to forecast and offers a benchmark for comparison. This chapter is concerned primarily with *ex ante* forecasts.

Ex post forecasts can be used to assess how well the model would do if the best possible forecasts were made of the causal variables. A comparison between the *ex post* and the *ex ante* forecast is useful in deciding whether further work should be done to improve the forecasts

Exhibit 13-1 TYPES OF CONDITIONAL AND UNCONDITIONAL FORECASTS

	Relationships	
Variables	Unknown	Known
Unknown	*Ex Ante*	Modified *Ex Ante*
Known	*Ex Post*	Calibration

of the causal variables. More detailed analyses could be made by using perfect forecasts for each of the causal variables, one at a time, to identify which variables are of greatest concern.

In the modified *ex ante* forecast, data from the forecast horizon are used to estimate the coefficients in the forecasting model. Then you go back to use the forecasts of the causal variables. A comparison can be made with the *ex ante* forecasts to assess the value of having the improved estimates of the relationships. You could use known data for only one of the relationships to test the sensitivity of the outputs.

Finally, **calibration**[G] forecasts can be examined to assess the net impact of better estimates of relationships and better forecasts of the causal variables.

These ideas on the use of conditional forecasts are relevant primarily for econometric and segmentation models. For judgmental models, it may be possible to create these tests by controlling the information that is given to the judges.

SITUATIONS FOR TESTING OUTPUTS

The testing situation should be as similar as possible to the forecasting situation. There should be a correspondence with respect to space, population, and time. For an important application of this principle, see the use of "work samples" for personnel selection [e.g., ROBERTSON and DOWNS, 1979; SMITH, 1976].

Space correspondence is the extent to which the environment of the test data corresponds to the environment of the forecasting problem. Obviously, the environment may have a large impact on the forecasting model. For example, Newton's laws do not apply throughout the universe, but they work well for the types of problems that we encounter in the earth's environment.

Many years ago, I was involved in a project to forecast the health of certain very important people (VIPs) around the world. Most of them would not have been interested in cooperating, and I am sure the C.I.A. did not bother to ask them. (Oh, they never told me who paid for the study. However, *Parade* magazine said that it was the C.I.A.) So, instead, we used executives in Minneapolis. Now Minneapolis is cold, but it isn't Moscow.

With respect to population, it is desirable to obtain data on the decision units that are the subject of the forecasting problem. Sometimes, however, it is necessary to compromise and to select analogous decision units. That is what we did in our VIP forecasting project. We were not really interested in those Minneapolis executives.

Generally, the selection of a relevant population is straightforward. One practice, however, has led to serious errors in assessing forecast validity. This is the selection of extreme groups from a population. Predictions made about these extreme groups lead to a phenomenon called regression toward the mean.

Regression toward the mean[G] occurs because, although extreme groups contain observations that differ significantly from the mean, they also contain observations that are due to errors in measurement. When later measurements are made, it is unlikely that the given observation will be subject to the same large error in measurement. Thus the mean of the extreme group regresses toward the population mean.

One well-known case of regression toward the mean has occurred in speed-reading courses. A reading test is given to a group of people. The slow readers then take the speed-reading course, after which they are retested. The forecast by the speed-reading advocates is upheld; they have been able to increase the speed of these slow readers. Of course, the measured speed would have been higher on the second test even if no speed-reading course had been given because the slow reader group would have included people who were not feeling good on the first test or misunderstood the directions or were tired or got bogged down by the selected passages to be read. Many of these transient factors may not occur with the same impact on the second testing. (Incidentally, according to Carver (1972), there is little evidence that speed-reading courses increase speed without reducing comprehension. Anecdotal evidence is provided by Woody Allen: he took a speed-reading course and was able to read *War and Peace* in 20 minutes. "It's about Russia," said Woody.)

I hope that none of you are speed-readers; but if you are, you can check your comprehension so far. Select the best answer: *Long-Range Forecasting* is about . . .

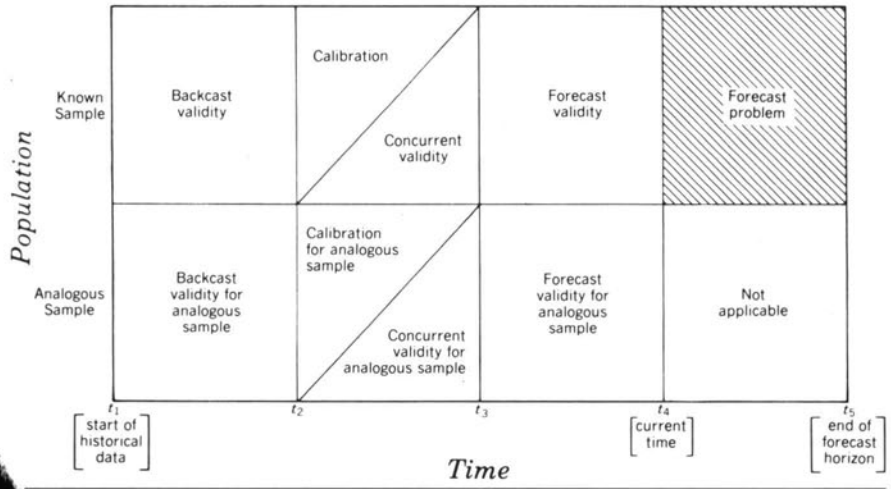
(a) The origin of Tonto's friend's name.
(b) The history of amateur fishing.
(c) The use and abuse of forecasting methods.
(d) Helpful hints on getting through life.

The correct answer is given in Appendix F.

The critical aspect of the test situation is generally the selection of a suitable time period. What time correspondence will allow for the best assessment of forecast validity?

To assist in the selection of appropriate test data, a "validation matrix" is presented in Exhibit 13-2. This matrix considers forecasts in cases where data on the forecast sample are available (the "known sample"); the data differ with respect to space (same population, but new environment); population (same space, but new population); both space and population. The last three categories are grouped into the "analogous sample" in Exhibit 13-2. Each type of data is examined for three time periods. Data from the time period immediately preceding the current time can be withheld to test forecast validity. The time period preceding that one can be used to calibrate the model and also to test for concurrent validity. The time period preceding that one can be used for a test of backcast validity. The time period after the current time represents the typical forecasting problem (shaded area).

Exhibit 13-2 THE VALIDATION MATRIX



As can be seen from Exhibit 13-2, eight situations can be used to test forecast validity. Six situations allow for testing unconditional forecasts (all but the calibration samples). If one were to include conditional forecasts (from Exhibit 13-1), there would be many additional possible tests.

After describing the use of the calibration sample, I discuss concurrent, forecast, and backcast validity. The relative advantages and disadvantages of each test are examined.

Calibration

The data used to calibrate (estimate) the model have frequently been used to test forecast validity. The fit to the calibration data is used to draw inferences about forecast validity. Although tests on the calibration data are easily available, inexpensive, and popular, this approach is the least useful of the tests listed in the validation matrix. Fortunately, it is seldom necessary to use the calibration data because of the availability of the other tests in Exhibit 13-2.

A good fit between the model and the calibration sample may be the result of chance. This is especially likely for complex models. Furthermore, the researcher inevitably finds ways (he thinks) to improve the model, and these typically provide a better fit. The net result is that the fit between the model and the calibration data provides only a crude estimate of forecast validity. In some cases it may be considered as an upper limit to the predictive power of the model.

The relationship between the ability of a model to explain the variation in the calibration sample and its forecast validity has been studied for almost half a century. One of these studies, Williams and Goodman (1971), found that the fit to the calibration sample provided a fair estimate of predictive ability for short-range forecasts with an extrapolation model. However, the following studies, found that ability to fit the calibration data was a poor measure of forecast validity:

> The U.S. National Resources Committee (1938) used econometric models to forecast changes in employment and consumption in 81 sectors of the economy for 1933–1936. Data from 1918 to 1932 were used to calibrate the model. A plot of the standard error of the regressions for 1918–1932 (the calibration sample) showed no relationship to the MAPEs for the 138 forecasts over the forecast horizon from 1933 to 1936. Thus the fit to the calibration data provided a poor estimate of forecast accuracy.

Ohlin and Duncan (1949) reviewed studies that attempted to forecast which prisoners could be successfully released on parole. In all six studies that allowed for a comparison, a causal model was superior to a naive extrapolation when the comparisons were made on the calibration data (as they had been in the original publications of the six studies). When these models were used to predict outside of the calibration data, the causal model was superior to the extrapolation in two cases, it was inferior in three cases, and there was one tie. Overall, the causal model was not more accurate.

Ferber (1956) examined forecasts of the total savings in the economy. Seven different models were calibrated from data for 1923–1940. Forecasts were then made for 1947–1949. There was only a small relationship between the R^2 for the calibration data and the forecast error.

Schupack (1962), in a study involving short-range sales forecasts for food and household products, found only a slight positive relationship between the fit of regression models to the calibration sample and their forecast accuracy.

In ELLIOTT and BAIER [1979], six econometric models provided good explanations of the changes in interest rates for a calibration sample. The R^2 for these models ranged from .858 to .996. However, for one-month ahead forecasts, these models were inferior to a no-change model.

Concurrent Validity

Locke (1961) asked 29 subjects to describe themselves using 81 adjectives. The data were analyzed by comparing subjects with long last names and those with short last names (six or fewer letters). Eighteen of the adjectives were significant in distinguishing between those with long and short names. In fact, the discriminant analysis distinguished almost perfectly between groups. Of course, when Locke cross-validated

this model on a hold-out sample of 30 subjects, there was no forecast validity. Einhorn (1972b) presented a similar illustration using random data in a segmentation analysis.

There is a moral: predictive studies often fail to go beyond the calibration sample. Two examples provided here show the test of concurrent validity to be useful:

Kurtz (1948) reexamined a study that used the **Rorschach test**[G] to predict the success of life insurance sales managers. The use of 32 variables from this test led to almost perfect predictions for the 42 successful and 38 unsuccessful sales managers in the calibration sample. But when Kurtz tested concurrent validity, the Rorschach test was completely lacking in predictive power. (Interestingly, Kurtz reported that some advocates of the Rorschach test refused to accept this result as a failure.)

Frank, Massy, and Morrison (1965) reanalyzed an earlier study by Frank and Massy (1963) which had concluded that adopters of a new brand of coffee, Folger's, could be predicted from socioeconomic and purchasing data. This study had correctly classified about 73% of the purchasers in the calibration sample, but when the test was cross-validated only 48% were correctly classified. This was no better than chance. (In contrast to the preceding study, Frank, Massy, and Morrison were able to accept their negative result. But, of course, they found the error by themselves.)

The test of concurrent validity is more important to the extent that heavy reliance is placed on the calibration sample. Alternatively, if the calibration model is not used to a great extent, there is less need for a test of concurrent validity. In the extreme, when prior theory is used to calibrate the model, there is no calibration test, only a test of concurrent validity. The following is an example:

Armstrong and Overton (1977) used five models to predict the effect of nonresponse bias in mail survey results. Each model was calibrated *a priori*. For example, one model assumed that the effect of nonresponse would be zero. Another assumed that the

average response in the latest wave of the questionnaire would provide the best prediction as to how nonrespondents would have answered. The concurrent validity of each of the five models was then tested on actual data.

It is seldom possible to test concurrent validity with time series data. If data are available for only one unit of observation (or if the data are not available on a disaggregate basis), concurrent validity cannot be tested.

A related problem in the use of concurrent validity arises from the use of small samples; that is, few data may be available for both calibrating the model and testing concurrent validity. There are solutions to this problem. The simplest one is to use **double cross validation**[G]. For this, one first examines concurrent validity. The names of the two subsamples are then reversed, and the process is repeated. In other words, the calibration is done on what was originally the concurrent validity subsample, and concurrent validity is tested on the original calibration sample. This procedure yields two estimates of concurrent validity. Two estimates are better than one. Right?

If sample size is very small, it is possible to use N-way cross validation. Here, all except one of the observations are used to calibrate the model. A forecast is made for the remaining observation. Next, this observation is replaced, and again, all the data but one are used to calibrate a new model; this model is then used to forecast for the one excluded observation. This process is repeated until each of the observations has been used for validation. (A discussion on N-way cross validation can be found in Frank, Massy, and Morrison (1965).) This procedure is also called the jackknife [MOSTELLER and TUKEY, 1977, p. 135].

If there are no concurrent data, you might think all is lost. But there is hope. One strategy is to create random data and to repeat the same procedures on these data that were used on the calibration sample. Comparisons can then be made between the fit to the calibration sample and the fit to the random data. For a rough idea of what will happen, Ando and Kaufman (1966) provided tables obtained by analyzing random data for sample sizes from 10 to 200, and using from 10 to 50 causal variables. It is better, however, to replicate the process than to use the tables. Examples of replication with random data are provided by Payne and Dyer (1975) and Armstrong (1975a).

An improvement of the random data validation was suggested by Montgomery (1975). Rather than using completely random data, you

can randomize the data for the dependent variable only; that is, the values for the dependent variable are removed and then randomly reassigned to the observations. Again, the process used to calibrate the model is applied to these randomized data. This procedure has some appeal because the distribution for the dependent variable is retained.

A checklist for tests of concurrent validity is provided in Exhibit 13-3. This exhibit indicates when each procedure is appropriate and cites a reference describing an early application of the procedure.

Exhibit 13-3 TESTING CONCURRENT VALIDITY

Sample Size	Recommended Procedure	Early Applications
Large	Cross validation	Minor (1958)
Moderate	Double cross validation	Roach (1971)
Small	*N*-way cross validation	Wiseman (1972b)
None	Random data validation	Montgomery (1975)

The test of concurrent validity is a more appropriate test of forecast validity than is the fit to the calibration sample. The value of such a test is especially great when complex models are used. For example, Ward (1954) found that the **shrinkage**[G] from calibration to concurrent data was much greater for complex regression models. Concurrent validity is also useful when the calibration sample has been used to estimate a number of alternative models. Although the calibration sample is adequate in some situations [CATTIN, 1980; BARRETT, PHILLIPS, and ALEXANDER, 1981], it is safer to test concurrent validity.

This discussion assumed that data were available on the relevant population. An alternative is to use an analogous sample, as we did in the previously cited study on the health of VIPs. The model was calibrated on this group, and concurrent validity was also tested.

Forecast Validity

To obtain the best measure of forecast validity, one could use the forecasting models and monitor their success. A shortcut is to withhold data from a recent time period and to pretend that the forecasts are being made. This test is almost as good as the real thing, although it is possible that the model may be contaminated by the researcher's knowledge of the time period that is being withheld.

Preferably, the test of forecast validity is run on the decision units

that are of interest. An unconditional test of forecast validity on these units would provide the ideal test of forecast validity for the problem.

Small samples cause problems in obtaining a good test of forecast validity. This is a common difficulty with time series data. One procedure that helps to make the most effective use of limited time series data is called **successive updating**[G]. This involves the removal of data from the validation sample and their addition to the calibration sample after each forecast.

Successive updating is illustrated in Exhibit 13-4. In period t, forecasts are made over the horizon from $t + 1$ to $t + h$. In period $t + 1$, data from period $t + 1$ are used in the calibration sample and forecasts are made for periods $t + 2$ to $t + h$. This procedure is repeated until there is only a one-period forecast. The one-period-ahead forecasts are designated in Exhibit 13-4 by F_1, the two-period-ahead forecasts by F_2, and so on. This yields h of the one-period forecasts, $h - 1$ of the two-period forecasts, and so on down to one of the h-period-ahead forecasts. Examples of this procedure can be found in Armstrong and Grohman (1972), Williams and Goodman (1971), MABERT [1976], and SCHNAARS [1984]. This procedure, with its different starting points, provides a good test of validity.

Exhibit 13-4 SUCCESSIVE UPDATING

Calibration Data Ends With:	Forecast Periods					
	$t + 1$	$t + 2$	$t + 3$	$t + 4$	$\cdots$	$t + h$
t	F_1	F_2	F_3	F_4	$\cdots$	F_h
$t + 1$	—	F_1	F_2	F_3	$\cdots$	F_{h-1}
$t + 2$	—	—	F_1	F_2	$\cdots$	F_{h-2}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$t + h - 1$	—	—	—	—		F_1

Backcast Validity

> The White Queen lives backward through time. She begins to cry before she sticks herself with her broach and stops immediately afterward. Living backward in time, she explains to Alice, "always makes one a little giddy at first . . . but there's one great advantage in it—that one's memory works both ways."
>
> Lewis Carroll
> *Through the Looking Glass*

It is sometimes difficult to obtain data for forecast validation. Often, one plans to use a causal model but finds that historical data on the *causal* variables do not exist, although current data are available. The use of backcasting is suggested here; this involves withholding data from some time in the distant past and using only the most recent data to develop the forecasting models. You then backcast what happened in the earlier period. Obviously, this approach may suffer from contamination because the researcher may be affected by the knowledge of what has happened. An example of backcasting is provided in financial forecasting by SMITH and BRAINARD [1976]. Another example is described here:

> In Armstrong (1968b), data from 1965 to 1960 were used to calibrate models to forecast camera sales by country. These models were then used to make six-year backcasts for camera sales in 1954. The accuracy of these backcasts served to assess how well the models would do in a six-year forecast.

The White Queen was right: backcasting does make people a little giddy at first. Some researchers are skeptical of backcasting as a measure of forecast validity. In my opinion, backcast validity provides a useful approach to the assessment of forecast validity. The only empirical evidence that I found is from Theil, who found a close correspondence between conditional forecasts and conditional backcasts:

> Theil (1966) examined input-output forecasts for two industries. He compared **mean square errors**[G] for conditional forecasts and backcasts. I converted these to root mean square errors for ease of interpretation. The correspondence was close:

Forecast Horizon (years)	Mean Square Errors			
	Agriculture, Forestry, Fishing		Basic Metal Industries	
	Backcasts	Forecasts	Backcasts	Forecasts
1	4.0	4.0	10.6	10.4
2	6.0	6.2	15.7	15.4
3	7.9	7.9	20.1	21.5
4	9.9	9.5	20.7	25.7
5.5	12.0	11.6	29.9	32.2
8	12.5	11.0	39.8	48.2

Backcasting is seldom used in published studies. That is unfortunate. It can add substantially to the assessment of predictive validity.

Ranking the Validation Tests

Some tests are more representative of forecast validity than others. I have ranked these tests in Exhibit 13-5. These rankings were drawn primarily from Gerstenfeld's law of trying because I could not find much empirical evidence. The rankings might be expected to vary, depending on the situation. Nevertheless, two things stand out from these rankings: first, the most popular method, the fit to the calibration data, is least valuable; and second, tests of backcast validity are highly rated, although they are not popular.

When more than one test is used, consideration must be given to the time sequence of the testing. For example, tests of concurrent validity should be carried out first because these data can then be incorporated into the calibration data to update the models. Similarly, tests of backcasting should be done before tests of forecasting so that the models can be recalibrated with the additional data included. The general principle is to save the most appropriate test for last. These ideas on time sequencing are summarized in Exhibit 13-5. Use more than one of the tests. One gains confidence in a model if it performs well in a variety of tests.

Exhibit 13-5 RANKING THE VALIDATION TESTS

Value (1 = most useful)	Time Sequence (1 = do first)	Test
1	4	Forecast validity
2	3	Backcast validity
3	4	Concurrent validity
4	3	Forecast validity with analogous data
5	2	Backcast validity with analogous data
6	2	Concurrent validity with analogous data
7	1	Calibration fit
8	1	Calibration fit with analogous data

TESTING ACCURACY

Accuracy is one of the most important criteria for a forecasting model. Measures of accuracy can help to assess costs and benefits in a situation.

A single measure of accuracy sometimes falls short. Is it better to forecast too high or too low? Do changes in direction have particular importance? Is it worse to err by forecasting that something will happen when it does not happen, or by forecasting that it will not happen when it does happen? It is difficult to develop generalizations on these points because they depend on the specific situation being examined.

This section describes some measures of forecast accuracy and discusses their advantages and disadvantages. You should be able to find some measures that will fit your problem.

Measures of Accuracy

This section of the book is not the most interesting. If you are not looking for a measure of accuracy, you may prefer to skip this section and go to the one on statistical and practical significance (p. 356). A review of accuracy measures can be found in Exhibit 13-7 (*LRF* p. 355).

A consistent notation will be followed. It is noted in the symbol list in the glossary and, for your convenience, is summarized here:

A	is the actual result
F	is the forecast
t	is the time interval
h	is the number of periods in the forecast horizon

For cross-sectional data, t represents the element to be forecast, and h is the number of elements. Unless otherwise noted, A and F are assumed to be equal to or greater than 0, and the data are ratio scaled (i.e., the intervals are meaningful, and there is a meaningful zero point).

1. **Mean error (ME) is calculated from**

$$\text{ME} = \frac{\sum_{t=1}^{h} (A_t - F_t)}{h}$$

The mean error is primarily a test of systematic error (bias). It assumes that the cost of the errors is symmetrical. It should not be used by itself because it provides no measure of the error variance.

Notice that you could have a ME = 0 when large errors are present if the large errors had different signs and canceled one another. For this measure, F and A can take on negative values. Interval-scaled data are sufficient.

2. Mean absolute deviation (MAD) is calculated from

$$\text{MAD} = \frac{\sum_{t=1}^{h} (\,|\, A_t - F_t \,|\,)}{h}$$

The MAD reflects the typical error. It does not distinguish between variance and bias. It is appropriate when the cost function is linear (e.g., when the cost of an error of 10 units is twice that of an error of five units).

3. Root mean square error (RMSE) is calculated from

$$\text{RMSE} = \left[\frac{\sum_{t=1}^{h} (A_t - F_t)^2}{h} \right]^{1/2}$$

The RMSE is similar to the MAD; for unbiased errors, you can get a *crude* approximation of the RMSE from the MAD as follows (see Brown, 1963, pp. 282–283).

$$\text{RMSE} = 1.25\ \text{MAD}$$

(In practice, this ratio often fluctuates by a significant amount.) Why would someone want to calculate the RMSE? For one thing, it is nice when you want to impress people (some forecasters prefer a complex measure to a simple one). Another reason for using the RMSE arises when there is a quadratic loss function; in other words, when the cost associated with an error increases as the square of the error. For instance, maybe the degree to which your client gets upset at bad forecasts increases with the square of the error? You should avoid the RMSE when the assessment of error is so crude that there are outliers (large measurement errors). These outliers will have a strong effect on the measure when they are squared.

4. Mean absolute percentage error (MAPE) is calculated from

$$\text{MAPE} = \left[\frac{\sum_{t=1}^{h} \left[\dfrac{|\, A_t - F_t \,|}{A_t} \right]}{h} \right] \cdot [100]$$

The MAPE is similar to the MAD except that it is dimensionless. This makes it nice for communication and helpful in making comparisons among forecasts from different situations. For example, to compare forecasting methods in two different situations with different units of measure, one can calculate the MAPEs and then average across situations.

When the cost of errors is more closely related to the percentage error than to the unit error, the MAPE is appropriate. This assumption is often reasonable. Problems may arise if values close to 0 are encountered for actual results. Finally, it helps if the actual results can be measured accurately.

The MAPE has a bias favoring estimates that are below the actual values. This can be seen by looking at the extremes: a forecast of 0 can never be off by more than 100%, but there is no limit to errors on the high side.

The *median* absolute percentage error might be preferred in situations where larger errors are not so costly. This measure is also useful in comparing models [as shown in MAKRIDAKIS, et al. 1982].

5. Adjusted mean absolute percentage error ($\overline{\text{MAPE}}$) is calculated from

$$\overline{\text{MAPE}} = \left[\frac{\sum_{t=1}^{h} \frac{|A_t - F_t|}{\frac{1}{2}(A_t + F_t)}}{h} \right] \cdot [100]$$

The $\overline{\text{MAPE}}$ is similar to the MAPE, but it does not favor low estimates. If the actual value is 100, a forecast of 50 is as good (or as bad) as a forecast of 200. At the extremes, the $\overline{\text{MAPE}}$ goes from 0 for a perfect forecast to 200 for an infinitely bad forecast. This is advantageous at times; for example, when working with judgmental forecasters, it will help to prevent an intentional biasing of the forecast. Another advantage is that the $\overline{\text{MAPE}}$ is less sensitive than the MAPE to measurement errors in the actual data.

The disadvantages are that the $\overline{\text{MAPE}}$ is more difficult to understand and more difficult to relate to decision making than is the MAPE. Thus, if the problems cited in the preceding paragraph are not serious, you should use the MAPE.

6. There are two measures of Theil's inequality (U). They are calculated as follows:

$$U1 = \frac{[(1/h)\sum_{t=1}^{h}(A_t - F_t)^2]^{1/2}}{[(1/h)\sum_{t=1}^{h}A_t^2]^{1/2} + [(1/h)\sum_{t=1}^{h}F_t^2]^{1/2}}$$

$$U2 = \frac{[(1/h)\sum_{t=1}^{h}(\Delta F_t - \Delta A_t)^2]^{1/2}}{[(1/h)\sum_{t=1}^{h}(\Delta A_t^2)]^{1/2}}$$

where Δ refers to changes.

There is some confusion about these measures because Theil proposed both, but at different times and under the same symbol. $U1$ is taken from Theil (1958, pp. 31–42), where he calls it a measure of forecast accuracy; it is bounded between 0 and 1, with 0 being a perfect forecast. $U2$ is from Theil (1966, Chapter 2), where he refers to it as a measure of forecast quality; for this, 0 represents a perfect forecast, 1 is the no-change forecast, and there is no upper limit. $U2$ can be interpreted as the RMSE of the proposed forecasting model divided by the RMSE of a no-change model.

Bliemel (1973) analyzed Theil's measures of inequality. He concluded that $U1$ was not very enlightening whether one is dealing with absolute values or with changes. For example, when changes are examined, *all* forecasting models will do better than a naive model that predicts no change! Yet Theil himself used $U1$ for change data. $U1$ is also difficult to understand and to relate to decision making. $U1$ can be viewed as an historical oddity; it should not be used for predicting either changes or absolute values.

$U2$ seems reasonable and is easier to interpret because the $U2 = 1.0$ benchmark is based on a no-change model. Values for $U2$ of less than 1 represent improvements over the no-change model. Furthermore, $U2$ has no serious defects. However, $U2$ does not pick up errors in forecasting levels but measures only errors in changes. Despite the advantages of $U2$ over $U1$, Bliemel says that $U1$ has been more widely advocated.

The confusion between $U1$ and $U2$ implies that these measures are not well understood. Imagine how much difficulty the *clients* will have with them. Fortunately, as will be noted in the summary, there is little need for Theil's U; other measures can do the job and do it more simply.

7. Coefficient of variation (CV) is calculated from

$$\text{CV} = \frac{\text{RMSE}}{[\sum_{t=1}^{h}A/h]}$$

The CV relates the root mean square error to the average value of the actual data. This unit-free measure is similar to the MAPE, but the MAPE is easier to interpret.

8. Coefficient of determination (R^2) is calculated from

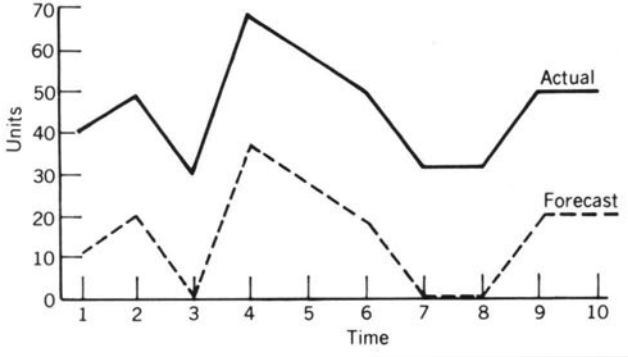
$$R^2 = \frac{\sum_{t=1}^{h} [(F_t - \overline{F})\,(A_t - \overline{A})]^2}{[\sum_{t=1}^{h} (F_t - \overline{F})^2]\,[\sum_{t=1}^{h} (A_t - \overline{A})^2]}$$

where $\overline{A}$ is the average A_t and $\overline{F}$ is the average F_t.

Although R^2 can be interpreted as the "proportion of variance explained," it is not an easy concept to explain to clients, nor is it easy to translate into decision-oriented terms. For example, an R^2 of 0 means that the model is not useful in explaining fluctuations, but it does not reflect its ability to forecast levels; thus, $R^2 = 0$ may not be a completely bad forecast. Similarly an R^2 of 1.0 does not represent a perfectly good forecast, as shown in Exhibit 13-6. Also, R^2 depends not only on the fit of the data but also on the steepness of the regression line (Barrett, 1974), implying that R^2 will be higher when changes are greater.

R^2 should not be used for the calibration sample. Instead, you should use $\overline{R}^2$ (the adjusted R^2). This adjustment allows for the fact that the regression model capitalizes on chance in finding the best fit to the calibration sample. In other words, $\overline{R}^2$ compensates for the loss in **degrees of freedom**[G] in fitting the model. (In the extreme, you can get a perfect R^2 by using a large enough number of independent variables.)

Exhibit 13-6 PERFECT R^2 DOES NOT IMPLY PERFECT FORECAST



At least three methods have been proposed for calculating $\bar{R}^2$. These were summarized by Uhl and Eisenberg (1970) as follows:

Formula	Source
$\bar{R}^2 = 1 - (1 - R^2)\dfrac{h - 1}{h - v}$	Wherry (1931)
$\bar{R}^2 = 1 - (1 - R^2)\dfrac{h - 1}{h - v - 1}$	McNemar (1962)
$\bar{R}^2 = 1 - (1 - R^2)\dfrac{h + v - 1}{h - v - 1}$	Lord (1950)

where h is the number of observations, and v is the number of independent variables.

$\bar{R}^2$ is constrained to be less than 1.0, but it can be negative. Try not to let this bother you; if it comes out negative, use $\bar{R}^2 = 0$.

Uhl and Eisenberg analyzed the measures of $\bar{R}^2$. They estimated the fit to calibration data, and then examined the question of which measure provided the best estimate of R^2 for a concurrent validation sample. Although Wherry's formula is the most popular, it was the least effective of the three measures in estimating the shrinkage that occurred for R^2 when going from calibration to validation sample. The McNemar formula (sometimes called a "modified Wherry") was somewhat better, but Lord's formula worked best. To remember this, just praise the Lord!

If the analyst examines many variables to obtain a good fit to the calibration data (as occurs in stepwise regression), the v in the previous formulas should be the number of variables examined, rather than the number of variables included in the final version (Uhl and Eisenberg, 1970). This is not commonly done. It is a conservative procedure, although a bit harsh. Better yet, use the tables in McINTYRE, et al. [1983].

Because some forecasters make the mistake of using R^2 rather than $\bar{R}^2$ for the calibration sample, it is worthwhile to examine the dangers of R^2. R^2 is an inflated measure of the goodness of fit. Montgomery and Morrison (1973) describe how to estimate this inflation in R^2 for various sample sizes, different numbers of variables, and different assumptions as to the true R^2. For example, for 10 observations, three independent variables, and a true $R^2 = .25$, the calculated R^2 for a sample would be expected to be .54. If one assumes the true R^2 to be 0 (the traditional

null hypothesis), there is a convenient rule of thumb for the inflation in R^2:

$$R^2 = \frac{v}{h}$$

where v is the number of independent variables, and h is the number of observations.

In most fields R^2 is more popular than R, perhaps because R^2 is a bit easier to explain. Curtis and Alf (1969) argue that R is a better measure of practical significance than is R^2. Suit yourself. Tradition is probably not worth fighting here. Furthermore, if practical importance is the issue, use the MAPE instead.

Here is a bad question to ask: "What is a good R^2?" R^2 means little in absolute terms. If you do not believe this, or if your friends are impressed by high R^2's, you may want to read Appendix G, called "Rules for Cheaters." Remember that Tom Swift had no trouble getting an R^2 of .85 with random data (Armstrong, 1970b). Also, Ames and Reiter (1961) showed that an R^2 in excess of .5 could be obtained by selecting an economic time series and regressing it against two to six other randomly selected economic time series. R^2 is most suitable when comparing different models in a given situation.

Finally, although it was assumed above that the data are interval scaled (or two categories of nominal data) in calculating R^2, other correlation coefficients assume only that the rankings are meaningful (i.e., the data are "ordinal"). Two of these measures, the Spearman and the Kendall coefficients, are described in textbooks on nonparametric statistics such as Siegel (1956).

9. Accuracy ratio ($\overline{Q}$) is calculated from

$$\overline{Q} = \frac{\sum_{t=1}^{h} Q}{h}$$

where

$$Q = \begin{cases} \dfrac{A_t}{F_t} & \text{if } A_t > F_t \\ \dfrac{F_t}{A_t} & \text{if } F_t > A_t \end{cases}$$

$\overline{Q}$ is similar to the MAPE. One key difference is that it is not bounded on both sides; it can range from 1 to infinity. For this reason, the accuracy ratio is preferable to the $\overline{\text{MAPE}}$ when dealing with large errors. In addition to being simple and easy to understand, the accuracy ratio is unit free. The loss function is assumed to be symmetrical in percentage terms; for example, if the actual value for a given problem was 10, answers of both 2 and 50 would have accuracy ratios of 5.

10a. Turning point (TP) errors may be of interest if special costs are associated with missing the direction of change. The types of turning point errors are illustrated here:

		Did a Directional Change Occur?	
		No	Yes
Was a Change Predicted?	No	a	b
	Yes	c	d

The a's represent successful predictions when no change was forecast and none occurred. The d's represent a successful prediction of a change. The b's represent errors when a change occurs that was not predicted. Finally, the c's represent errors when change was predicted but did not occur. (As the famous economist said, "We predicted nine of the last five recessions.")

The translation of TP errors into a common unit of measure depends upon the situation, that is, the relative costs and benefits of the a's, b's, c's, and d's must be assessed. But this is a difficult task if the magnitudes of the changes vary. Because magnitudes do vary in most situations, and because the magnitudes affect costs and benefits, the TP should be supplemented by other measures. Fortunately, it is seldom necessary to use TP errors. Still, turning points have been used for many years in economic problems. Descriptions of turning points can be found in Smyth (1966), and applications of TPs are given in Fels and Hinshaw (1974).

10b. "Hits and misses" is analogous to TPs but is more general; it is concerned, not with a change in direction, but with whether or not an event occurs. For example, consider a jury that faces the following possibilities in regard to a person tried for murder (considering only the simplest cases):

The Jury . . .	The Defendant . . . Did Not Commit Murder	Did Commit Murder
Does Not Convict	*a*	*b*
Convicts	*c*	*d*

In such an example, when there are "either-or" situations, the hits and misses table provides a good way to structure the problem. Costs and benefits can be assigned to each of the four categories.

Overall measures from the "hits and misses" table cause problems. Hedlund et al. (1973) discussed these problems in the context of predicting which mental patients are dangerous. So few of the patients are dangerous that an excellent overall prediction is obtained by assuming that none are dangerous. (That isn't a bad strategy; you would have to lock up a lot of nice people to detain a single dangerous one.) Joy and Tollefson (1975) examined hits and misses in the context of financial problems and Rosen (1954) used this measure in predicting suicides.

Rating the Measures of Accuracy

Although there are other measures of accuracy, the list above should be sufficient for most forecasting problems. A checklist is provided in Exhibit 13-7 to help in the selection of an appropriate measure of accuracy.

The typical procedure is to compare the various forecasting models on each of the selected measures. Ohlin and Duncan (1949) used an "index of predictive efficiency" to consider two models simultaneously. This measure can be calculated from

$$\text{IPE} = \frac{M1 - M2}{M1}$$

where $M1$ is the error measure for the current model, and $M2$ is the error measure for the proposed model. Such an approach would be helpful for some of the accuracy measures in Exhibit 13-7. For example, for the MAPE, IPE could be calculated from

$$\text{IPE} = \frac{(\text{MAPE})_1 - (\text{MAPE})_2}{(\text{MAPE})_1}$$

Exhibit 13-7 RATINGS OF THE MEASURES OF ACCURACY

Accuracy Measures	Scaling Require-ments	Unit Free?	Loss Function Symmetrical?	Assesses Both Levels and Changes?	Closely Related to Decision Making?	Easy to Understand?
1. Mean error (ME)	Interval	No	Yes	No	No	Fairly
2. Mean absolute deviation (MAD)	Interval	No	Yes	Yes	Yes	Yes
3. Root mean square error (RMSE)	Interval	No	Yes	Yes	Yes	Fairly
4a. Mean absolute percentage error (MAPE)	Ratio	Yes	No	Yes	Yes	Yes
4b. Median absolute percentage error	Ratio	Yes	Yes	Yes	Yes	Yes
5. Adjusted MAPE ($\overline{\text{MAPE}}$)	Ratio	Yes	Yes	Yes	Fairly	Yes
6. Theil's measure of inequality ($U2$)	Interval	Yes	Yes	No	No	No
7. Coefficient of variation (CV)	Ratio	Yes	Yes	Yes	Yes	Yes
8. Coefficient of determination (R^2)	Varies	Yes	Yes	No	No	Fairly
9. Accuracy ratio ($\overline{Q}$)	Ratio	Yes	Yes	Yes	Fairly	Yes
10a. Turning points (TPs) 10b. Hits and misses	Nominal	Yes	No	No	No	Fairly

When possible, it is preferable to translate the error scores into economic terms. One would like to compare the loss from forecasts using the current model with the loss that would occur under the proposed model. Dawes (1971) provided a good example of this for the prediction of which applicants to graduate school will be successful.

Once again, the strategy of eclectic research is recommended. If more than one of the measures is appropriate, use all of them. Certainly it leads to increased confidence if the good performance of a forecasting model holds up over a variety of measures. As an example, Schupack (1962) used various error measures in his study of econometric vs. extrapolation models on one-year forecasts of consumer products.

Statistical and Practical Significance

How much confidence can be placed in the results from a forecasting test? Measures of **statistical significance**[G] may be useful here. Many books describe appropriate tests: my favorites are Blalock (1979) for the standard tests and Siegel (1956) for **nonparametric tests**[G]. This section describes the philosophy behind tests of statistical significance. (For a longer discussion, see Bakan, 1966.)

Testing for statistical significance examines whether the superiority of a model is due to luck on the part of the researcher. He adopts the position that unusual events (e.g., things that happen, say, 5% of the time) do not happen to him. This 5% limit is set before analyzing the data. If the results fall in this unlikely region, he decides that luck was not responsible and that the proposed model produces different results from the current model. If the results do not fall in this ".05 level of statistical significance" region, he assumes that the proposed model offers no reliable advantage.

This procedure represents the classical approach to statistical significance. The proposed model is either accepted or rejected. Such a black and white viewpoint is difficult to use because the prior selection of the level of significance seldom is based on a cost-benefit analysis. (Quality control is one of the few areas where levels of statistical significance have been translated into cost-benefit terms.) Still, statistical significance can be of some value if one tries to set a reasonable level for significance before examining the data. The major role of significance tests is to decide how much confidence to place in a result. This is important because most of us, including trained statisticians, have a poor intuitive feel of the confidence we should place in a result. Consider the baby boy problem from Tversky and Kahneman (1971):

Baby Boy Problem

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50% of all babies are boys. The exact percentage of baby boys, however, varies from day to day. Sometimes it may be higher than 50%, sometimes lower.

For a period of one year, each hospital recorded the days on which more than 60% of the babies born were boys. Which hospital do you think recorded more such days? (Check one)

_____ The larger hospital?
_____ The smaller hospital?
_____ About the same? (i.e., within 5% of each other)

According to Tversky and Kahneman, most people do not get the correct answer (see Appendix F for the answer). This simple example (and some more complex ones from Tversky and Kahneman) suggests that we should rely upon formal tests of statistical significance and not upon our intuitions (even though these intuitions may be held strongly).

This advice, to rely on explicit statistical tests, is often ignored in forecasting studies. Significance tests are frequently performed on the calibration sample but then are ignored in comparing the accuracy of the forecasting models. It is far better to do just the reverse!

If the accuracy of more than two models is being examined, an adjustment should be made in the statistical testing. As the number of models is increased, a likelihood arises that some models will, by chance, look significant. For example, assume that 20 models are being compared with current practice. It is likely that one of these 20 comparisons will show up as being significant at the .05 level even if there are no real differences. The solution is to use tables for multiple comparisons. This too is advice that is often ignored. (Why, even I have been guilty of this oversight! Perhaps I will err in the future—but it would be wrong). Appendix C provides tables to be used for statistical significance for two or more alternative models. A discussion of relevant tests is provided in Patterson (1955).

How much of a difference makes a difference? Though a difference may be significant in a statistical sense, is it also significant in a practical sense? Such an evaluation depends upon the economics of the

situation in which the forecasts are to be used. How much money or how many lives or how many resources can be saved if a proposed forecasting model is adopted, in comparison to the current model? How much more (less) does the proposed model cost? This is what cost-benefit analysis is all about.

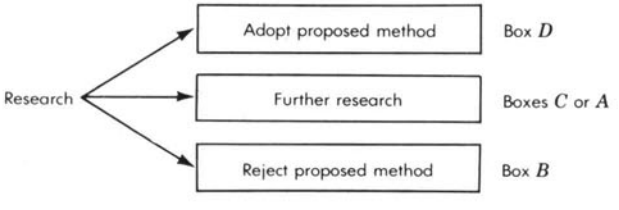
Although **practical significance**[G] and statistical significance are discussed separately, it helps to consider them jointly when interpreting the results. The key decision is whether to adopt a proposed forecasting model, to keep the old model, or to pursue furtherresearch.

A guide to the interpretation of statistical and practical significance is provided in Exhibit 13-8. The test of statistical significance helps to assess confidence, and the test of practical significance measures importance. The only surprising guideline from Exhibit 13-8 is that statistically significant but practically insignificant models, Box *B*, should be discarded; confidence is high, but the proposed model lacks practical value. This is a rule of thumb. It is preferable to the common assumption that statistical significance implies practical significance. Box *D*,

Exhibit 13-8 INTERPRETING PRACTICAL AND STATISTICAL SIGNIFICANCE

		Statistically Significant? No	Statistically Significant? Yes
Practically Significant?	No	*A* Do more work on proposed method . . . maybe.	*B* Reject proposed method.
	Yes	*C* Do more work on proposed method.	*D* Adopt proposed method.

Another way to examine the advice in Exhibit 13-8 is as follows:



significant in both a practical and a statistical sense, provides an adequate basis for adopting a proposed method. Box *A*, significant in neither a statistical nor a practical sense, provides a dilemma; it may be that there is no value (in which case one should stick with the existing method), or it may be that insufficient data have been collected. The researcher should examine the latter possibility before discarding the proposed method. Finally, in Box *C*, the results are practically but not statistically significant. This indicates the need for more research to determine whether the superiority was due to luck.

In addition to testing against a preset level of significance, it helps to report information on the calculated level of significance. The reader of a report can then see whether the results were significant in terms of a critical level of significance for the problem at hand.

Popularity of the Criteria

What criteria are most preferred by academics and practitioners? CARBONE and ARMSTRONG [1982] asked a group of experts (attendees at the first International Symposium on Forecasting in Quebec in 1981) to list the criteria they would like to use for evaluating extrapolative methods. This focus on extrapolative methods ruled out criteria such as "better understanding" and "improved policies." In retrospect, the obvious answer would seem to be "potential savings from improved decision making." However, this answer was never used. Instead, the criteria that were chosen focused upon the forecasts. The most frequently mentioned criterion was accuracy, and within this category, mean square error (MSE) (or root mean square error) was most popular, especially for academics. Second in importance were criteria related to implementation, including "ease of interpretation" (mentioned by 38%) and "ease of use/implementation" (30%). Exhibit 13-9 summarizes the responses. I was surprised by the popularity of MSE and by the lack of popularity of R^2.

Given the opportunity, one would like the client to specify the criteria. Presumably, more than one criterion would be of interest. Exhibit 13-9 indicates some typical preferences.

Benchmarks For Accuracy

This chapter has emphasized the need for assessing alternative methods in a given situation. An alternative approach is to judge accuracy against that achieved by others in similar situations. To obtain such a benchmark, you could conduct a survey of experts to determine typ-

Exhibit 13-9 RELATIVE POPULARITY OF MEASURES OF ACCURACY (Percentage of times mentioned)

	Percentage of	
	Academics (n = 63)	Practitioners (n = 63)
Mean square error (MSE or RMSE)	48	32
Mean absolute error (MAE)	19	22
Mean absolute percentage error (MAPE)	24	14
Mean percentage error (MPE)	8	8
Theil's *U*	5	2
R^2	0	3

Notes. The percentages are based on those who said accuracy was relevant. Some did not mention a specific measure, while others mentioned more than one, thus the columns do not sum to 100%.

ical levels of accuracy. Or you might find such benchmarks in the literature. For an example of the latter, the typical error in a one-year-ahead forecast of corporate earnings was determined [in ARMSTRONG, 1983b] to have a MAPE of 21.4 for judgmental forecasts. This estimate was based on many years of forecasts for 1,250 companies.

Evidence on the accuracy of sales forecasts was obtained from a survey of companies conducted by MENTZER and COX [1984]. As might be expected, the typical error increases as the forecast horizon increases and also as the forecast becomes more detailed. Exhibit 13-10 presents a summary of the typical MAPEs for sales forecasts. I do not, however, know what methods were used by these respondents. These should be treated as *very crude estimates*. ZARNOWITZ [1979], based on his studies of forecasts of GNP, suggests a rule of thumb that the forecast error is constant over the cumulative horizon. That is, the expected percentage error for a one-month forecast would be the same as for a one-quarter, one-year, or five-year forecast.

SUMMARY

Conditional forecasts can help in the analysis of the inputs to a model. Three types of conditional forecasts were summarized in Exhibit 13-1: *ex post*, modified *ex ante*, and calibration. However, the primary interest in testing outputs is to help in the comparison of alternative models. For this purpose, *ex ante* forecasts should be examined.

The testing situation should be similar to the actual forecasting

Exhibit 13-10 TYPICAL ERRORS FOR SALES FORECASTS (Entries are MAPEs)

Level	Forecast Horizon		
	Under 3 Months	3 Months to 2 Years	Over 2 Years
Industry	8	11	15
Corporate	7	11	18
Product group	10	15	20
Product line	11	16	20
Product	16	21	26

Source. MENTZER and COX [1984] survey results from 160 corporations. These results are crude estimates because most firms do not keep systematic records. Further, the report of the study was ambiguous in its definitions of the time interval. We suppose that "Under 3 months" is intended to mean 'monthly,' but the length of time is not apparent for "Over 2 years."

problem. The validation matrix was provided in Exhibit 13-2 to help in the selection of appropriate situations, and these various situations were ranked in Exhibit 13-5. Most appropriate, by far, is to simulate the actual forecasting situation, using the decision units that are of interest. Least appropriate, but most popular, is to infer the forecast validity from the calibration sample. Backcast validity is a useful though seldom used test. Much attention was given to tests of concurrent validation because these tests are nearly always feasible; procedures were summarized in Exhibit 13-3 for examining concurrent validity for large and small samples.

The selection of a measure of accuracy depends to a great extent upon the situation. Twelve measures of accuracy were described, and ratings of these measures were provided in Exhibit 13-7. The MAPE provides a good measure for most situations. R^2, with all its problems, may do more harm than good.

Formal tests of statistical significance are preferable to the use of intuitive measures. Our common sense frequently is misleading. Tests of statistical significance were related to practical significance in Exhibit 13-8.

The popularity of criteria among academics and practitioners showed much agreement; the preferred measure, in general, was the root mean square error.(See Exhibit 13-9.)

The chapter concluded with a summary of typical errors for various levels of aggregation and for various time horizons (Exhibit 13-10).