Chapter 10

Point Forecast Evaluation

As we've stressed repeatedly, good forecasts lead to good decisions. The importance of forecast evaluation techniques follows immediately. Given a track record of forecasts, $y_{t+h,t}$, and corresponding realizations, y_{t+h}, we naturally want to monitor and improve forecast performance. In this chapter we show how to do so. We discuss both absolute aspects of forecast evaluation, focusing on methods for checking forecast optimality, and relative aspects, focusing on methods for ranking forecast accuracy, quite apart from optimality.

10.1 Absolute Standards for Point Forecasts

Think about evaluating a single forecast, in isolation. Evaluating a single forecast amounts to checking whether it has the properties expected of an optimal forecast. Denote by y_t the covariance stationary time series to be forecast. The Wold representation is

$$y_t = \mu + \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + ...$$

$$\varepsilon_t \sim WN(0,\ \sigma^2).$$

Thus the h-step-ahead linear least-squares forecast is

$$y_{t+h,t} = \mu + b_h\varepsilon_t + b_{h+1}\varepsilon_{t-1} + ...$$

and the corresponding h-step-ahead forecast error is

$$e_{t+h,t} = y_{t+h} - y_{t+h,t} = \varepsilon_{t+h} + b_1\varepsilon_{t+h-1} + ... + b_{h-1}\varepsilon_{t+1},$$

with variance

$$\sigma_h^2 = \sigma^2 \left(1 + \sum_{i=1}^{h-1} b_i^2\right).$$

The key property of optimal forecast errors, from which all others follow, (including those cataloged below), is that they should be unforecastable on the basis of information available at the time the forecast was made. This **unforecastability principle** is valid in great generality; it holds, for example, regardless of whether linear-projection optimality or conditional-mean optimality is of interest, regardless of whether the relevant loss function is quadratic, and regardless of whether the series being forecast is stationary.

Many tests of aspects of optimality are based on the unforecastability principle. 1-step-ahead errors, for example, had better be white noise, because otherwise we could forecast the errors using information readily available when the forecast is made. Indeed at least four key properties of optimal forecasts, which we can easily check, follow immediately from the unforecastability principle:

a. Optimal forecasts are unbiased

b. Optimal forecasts have 1-step-ahead errors that are white noise

c. Optimal forecasts have h-step-ahead errors that are at most $MA(h-1)$

d. Optimal forecasts have h-step-ahead errors with variances that are nondecreasing in h and that converge to the unconditional variance of the process.

10.1.1 Are errors zero-mean?

If the forecast is unbiased, then the forecast error has a zero mean. A variety of tests of the zero-mean hypothesis can be performed, depending on the assumptions we're willing to maintain. For example, if $e_{t+h,t}$ is Gaussian white noise (as might be reasonably the case for 1-step-ahead errors), then the standard t-test is the obvious choice. We would simply regress the forecast error series on a constant and use the reported t-statistic to test the hypothesis that the population mean is zero. If the errors are non-Gaussian but remain iid, then the t-test is still applicable in large samples.

If the forecast errors are dependent, then more sophisticated procedures are required. We maintain the framework of regressing on a constant, but we must "correct" for any serial correlation in the disturbances. Serial correlation in forecast errors can arise for many reasons. Multi-step-ahead forecast errors will be serially correlated, even if the forecasts are optimal, because of the forecast-period overlap associated with multi-step-ahead forecasts. More generally, serial correlation in forecast errors may indicate that the forecasts are suboptimal. The upshot is simply that when regressing forecast errors on an intercept, we need to be sure that any serial correlation in the disturbance is appropriately modeled. A reasonable starting point for a regression involving h-step-ahead forecast errors is $MA(h-1)$ disturbances, which we'd expect if the forecast were optimal. The forecast may, of course, *not* be optimal, so we don't adopt $MA(h-1)$ disturbances uncritically; instead, we try a variety of models using the AIC and SIC to guide selection in the usual way.

10.1.2 Are 1-step-ahead errors white noise?

Under various sets of maintained assumptions, we can use standard tests of the white noise hypothesis. For example, the sample autocorrelation and

partial autocorrelation functions, together with Bartlett asymptotic standard errors, are often useful in that regard. Tests based on the first autocorrelation (e.g., the Durbin-Watson test), as well as more general tests, such as the Box-Pierce and Ljung-Box tests, are useful as well.

10.1.3 Are h-step-ahead errors are at most $MA(h-1)$?

The $MA(h-1)$ structure implies a cutoff in the forecast error's autocorrelation function beyond displacement $h-1$. This immediately suggests examining the statistical significance of the sample autocorrelations beyond displacement $h-1$ using the Bartlett standard errors. In addition, we can regress the errors on a constant, allowing for $MA(q)$ disturbances with $q > (h-1)$, and test whether the moving-average parameters beyond lag $h-1$ are zero.

10.1.4 Are h-step-ahead error variances non-decreasing in h?

It's often useful to examine the sample h-step-ahead forecast error variances as a function of h, both to be sure they're non-decreasing in h and to see their *pattern*, which may convey useful information.

10.1.5 Are errors orthogonal to available information?

The tests above make incomplete use of the unforecastability principle, insofar as they assess only the *univariate* properties of the errors. We can make a more complete assessment by broadening the information set and assessing optimality with respect to various sets of information, by estimating regressions of the form

$$e_{t+h,t} = \alpha_0 + \sum \alpha_i x_{it} + u_t.$$

The hypothesis of interest is that all the α's are zero, which is a necessary condition for forecast optimality (orthogonality) with respect to available

information.

The particular case of testing optimality with respect to $y_{t+h,t}$ is very important in practice. (Note that $y_{t+h,t}$ is obviously in the time-t information set.) The relevant regression is

$$e_{t+h,t} = \alpha_0 + \alpha_1 y_{t+h,t} + u_t,$$

and optimality corresponds to $(\alpha_0,\ \alpha_1) = (0,\ 0)$.

If the above regression seems a little strange to you, consider what may seem like a more natural approach to testing optimality, regression of the realization on the forecast:

$$y_{t+h} = \beta_0 + \beta_1 y_{t+h,t} + u_t.$$

This is called a "**Mincer-Zarnowitz regression**." If the forecast is optimal with respect to the information used to construct it, then we'd expect $(\beta_0,\ \beta_1) = (0,\ 1)$, in which case

$$y_{t+h} = y_{t+h,t} + u_t.$$

Note, however, that if we start with the regression

$$y_{t+h} = \beta_0 + \beta_1 y_{t+h,t} + u_t,$$

and then subtract $y_{t+h,t}$ from each side, we obtain

$$e_{t+h,t} = \alpha_0 + \alpha_1 y_{t+h,t} + u_t,$$

where $(\alpha_0,\ \alpha_1) = (0,\ 0)$ when $(\beta_0,\ \beta_1) = (0,\ 1)$. Thus, the two approaches are identical. We can regress the error on an intercept and the forecast and test (0, 0), or we can regress the realization on an intercept and the forecast and test (0, 1).

10.2 Relative Standards for Point Forecasts

Now think about ranking a set of forecasts, quite apart from how any or all of them fare regarding the absolute optimality criteria assessed in section 10.1.

10.2.1 Accuracy Rankings via Expected Loss

The crucial object in measuring forecast accuracy is the loss function, $L(y_{t+h},\ y_{t+h,t})$, often restricted to $L(e_{t+h,t})$, which charts the "loss," "cost," or "disutility" associated with various pairs of forecasts and realizations.[1] In addition to the shape of the loss function, the forecast horizon h is of crucial importance. Rankings of forecast accuracy may of course be very different across different loss functions and different horizons.

Let's discuss a few accuracy measures that are important and popular. Accuracy measures are usually defined on the forecast errors,

$$e_{t+h,t} = y_{t+h} - y_{t+h,t},$$

or percent errors,

$$p_{t+h,t} = (y_{t+h} - y_{t+h,t})/y_{t+h}.$$

Mean error measures forecast-error location, which is one component of accuracy. In population we write

$$\mu_{e_{t+h,t}} = E(e_{t+h,t}),$$

and in sample we write

$$\hat{\mu}_{e_{t+h,t}} = \frac{1}{T}\sum_{t=1}^{T} e_{t+h,t}.$$

The mean error is the forecast **bias**. Other things the same, we prefer a

[1] Because in many applications the loss function will be a direct function of the forecast error, $L(y_t,\ y_{t+h,t}) = L(e_{t+h,t})$, we write $L(e_{t+h,t})$ from this point on to economize on notation, while recognizing that certain loss functions (such as direction-of-change) don't collapse to the $L(e_{t+h,t})$ form.

forecast with small bias.

Error variance measures dispersion of the forecast errors, which is another component of accuracy. In population we write

$$\sigma^2_{e_{t+h,t}} = E(e_{t+h,t} - \mu_{e_{t+h,t}})^2,$$

and in sample we write

$$\hat{\sigma}^2_{e_{t+h,t}} = \frac{1}{T}\sum_{t=1}^{T}(e_{t+h,t} - \hat{\mu}_{e_{t+h,t}})^2.$$

Other things the same, we prefer a forecast with small error variance.

Although the mean error and the error variance are components of accuracy, neither provides an overall accuracy measure. For example, one forecast might have a small $\hat{\mu}_{e_{t+h,t}}$ but a large $\hat{\sigma}^2_{e_{t+h,t}}$, and another might have a large $\hat{\mu}_{e_{t+h,t}}$ and a small $\hat{\sigma}^2_{e_{t+h,t}}$. Hence we would like an accuracy measure that somehow incorporates *both* the mean error and error variance.

The **mean squared error** does just that. It is the most common overall accuracy measure, by far. In population we write

$$MSE_{e_{t+h,t}} = E(e_{t+h,t})^2,$$

and in sample we write

$$\widehat{MSE}_{e_{t+h,t}} = \frac{1}{T}\sum_{t=1}^{T} e^2_{t+h,t}.$$

This "bias-variance tradeoff" is a crucially important insight for forecasting. Among other things, it highlights the fact that bias is not necessarily "bad," under quadratic loss (MSE). We'd be happy, for example, to take a small bias increase in exchange for a massive variance reduction.

We sometimes take square roots to preserve units, yielding the **root mean**

squared error. In population we write

$$RMSE_{e_{t+h,t}} = \sqrt{E(e_{t+h,t})^2},$$

and in sample we write

$$\widehat{RMSE}_{e_{t+h,t}} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e_{t+h,t}^2}.$$

To understand the meaning of "preserving units," and why it's sometimes helpful to do so, suppose that the forecast errors are measured in dollars. Then the mean squared error, which is built up from *squared* errors, is measured in dollars *squared*. Taking square roots – that is, moving from MSE to RMSE – brings the units back to dollars.

MSE can be decomposed into bias and variance components, reflecting the tradeoff between bias and variance forecast accuracy under quadratic loss. In particular, MSE can be decomposed into the sum of variance and squared bias. In population we write

$$MSE_{e_{t+h,t}} = \sigma^2_{e_{t+h,t}} + \mu^2_{e_{t+h,t}},$$

and in sample we write

$$\widehat{MSE}_{e_{t+h,t}} = \hat{\sigma}^2_{e_{t+h,t}} + \hat{\mu}^2_{e_{t+h,t}}.$$

Mean absolute error is a less popular, but nevertheless common, overall accuracy measure. In population we write

$$MAE_{e_{t+h,t}} = E|e_{t+h,t}|,$$

and in sample we write

$$\widehat{MAE} = \frac{1}{T}\sum_{t=1}^{T} |e_{t+h,t}|.$$

When using MAE we don't have to take square roots to preserve units.

10.2.2 On MSE vs. MAE

Introspection suggests using MAE – not MSE – as the canonical benchmark loss function. Consider using the distribution of e directly, ranking forecasts by the distance of $F(e)$ from $F^*(\cdot)$, the unit step function at 0 (the cdf of errors from a perfect forecast, which are 0 w.p. 1). That is, rank forecasts by

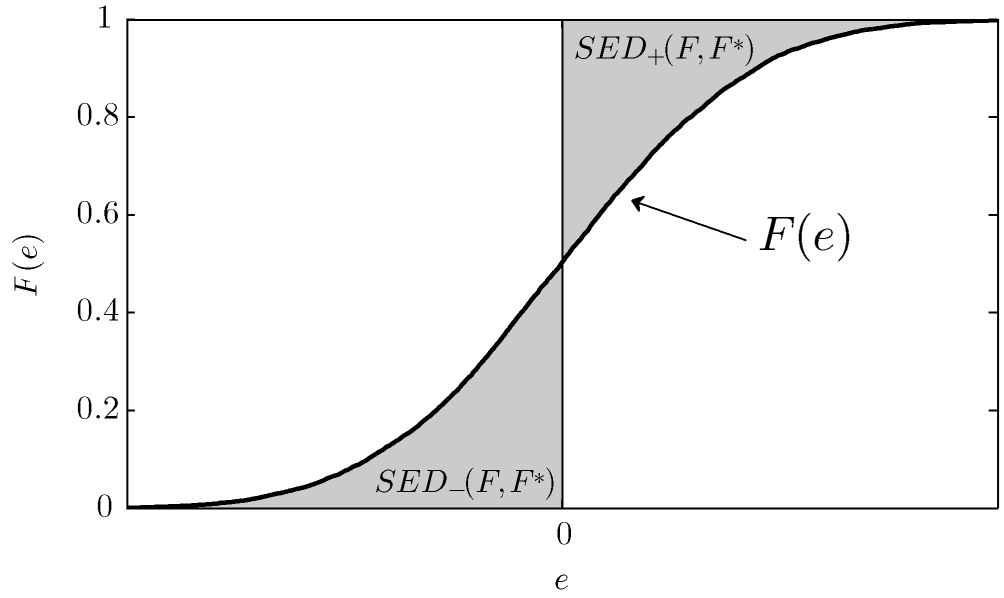
$$SED(F, F^*) = \int_{-\infty}^{\infty} |F(e) - F^*(e)|\, de,$$

where smaller is better. We call $SED(F, F^*)$ the *stochastic error distance*. In Figure 10.1a we show $SED(F, F^*)$, and in Figure 10.1b we provide an example of two error distributions such that one would prefer F_1 to F_2 under $SED(F, F^*)$.

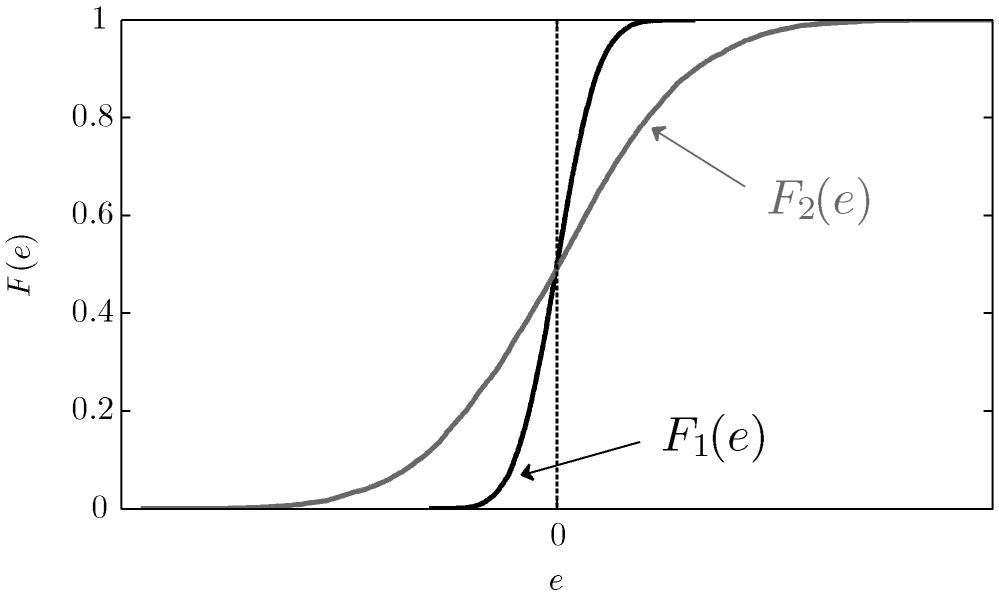
We motivated $SED(F, F^*)$ as directly appealing and intuitive. It turns out, moreover, that $SED(F, F^*)$ is intimately connected to one, and only one, traditionally-invoked loss function, and it is not quadratic. In particular, for any forecast error e, with cumulative distribution function $F(e)$ such that $E(|e|) < \infty$, we have

$$SED(F, F^*) = \int_{-\infty}^{0} F(e)\, de + \int_{0}^{\infty} [1 - F(e)]\, d\, e = E(|e|). \tag{10.1}$$

That is, $SED(F, F^*)$ equals expected absolute loss for any error distribution. Hence if one is comfortable with $SED(F, F^*)$ and wants to use it to evaluate forecast accuracy, then one must also be comfortable with expected absolute-error loss and want to use it to evaluate forecast accuracy. The two criteria



(a) c.d.f. of e. Under the $SED(F, F^*)$ criterion, we prefer smaller $SED(F, F^*) = SED_-(F, F^*) + SED_+(F, F^*)$.



(b) Two forecast error distributions. Under the $SED(F, F^*)$ criterion, we prefer $F_1(e)$ to $F_2(e)$.

Figure 10.1: Stochastic Error Distance ($SED(F, F^*)$)

are *identical*.

10.2.3 Benchmark Comparisons

It is sometimes of interest to compare forecast performance to that of an allegedly-naive benchmark.

Predictive R^2

Recall the formula for R^2,

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2},$$

where e_t is the in-sample regression residual. If we replace the e_t's with $e_{t,t-1}$'s, out-of-sample 1-step forecast errors, then we get the predictive R^2,

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_{t,t-1}^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2},$$

Predictive R^2 compares an estimate of 1-step-ahead out-of-sample forecast error variance to an estimate of unconditional variance. Put differently, it compares actual 1-step forecast accuracy to that of the historical mean forecast, $\bar{y}$. The hope is that the former is much smaller than the latter, in which case the predictive R^2 will be near 1.

h-step-ahead versions of predictive R^2's are immediate. We simply replace $e_{t,t-1}$ with $e_{t,t-h}$ in the formulas.

Theil's U-Statistic

The so-called "Theil U-statistic" is just a predictive R^2, but we change the benchmark from the historical mean forecast, $\bar{y}$, to a "no change" forecast,

y_{t-1},

$$U = 1 - \frac{\sum_{t=1}^{T} e_{t,t-1}^2}{\sum_{t=1}^{T} (y_t - y_{t-1})^2}.$$

In the meteorological literature measures like U are called "skill scores," because they assess actual skill relative to a potentially-naive forecast.

It is important to note that allegedly-naive benchmarks may not be so naive. For example, many economic variables may in fact be nearly random walks, in which case forecasters will have great difficulty beating the random walk through no fault of their own (i.e., the predictive R^2 relative to a random walk "no-change" forecast given by Theil's U may be near 0)!

10.2.4 Measures of Forecastability

Forecastability measures are a leading example of benchmark comparisons, as we discuss them here.

It is natural and informative to judge forecasts by their accuracy. However, actual and forecasted values will differ, even for good forecasts. To take an extreme example, consider a zero-mean white noise process. The optimal linear forecast under quadratic loss in this case is simply zero, so the paths of forecasts and realizations will clearly look different. These differences illustrate the inherent limits to predictability, even when using optimal forecasts. The extent of a series' predictability depends on how much information the past conveys regarding future values of this series; as a result, some processes are inherently easy to forecast, and others are more difficult. Note also that predictability and volatility are different concepts; predictability is about the *ratio* of conditional to unconditional variance, whereas volatility is simply about unconditional variance.

Below we discuss some of the difficulties involved in predictability measurement and propose a simple measure of relative predictability based on the ratio of the expected loss of an optimal short-run forecast to the expected loss

of an optimal long-run forecast. Our measure allows for covariance stationary or difference stationary processes, univariate or multivariate information sets, general loss functions, and different forecast horizons of interest. First we propose parametric methods for estimating the predictability of observed series, and then we discuss alternative nonparametric measures, survey-based measures, and more.

Population Measures

The expected loss of an optimal forecast will in general exceed zero, which illustrates the inherent limits to predictability, even when using optimal forecasts. Put differently, poor forecast accuracy does not necessarily imply that the forecaster failed. The extent of a series' predictability in population depends on how much information the past conveys regarding the future; given an information set, some processes are inherently easy to forecast, and others are more difficult.

In measuring predictability it is important to keep two points in mind. First, the question of whether a series is predictable or not should be replaced by one of *how* predictable it is. Predictability is always a matter of degree. Second, the question of how predictable a series is cannot be answered in general. We have to be clear about the relevant forecast horizon and loss function. For example, a series may be highly predictable at short horizons, but not at long horizons.

A natural measure of the forecastability of covariance stationary series under squared-error loss, patterned after the familiar regression R^2, is

$$G = 1 - \frac{var(e_{t+j,t})}{var(y_{t+j})},$$

where $\hat{y}_{t+j,t}$ is the optimal (i.e., conditional mean) forecast and $e_{t+j,t} = y_{t+j} - \hat{y}_{t+j,t}$.

We can also relax several constraints that limit the broad applicability of the predictive R^2 above. Its essence is basing measures of predictability on

the difference between the conditionally expected loss of an optimal short-run forecast, $E(L(e_{t+j,t}))$, and that of an optimal long-run forecast, $E(L(e_{t+k,t}))$,, $j \ll k$, , where $E(\cdot)$ denotes the mathematical expectation conditional on the information set Ω. If $E(L(e_{t+j,t})) \ll E(L(e_{t+k,t}))$, we say that the series is highly predictable at horizon j relative to k, and if $E(L(e_{t+j,t})) \approx E(L(e_{t+k,t}))$. we say that the series is nearly unpredictable at horizon j relative to k. Thus, we define a general measure of predictability as

$$P(L, \Omega, j, k) = 1 - \frac{E(L(e_{t+j,t})}{E(L(e_{t+k,t})},$$

where the information set Ω can be univariate or multivariate, as desired. The predictive R^2 measure emerges when the series is covariance stationary, $L(x) = x^2$ (and hence the optimal forecast is the conditional mean), the information set is univariate, and $k = \infty$. The advantages of our generalization include: (1) It is valid for both covariance stationary and difference stationary series, so long as $k < \infty$. (2) It allows for general loss functions. The loss function $L(\cdot)$ need not be quadratic or even symmetric; we only require that L(0) = 0 and that $L(\cdot)$ be strictly monotone on each side of the origin. By the restrictions imposed on $L(\cdot)$, we have that for all covariance stationary or difference stationary processes $P(L(\cdot), \Omega, j, k) \in [0, 1]$, with larger values indicating greater predictability. (3) It allows for univariate or multivariate information sets, and economic theory may suggest relevant multivariate information sets. (4) It allows for flexibility in the choice of j and k and enables one to tailor the predictability measure to the horizons of economic interest.

Our predictability measure is closely related to Theil's U statistic, which we define for the 1-step-ahead horizon as

$$U = \frac{E(e_{t,t-1}^2)}{E((y_t - y_{t-1})^2)}.$$

To see this, specialize P to the quadratic, univariate, $j = 1$ case and write it

as

$$P(quadratic, univariate, 1, k) = 1 - \frac{E(e_{t,t-1}^2)}{E(e_{t,t-k}^2)},$$

or

$$1 - P = \frac{E(e_{t,t-1}^2)}{E(e_{t,t-k}^2)}.$$

Thus, under certain conditions, $1-P$ is similar in spirit to Theil's U. The key difference is that Theil's U assesses 1-step forecast accuracy relative to that of a "naive" no-change forecast, whereas P assesses 1-step accuracy relative to that of a long-horizon (k-step) forecast. In the general case,

$$P(L(\cdot), \Omega, j, k) = 1 - \frac{E(L(e_{t,t-j}))}{E(L(e_{t,t-k}))}.$$

Thus, $P(L(\cdot), \Omega, j, k)$ is effectively one minus the ratio of expected losses of two forecasts of the same object, y_t. Typically, one forecast, $\hat{y}_{t,t-j}$, is based on a rich information set, while the other forecast, $\hat{y}_{t,t-k}$, is based on a sparse information set.

The formula for $P(L(\cdot), \Omega, j, k)$ also makes clear that the concept of predictability is related to, but distinct from, the concept of persistence of a series. Suppose, for example, that the series y_t is a random walk. Then

$$P(e^2, \text{univariate}, j, k) = 1 - \frac{j}{k},$$

as will be shown later. The corresponding j-step variance ratio, a common persistence measure, is

$$V_j = \frac{var(y_t - y_{t-j})}{var(y_t - y_{t-1})} = j.$$

It is clear, however, that although $P(e^2, \text{univariate}, j, k)$ and V_j are deterministically related in the random walk case ($P = 1 - V/k$), they are not deterministically related in more general cases.

Sample Measures

Predictability is a population property of a series, not of any particular sample path, but predictability can be estimated from a sample path. We proceed by fitting a parametric model and then transforming estimates of the parameters into an estimate of P. To keep the discussion tractable, and in keeping with the empirical analysis of subsequent sections, we postulate a quadratic loss function $L(e) = e^2$ for estimation, prediction, model selection, and construction of predictability measures.

It is clear that parametric measures of predictability in general will depend on the specification of the parametric model. Here we focus on univariate autoregressive models, although one could easily generalize the discussion to other parametric models, such as vector $ARMA$ models. We construct P by simply reading off the appropriate diagonal elements of the forecast MSE matrices for forecast horizons j and k. To build intuition, consider a univariate $AR(1)$ population process with innovation variance Σ_u: $y_t = A_1 y_{t-1} + u_t$. Then for $A_1 = 0$ the model reduces to white noise, and short-run forecasts are just as accurate as long-run forecasts. As a result, relative predictability is zero: $P(j,k) = 1 - \Sigma_u \ / \ \Sigma_u = 0$, for all j. In contrast, for $A_1 = 1$ the model becomes a random walk, and relative predictability steadily declines as the forecast horizon increases: $P(j,k) = 1 - (j\Sigma_u)/(k\Sigma_u) = 1 - j/k$.

Forecast errors from consistently estimated processes and processes with known parameters are asymptotically equivalent. In practice, we estimate P by replacing the underlying unknown parameters by their least squares estimates.

10.2.5 Statistical Assessment of Accuracy Rankings

Once we've decided on a loss function, it is often of interest to know whether one forecast is more accurate than another. In hypothesis testing terms, we might want to test the equal accuracy hypothesis,

$$E[L(e^a_{t+h,t})] = E[L(e^b_{t+h,t})],$$

against the alternative hypothesis that one or the other is better. Equivalently, we might want to test the hypothesis that the expected loss differential is zero,

$$E(d_t) = E[L(e^a_{t+h,t})] - E[L(e^b_{t+h,t})] = 0.$$

The hypothesis concerns population expected loss; we test it using sample average loss.

A Motivational Example

Consider a model-free forecasting environment, as for example with forecasts based on surveys, forecasts extracted from financial markets, forecasts obtained from prediction markets, or forecasts based on expert judgment. One routinely has competing model-free forecasts of the same object, gleaned for example from surveys or financial markets, and seeks to determine which is better.

To take a concrete example, consider U.S. inflation forecasting. One might obtain survey-based forecasts from the Survey of Professional Forecasters (S), $\{\pi^S_t\}^T_{t=1}$, and simultaneously one might obtain market-based forecasts from inflation-indexed bonds (B), $\{\pi^B_t\}^T_{t=1}$. Suppose that loss is quadratic and that during $t = 1, ..., T$ the sample mean-squared errors are $\widehat{MSE}(\pi^S_t) = 1.80$ and $\widehat{MSE}(\pi^B_t) = 1.92$. Evidently "$S$ wins," and one is tempted to conclude that S provides better inflation forecasts than does B. The forecasting literature is filled with such horse races, with associated declarations of superiority based

on outcomes.

Obviously, however, the fact that $\widehat{MSE}(\pi_t^S) < \widehat{MSE}(\pi_t^B)$ in a particular sample realization does not mean that S is necessarily truly better than B in population. That is, even if in population $MSE(\pi_t^S) = MSE(\pi_t^B)$, in any particular sample realization $t = 1, ..., T$ one or the other of S and B must "win," so the question arises in any particular sample as to whether S is truly superior or merely lucky. The Diebold-Mariano test answers that question, allowing one to assess the significance of apparent predictive superiority. It provides a test of the hypothesis of equal expected loss (in our example, $MSE(\pi_t^S) = MSE(\pi_t^B)$), valid under quite general conditions including, for example, wide classes of loss functions and forecast-error serial correlation of unknown form.

The Diebold-Mariano Perspective

The essence of the DM approach is to take forecast errors as primitives, intentionally, and to make assumptions directly on those forecast errors. (In a model-free environment there are obviously no models about which to make assumptions.) More precisely, DM relies on assumptions made directly on the forecast error *loss differential*. Denote the loss associated with forecast error e_t by $L(e_t)$; hence, for example, time-t quadratic loss would be $L(e_t) = e_t^2$. The time-t loss differential between forecasts 1 and 2 is then $d_{12t} = L(e_{1t}) - L(e_{2t})$. DM requires only that the loss differential be covariance stationary.[2] That is, DM assumes that:

$$Assumption\ DM: \begin{cases} E(d_{12t}) = \mu, \ \forall t \\ cov(d_{12t}, d_{12(t-\tau)}) = \gamma(\tau), \ \forall t \\ 0 < var(d_{12t}) = \sigma^2 < \infty. \end{cases} \tag{10.2}$$

[2] Actually covariance stationarity is sufficient but may not be strictly necessary, as less-restrictive types of mixing conditions could presumably be invoked.

The key hypothesis of equal predictive accuracy (i.e., equal expected loss) corresponds to $E(d_{12t}) = 0$, in which case, under the maintained Assumption DM:

$$DM_{12} = \frac{\bar{d}_{12}}{\hat{\sigma}_{\bar{d}_{12}}} \xrightarrow{d} N(0,1), \tag{10.3}$$

where $\bar{d}_{12} = \frac{1}{T}\sum_{t=1}^{T} d_{12t}$ is the sample mean loss differential and $\hat{\sigma}_{\bar{d}_{12}}$ is a consistent estimate of the standard deviation of $\bar{d}_{12}$ (more on that shortly). That's all: If Assumption DM holds, then the $N(0,1)$ limiting distribution of test statistic DM *must* hold.

DM is simply an asymptotic z-test of the hypothesis that the mean of a constructed but observed series (the loss differential) is zero. The only wrinkle is that forecast errors, and hence loss differentials, may be serially correlated for a variety of reasons, the most obvious being forecast sub-optimality. Hence the standard error in the denominator of the DM statistic (10.3) should be calculated robustly. A simple approach is to recognize that DM is just a t-statistic for the hypothesis of a zero population mean loss differential, adjusted to reflect the fact that the loss differential series is not necessarily, so that we can compute it via HAC regression (e.g., Newey-West or Kiefer-Vogelsang) on an intercept. Perhaps an even simpler approach is to regress the loss differential on an intercept, allowing for $AR(p)$ disturbances, and using information criterion like AIC to select p.

DM is also readily extensible. The key is to recognize that the DM statistic can be trivially calculated by regression of the loss differential on an intercept, using heteroskedasticity and autocorrelation robust (HAC) standard errors. Immediately, then (and as noted in the original Diebold-Mariano paper), one can potentially extend the regression to condition on additional variables that may explain the loss differential, thereby moving from an un-

conditional to a conditional expected loss perspective.[3] For example, comparative predictive performance may differ by stage of the business cycle, in which case one might include a 0-1 NBER business cycle chronology variable (say) in the DM HAC regression.

Thoughts on Assumption *DM*

Thus far I have praised DM rather effusively, and its great simplicity and wide applicability certainly *are* virtues: There is just one Assumption DM, just one DM test statistic, and just one DM limiting distribution, always and everywhere. But of course everything hinges on Assumption DM. Here I offer some perspectives on the validity of Assumption DM.

First, as George Box (1979) famously and correctly noted, "All models are false, but some are useful." Precisely the same is true of *assumptions*. Indeed all areas of economics benefit from assumptions that are surely false if taken literally, but that are nevertheless useful. So too with Assumption DM. Surely d_t is likely never *precisely* covariance stationary, just as surely *no* economic time series is likely precisely covariance stationary. But in many cases Assumption DM may be a useful approximation.

Second, special forecasting considerations lend support to the validity of Assumption DM. Forecasters strive to achieve forecast optimality, which corresponds to unforecastable covariance-stationary errors (indeed white-noise errors in the canonical 1-step-ahead case), and hence unforecastable covariance-stationary loss differentials. Of course forecasters may not achieve optimality, resulting in serially-correlated, and indeed forecastable, forecast errors. But $I(1)$ non-stationarity of forecast errors takes serial correlation to the extreme.[4]

Third, even in the extreme case where nonstationary components somehow *do* exist in forecast errors, there is reason to suspect that they may be shared.

[3] Important subsequent work takes the conditional perspective farther; see Giacomini and White (2006).

[4] Even with apparent nonstationarity due to apparent breaks in the loss differential series, Assumption DM may nevertheless hold if the breaks have a stationary rhythm, as for example with hidden-Markov processes in the tradition of Hamilton (1989).

In particular, information sets overlap across forecasters, so that forecast-error nonstationarities may vanish from the loss differential. For example, two loss series, each integrated of order one, may nevertheless be cointegrated with cointegrating vector $(1, -1)$. Suppose for example that $L(e_{1t}) = x_t + \varepsilon_{1t}$ and $L(e_{2t}) = x_t + \varepsilon_{2t}$, where x_t is a common nonstationary $I(1)$ loss component, and ε_{1t} and ε_{2t} are idiosyncratic stationary $I(0)$ loss components. Then $d_{12t} = L(e_{1t}) - L(e_{2t}) = \varepsilon_{1t} - \varepsilon_{2t}$ is $I(0)$, so that the loss differential series is covariance stationary despite the fact that neither individual loss series is covariance stationary.

Fourth, and most importantly, standard and powerful tools enable empirical assessment of Assumption DM. That is, the approximate validity of Assumption DM is ultimately an empirical matter, and a wealth of diagnostic procedures are available to help assess its validity. One can plot the loss differential series, examine its sample autocorrelations and spectrum, test it for unit roots and other nonstationarities including trend, structural breaks or evolution, and so on.

10.3 OverSea Shipping

We'll work with an application to OverSea Services, Inc., a major international cargo shipper. To help guide fleet allocation decisions, each week OverSea makes forecasts of volume shipped over each of its major trade lanes, at horizons ranging from 1-week ahead through 16-weeks-ahead. In fact, OverSea produces two sets of forecasts – a quantitative forecast is produced using modern quantitative techniques, and a judgmental forecast is produced by soliciting the opinion of the sales representatives, many of whom have years of valuable experience.

Here we'll examine the realizations and 2-week-ahead forecasts of volume on the Atlantic East trade lane (North America to Europe). We have

nearly ten years of data on weekly realized volume (VOL) and weekly 2-week-ahead forecasts (the quantitative forecast $VOLQ$, and the judgmental forecast $VOLJ$), from January 1988 through mid-July 1997, for a total of 499 weeks.

In Figure 1, we plot realized volume vs. the quantitative forecast, and in Figure 2 we show realized volume vs. the judgmental forecast. The two plots look similar, and both forecasts appear quite accurate; it's not too hard to forecast shipping volume just two weeks ahead.

In Figures 3 and 4, we plot the errors from the quantitative and judgmental forecasts, which are more revealing. The quantitative error, in particular, appears roughly centered on zero, whereas the judgmental error seems to be a bit higher than zero on average. That is, the judgmental forecast appears biased in a pessimistic way – on average, actual realized volume is a bit higher than forecasted volume.

In Figures 5 and 6, we show histograms and related statistics for the quantitative and judgmental forecast errors. The histograms confirm our earlier suspicions based on the error plots; the histogram for the quantitative error is centered on a mean of -.03, whereas that for the judgmental error is centered on 1.02. The error standard deviations, however, reveal that the judgmental forecast errors vary a bit less around their mean than do the quantitative errors. Finally, the Jarque-Bera test can't reject the hypothesis that the errors are normally distributed.

In Tables 1 and 2 and Figures 7 and 8, we show the correlograms of the quantitative and judgmental forecast errors. In each case, the errors appear to have $MA(1)$ structure; the sample autocorrelations cut off at displacement 1, whereas the sample partial autocorrelations display damped oscillation, which is reasonable for 2-step-ahead forecast errors.

To test for the statistical significance of bias, we need to account for the $MA(1)$ serial correlation. To do so, we regress the forecast errors on a con-

stant, allowing for $MA(1)$ disturbances. We show the results for the quantitative forecast errors in Table 3, and those for the judgmental forecast errors in Table 4. The t-statistic indicates no bias in the quantitative forecasts, but sizable and highly statistically significant bias in the judgmental forecasts.

In Tables 5 and 6, we show the results of Mincer-Zarnowitz regressions; both forecasts fail miserably. We expected the judgmental forecast to fail, because it's biased, but until now no defects were found in the quantitative forecast.

Now let's compare forecast accuracy. We show the histogram and descriptive statistics for the squared quantitative and judgmental errors in Figures 9 and 10. The histogram for the squared judgmental error is pushed rightward relative to that of the quantitative error, due to bias. The $RMSE$ of the quantitative forecast is 1.26, while that of the judgmental forecast is 1.48.

In Figure 11 we show the (quadratic) loss differential; it's fairly small but looks a little negative. In Figure 12 we show the histogram of the loss differential; the mean is -.58, which is small relative to the standard deviation of the loss differential, but remember that we have not yet corrected for serial correlation. In Table 7 we show the correlogram of the loss differential, which strongly suggests $MA(1)$ structure. The sample autocorrelations and partial autocorrelations, shown in Figure 13, confirm that impression. Thus, to test for significance of the loss differential, we regress it on a constant and allow for $MA(1)$ disturbances; we show the results in Table 8. The mean loss differential is highly statistically significant, with a p-value less than .01; we conclude that the quantitative forecast is more accurate than the judgmental forecast under quadratic loss.

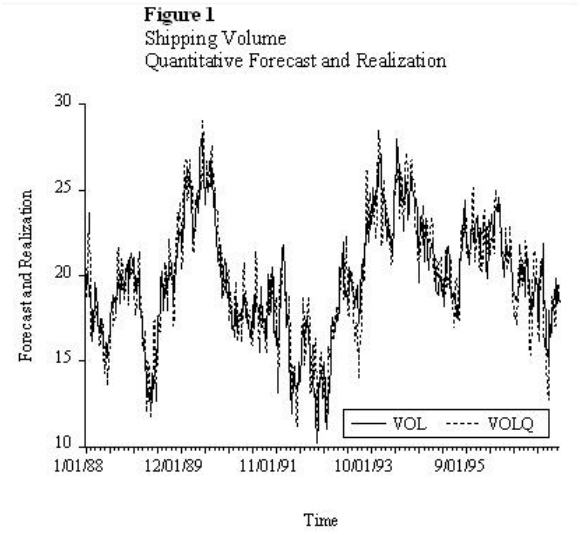
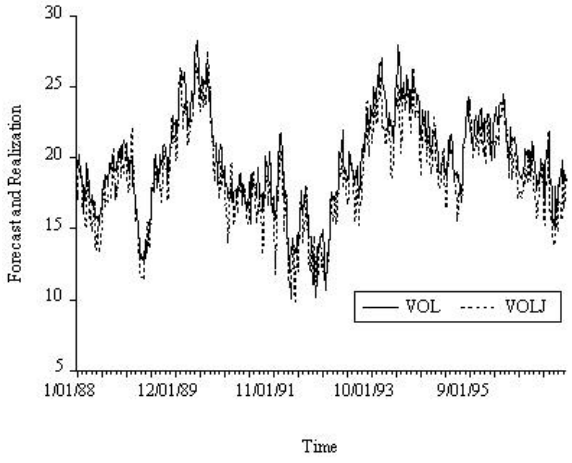


Figure 1
Shipping Volume
Quantitative Forecast and Realization
Forecast and Realization
30
25
20
15
10
VOL
VOLQ
1/01/88
12/01/89
11/01/91
10/01/93
9/01/95
Time

Figure 2
Shipping Volume
Judgmental Forecast and Realization



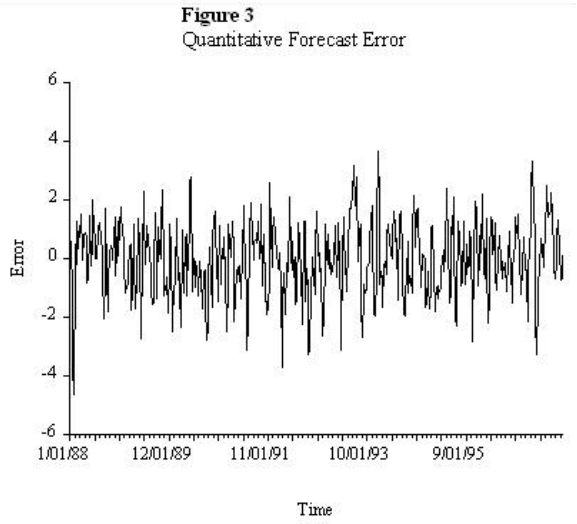


Figure 3
Quantitative Forecast Error
6
4
2
0
-2
-4
-6
Error
1/01/88
12/01/89
11/01/91
10/01/93
9/01/95
Time

Figure 4
Judgmental Forecast Error

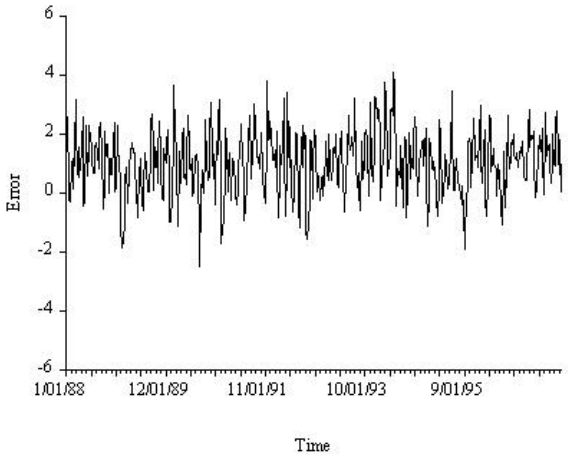


Figure 5
Histogram and Related Statistics
Quantitative Forecast Error

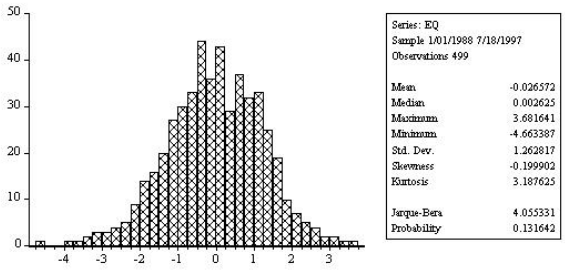


Figure 6
Histogram and Related Statistics
Judgmental Forecast Error

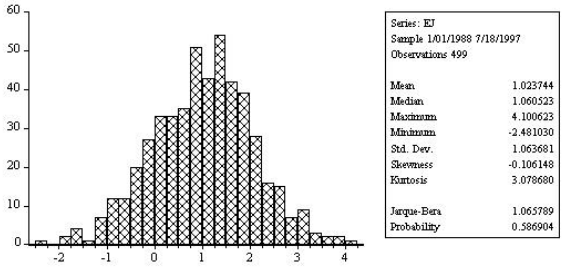
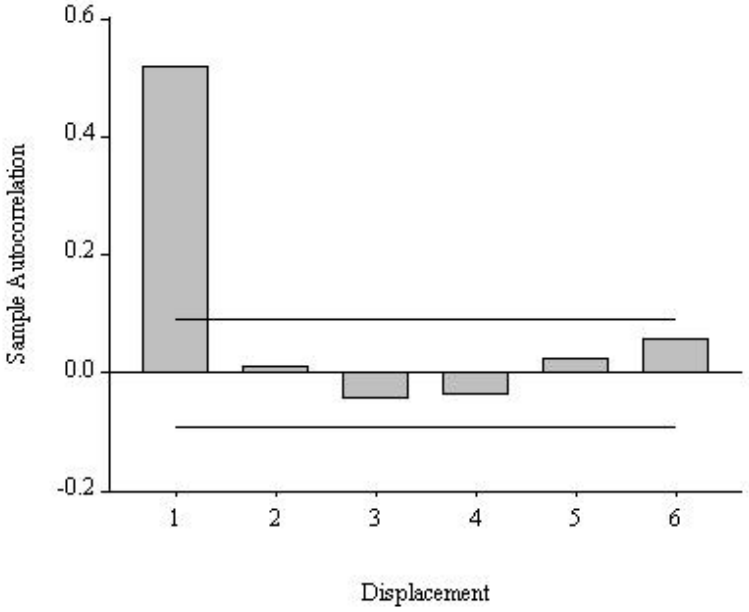


Table 1
Correlogram, Quantitative Forecast Error

Sample: 1/01/1988 7/18/1997
Included observations: 499

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.518	0.518	.045	134.62	0.000
2	0.010	-0.353	.045	134.67	0.000
3	-0.044	0.205	.045	135.65	0.000
4	-0.039	-0.172	.045	136.40	0.000
5	0.025	0.195	.045	136.73	0.000
6	0.057	-0.117	.045	138.36	0.000

Figure 7
Sample Autocorrelations and Partial Autocorrelations
Quantitative Forecast Error



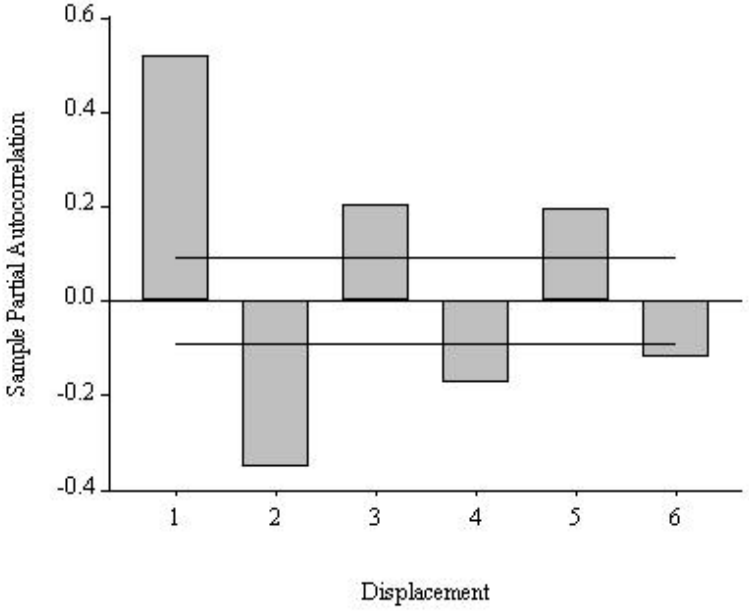
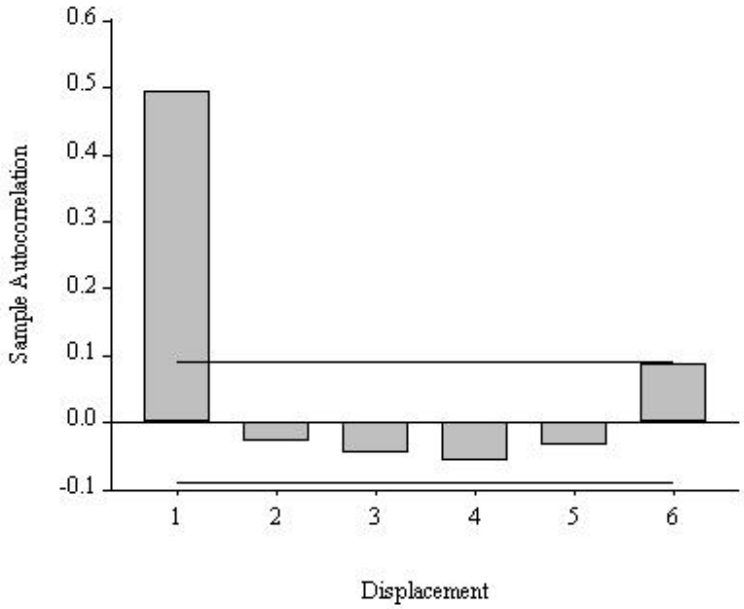


Table 2
Correlogram, Judgmental Forecast Error

Sample: 1/01/1988 7/18/1997
Included observations: 499

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.495	0.495	.045	122.90	0.000
2	-0.027	-0.360	.045	123.26	0.000
3	-0.045	0.229	.045	124.30	0.000
4	-0.056	-0.238	.045	125.87	0.000
5	-0.033	0.191	.045	126.41	0.000
6	0.087	-0.011	.045	130.22	0.000

Figure 8
Sample Autocorrelations and Partial Autocorrelations
Judgmental Forecast Error



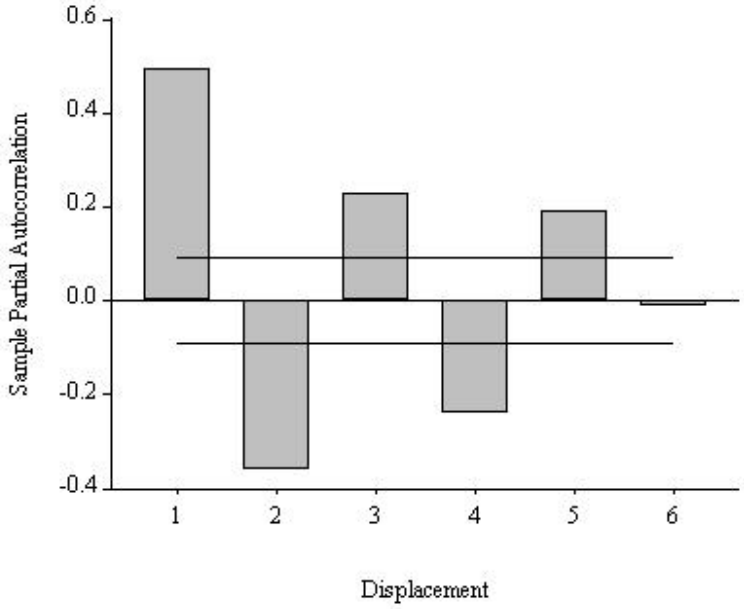


Table 3
Quantitative Forecast Error
Regression on Intercept, MA(1) Disturbances

LS // Dependent Variable is EQ
Sample: 1/01/1988 7/18/1997
Included observations: 499
Convergence achieved after 6 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.024770	0.079851	-0.310200	0.7565
MA(1)	0.935393	0.015850	59.01554	0.0000

R-squared	0.468347	Mean dependent var	-0.026572
Adjusted R-squared	0.467277	S.D. dependent var	1.262817
S.E. of regression	0.921703	Akaike info criterion	-0.159064
Sum squared resid	422.2198	Schwarz criterion	-0.142180
Log likelihood	-666.3639	F-statistic	437.8201
Durbin-Watson stat	1.988237	Prob(F-statistic)	0.000000

Inverted MA Roots -.94

Table 4
Judgmental Forecast Error
Regression on Intercept, MA(1) Disturbances

LS // Dependent Variable is EJ
Sample: 1/01/1988 7/18/1997
Included observations: 499
Convergence achieved after 7 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.026372	0.067191	15.27535	0.0000
MA(1)	0.961524	0.012470	77.10450	0.0000

R-squared	0.483514	Mean dependent var	1.023744
Adjusted R-squared	0.482475	S.D. dependent var	1.063681
S.E. of regression	0.765204	Akaike info criterion	-0.531226
Sum squared resid	291.0118	Schwarz criterion	-0.514342
Log likelihood	-573.5094	F-statistic	465.2721
Durbin-Watson stat	1.968750	Prob(F-statistic)	0.000000

Inverted MA Roots -.96

Table 5
Mincer-Zarnowitz Regression
Quantitative Forecast Error

LS // Dependent Variable is VOL
Sample: 1/01/1988 7/18/1997
Included observations: 499
Convergence achieved after 10 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.958191	0.341841	8.653696	0.0000
VOLQ	0.849559	0.016839	50.45317	0.0000
MA(1)	0.912559	0.018638	48.96181	0.0000

R-squared	0.936972	Mean dependent var	19.80609
Adjusted R-squared	0.936718	S.D. dependent var	3.403283
S.E. of regression	0.856125	Akaike info criterion	-0.304685
Sum squared resid	363.5429	Schwarz criterion	-0.279358
Log likelihood	-629.0315	F-statistic	3686.790
Durbin-Watson stat	1.815577	Prob(F-statistic)	0.000000

Inverted MA Roots -.91

Wald Test:
Null Hypothesis: C(1)=0C(2)=1

F-statistic	39.96862	Probability	0.000000
Chi-square	79.93723	Probability	0.000000

Table 6
Mincer-Zarnowitz Regression
Judgmental Forecast Error

LS // Dependent Variable is VOL
Sample: 1/01/1988 7/18/1997
Included observations: 499
Convergence achieved after 11 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.592648	0.271740	9.540928	0.0000
VOLJ	0.916576	0.014058	65.20021	0.0000
MA(1)	0.949690	0.014621	64.95242	0.0000

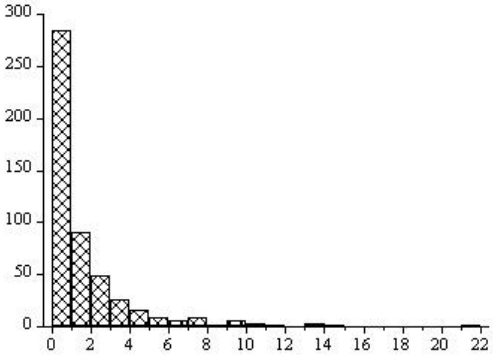
R-squared	0.952896	Mean dependent var	19.80609
Adjusted R-squared	0.952706	S.D. dependent var	3.403283
S.E. of regression	0.740114	Akaike info criterion	-0.595907
Sum squared resid	271.6936	Schwarz criterion	-0.570581
Log likelihood	-556.3715	F-statistic	5016.993
Durbin-Watson stat	1.917179	Prob(F-statistic)	0.000000

Inverted MA Roots -.95

Wald Test:
Null Hypothesis: C(1)=0C(2)=1

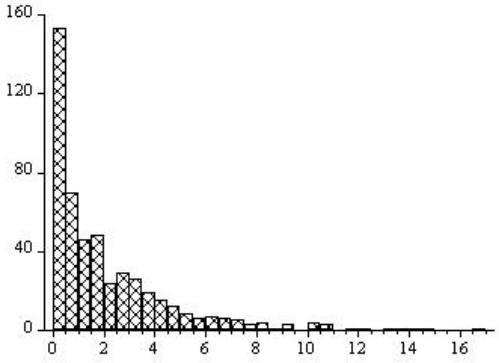
F-statistic	143.8323	Probability	0.000000
Chi-square	287.6647	Probability	0.000000

Figure 9
Histogram and Related Statistics
Squared Quantitative Forecast Error



Series: EQSQ	
Sample 1/01/1988 7/18/1997	
Observations 499	
Mean	1.592217
Median	0.763750
Maximum	21.74718
Minimum	5.61E-06
Std. Dev.	2.369751
Skewness	3.293315
Kurtosis	18.88079
Jarque-Bera	6145.666
Probability	0.000000

Figure 10
Histogram and Related Statistics
Squared Judgmental Forecast Error



Series: EJSQ	
Sample 1/01/1988 7/18/1997	
Observations 499	
Mean	2.177201
Median	1.308296
Maximum	16.81511
Minimum	4.63E-05
Std. Dev.	2.623644
Skewness	2.134551
Kurtosis	8.646748
Jarque-Bera	1041.891
Probability	0.000000

Figure 11
Loss Differential

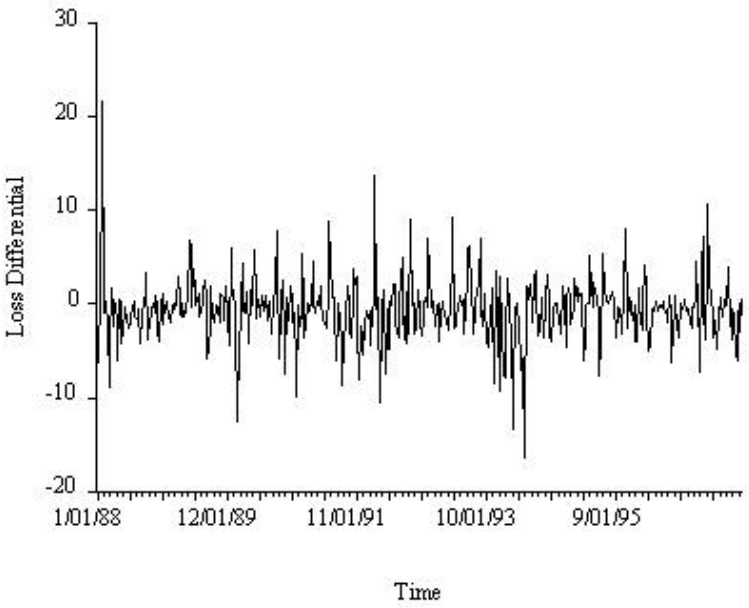


Figure 12
Histogram and Related Statistics
Loss Differential

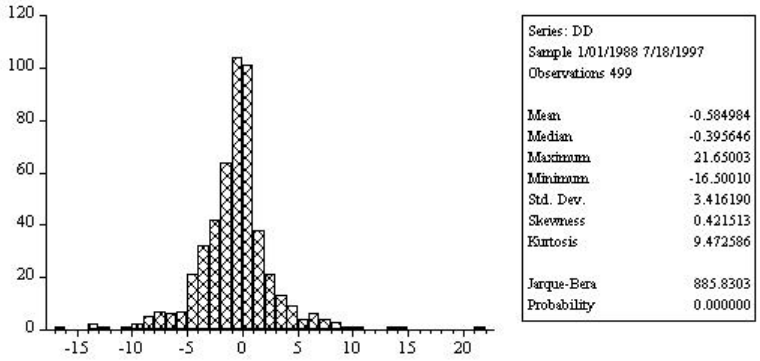
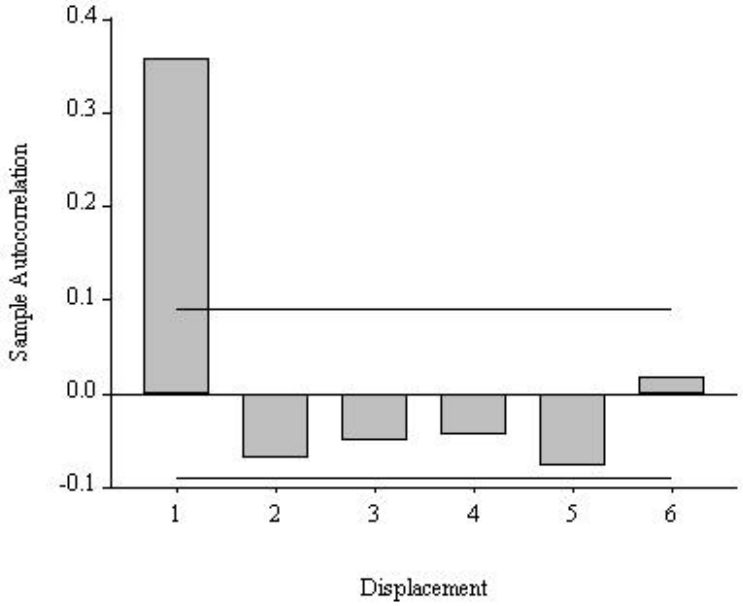


Table 7
Loss Differential Correlogram

Sample: 1/01/1988 7/18/1997
Included observations: 499

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	0.357	0.357	.045	64.113	0.000
2	-0.069	-0.226	.045	66.519	0.000
3	-0.050	0.074	.045	67.761	0.000
4	-0.044	-0.080	.045	68.746	0.000
5	-0.078	-0.043	.045	71.840	0.000
6	0.017	0.070	.045	71.989	0.000

Figure 13
Sample Autocorrelations and Partial Autocorrelations
Loss Differential



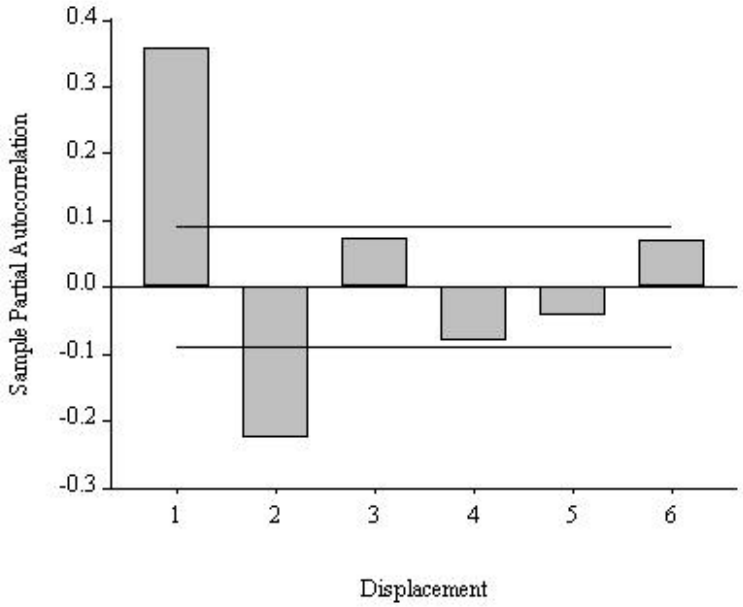


Table 8
Loss Differential
Regression on Intercept with MA(1) Disturbances

LS // Dependent Variable is DD
Sample: 1/01/1988 7/18/1997
Included observations: 499
Convergence achieved after 4 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.585333	0.204737	-2.858945	0.0044
MA(1)	0.472901	0.039526	11.96433	0.0000

R-squared	0.174750	Mean dependent var	-0.584984
Adjusted R-squared	0.173089	S.D. dependent var	3.416190
S.E. of regression	3.106500	Akaike info criterion	2.270994
Sum squared resid	4796.222	Schwarz criterion	2.287878
Log likelihood	-1272.663	F-statistic	105.2414
Durbin-Watson stat	2.023606	Prob(F-statistic)	0.000000

Inverted MA Roots -.47

10.4 Exercises, Problems and Complements

1. Forecast evaluation in action.

   Discuss in detail how you would use forecast evaluation techniques to address each of the following questions.

   a. Are asset returns (e.g., stocks, bonds, exchange rates) forecastable over long horizons?

   b. Do forward exchange rates provide unbiased forecasts of future spot exchange rates at all horizons?

   c. Are government budget projections systematically too optimistic, perhaps for strategic reasons?

   d. Can interest rates be used to provide good forecasts of future inflation?

2. Forecast error analysis.

   You work for a London-based hedge fund, Thompson Energy Investors, and your boss has assigned you to assess a model used to forecast U.S. crude oil imports. On the last day of each quarter, the model is used to forecast oil imports at horizons of 1-quarter-ahead through 4-quarters-ahead. Thompson has done this for each of 80 quarters and has kept the corresponding four forecast error series, which appear on the book's web page.

   a. Based on a correlogram analysis, assess whether the 1-quarter-ahead forecast errors are white noise. (Be sure to discuss all parts of the correlogram: sample autocorrelations, sample partial autocorrelations, Bartlett standard errors and Ljung-Box statistics.) Why care?

   b. Regress each of the four forecast error series on constants, in each case allowing for a $MA(5)$ disturbances. Comment on the significance of

the MA coefficients in each of the four cases and use the results to assess the optimality of the forecasts at each of the four horizons. Does your 1-step-ahead $MA(5)$-based assessment match the correlogram-based assessment obtained in part a? Do the multi-step forecasts appear optimal?

c. Overall, what do your results suggest about the model's ability to predict U.S. crude oil imports?

3. The mechanics of practical forecast evaluation.

For the following, use the time series of shipping volume, quantitative forecasts, and judgmental forecasts used in this chapter.

a. Replicate the empirical results reported in this chapter. Explore and discuss any variations or extensions that you find interesting.

b. Using the first 250 weeks of shipping volume data, specify and estimate a univariate autoregressive model of shipping volume (with trend and seasonality if necessary), and provide evidence to support the adequacy of your chosen specification.

c. Use your model each week to forecast two weeks ahead, each week estimating the model using all available data, producing forecasts for observations 252 through 499, made using information available at times 250 through 497. Calculate the corresponding series of 248 2-step-ahead recursive forecast errors.

d. Using the methods of this chapter, evaluate the quality of your forecasts, both in isolation and relative to the original quantitative and judgmental forecasts. Discuss.

4. Forecasting Competitions.

There are many forecasting competitions. Kaggle.com, for example, is a well-known online venue. Participants are given a "training sample" of

data and asked to forecast a "test sample"; that is, to make an out-of-sample forecast of hold-out data, which they are not shown

(a) Check out Kaggle. Also read "A Site for Data Scientists to Prove Their Skills and Make Money," by Claira Cain Miller, *New York Times*, November 3, 2011. What's good about the Kaggle approach? What's bad? What happened to Kaggle since its launch in 2011?

(b) "Kaggle competitions" effectively outsource forecasting. What are pros and cons of in-house experts vs. outsourcing?

(c) Kaggle strangely lets people peek at the test sample by re-submitting forecasts once per day.

(d) Kaggle scores extrapolation forecasts rather than h-step. This blends apples and oranges.

(e) Kaggle is wasteful from a combining viewpoint. One doesn't just want to find the "winner."

5. The Peso Problem.

Suppose someone assigns a very high probability to an event that fails to occur, or a very low probability to an event that does occur. Is the person a bad probability forecaster? The answer is perhaps, but not at all necessarily. Even events *correctly* forecast to occur with high probability may simply fail to occur, and conversely.

Thus, for example, a currency might sell forward at a large discount, indicating that the market has assigned a high probability of a large depreciation. In the event, that depreciation might fail to occur, but that does not necessarily mean that the market was in any sense "wrong" in assigning a high depreciation probability. The term "**Peso problem**" refers to exactly such issues in a long-ago situation involving the Mexican Peso.

6. Measuring forecastability with canonical correlations.

   One can measure forecastability via canonical correlation between "past" and "future," as in Jewell and Bloomfield 1983, Hannan and Poskitt 1988.

7. Forecast Evaluation When Realizations are Unobserved.

   Sometimes we never see the realization of the variable being forecast. This occurs for example in forecasting ultimate resource recovery, such as the total amount of oil in an underground reserve. The actual value, however, won't be known until the reserve is depleted, which may be decades away. Such situations obviously make for difficult accuracy evaluation!

   If the resource recovery example sounds a bit exotic, rest assured that it's not. In volatility forecasting, for example, "true" volatility is never observed. And in any sort of state-space model, such as a dynamic factor model, the true state vector is never observed. (See Chapters ***.)

   (a) Nordhaus tests.

      – Some optimality tests can be obtained even when the forecast target is unobservable (Patton and Timmermann 2010, building on Nordhaus 1987). In particular, (1) forecast revisions (for fixed target date) should be MDS, and (2) forecast variance (not error variance, but forecast variance) should decrease with distance from terminal date. PT 2010 also have a nice generalized MZ test that builds on these ideas.

   (b) Patton tests.

8. Nonparametric predictabilty assessment.

   We presented our autoregressive modeling approach as a parametric method. However, in general, we need not assume that the fitted autore-

gression is the true data-generating process; rather, it may be considered an approximation, the order of which can grow with sample size. Thus the autoregressive model can be viewed as a sieve, so our approach actually *is* nonparametric.

Nevertheless, the sieve approach has a parametric flavor. For any fixed sample size, we assess predictability through the lens of a particular autoregressive model. Hence it may be of interest to develop an approach with a more thoroughly nonparametric flavor by exploiting Kolmogorov's well-known spectral formula for the univariate innovation variance,

$$\sigma^2 = \exp(\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln 2\pi f(\omega) d\omega),$$

where f is the spectral density function. Kolmogorov's result has been extended to univariate h-step-ahead forecast error variances by Bhansali (1992).

9. Can unskilled density forecasters successfully disguise themselves as skilled?

10. Cross section forecast evaluation.

    Most of the basic lessons for time-series forecast evaluation introduced in this chapter are also relevant for cross-section forecast evaluation. Cross-section forecast errors (appropriately standardized if heteroskedasticity is present) should be iid white noise over space, and unpredictable using any available covariates. DM-type tests can be done for point forecasts, and DGT-type test for density forecasts.

11. Turning point forecasts into density forecasts.

    As we have shown, Mincer-Zarnowitz corrections can be used to "correct" sub-optimal point forecasts. They can also be used to produce density forecasts, by drawing from an estimate of the density of the MZ regression disturbances, as we did in a different context in section 4.1.

10.5 Notes