# The Theory of Forecasting

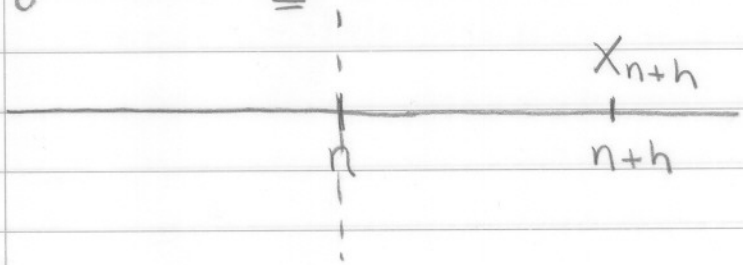
(Chapter 4, "Forecasting Economic Variables" by Granger and Newbold 1986)

## Some Basic Concepts

- **Information Set**

Let $\{X_t\}$ be some discrete-time stochastic process, which for the time will be assumed to be stationary

We are at time $n$ (≡ now) and wish to forecast $h$ time units ahead $(n+h)$.




$I_n$ ≡ information available at time $n$.

* univariate information set

A sample of previous values of the series $X_{n-j}$, $j = 0, 1, \ldots N$

\+

Some properties of the stochastic process $\{X_t\}$, for instance, $E(X_t) = 0$ and stationarity

* Multivariate information set

A sample of previous values of the series $X_{n-j}$, $Y_{n-j}$, $Z_{n-j}$, .. $j = 0, 1, \ldots N$

\+ Similar

- Conditional Variables

Since the variable to be forecasted $X_{n+h}$ is a random variable, it can be fully characterized only in terms of a probability density. However, since the information set $I_n$ is to be utilized one needs to use a conditional density function, i.e.,

$$\text{Prob}\left(x < X_{n+h} \leq x + dx \mid I_n\right) = g_c(x)\,dx$$

If $g_c(x)$ is available other properties of $X_{n+h}$ could be immediately determined.
If $g_c(x)$ is not available we attempt to find some confidence band for $X_{n+h}$ or some single value called a point forecast.

- Cost function

To obtain any kind of best value for a point forecast, one requires a criterion against which various alternatives can be judged.

$$C(e) \quad \text{with} \quad C(0) = 0$$

and

$$e_{n,h} = X_{n+h} - f_{n,h}$$

$f_{n,h}$ is a point forecast of $X_{n+h}$, based on $I_n$.

Goal: To choose $f_{n,h}$ such that

$$\min_{\{f_{n,h}\}} E_c\{C(e_{n,h})\} \qquad (1)$$

Example: $C(e) = ae^2$

We will prove that $f_{n,h} = E_c(X_{n+h})$

- Linear Forecasts

In practice one will rarely know the conditional density function sufficiently well for a complete solution of (1) to be possible. For this reason we put restrictions on the form of forecast to be considered; in particular by assuming that

> $f_{n,h}$ is a linear function of the data available in $I_n$.

## Generalized Cost functions : $C(e)$

(i) $C(0) = 0$

(ii) monotonic non decreasing for $e > 0$
i-e $C(e_1) \geq C(e_2)$ for $e_1 > e_2$

(iii) monotonic non increasing for $e < 0$

Note that $C(e)$ doesn't need to be symmetrical

Let $g_{c,h}(x)$ be the conditional probability density function of $X_{n+h} | I_n$, then the required optimal forecast $f_{n,h}$, which will be a function of only $I_n$, will be found by choosing $f_{n,h}$ such that minimizes

$$\boxed{\mathcal{I} = \int_{-\infty}^{\infty} C(x - f_{n,h})\, g_{c,h}(x)\, dx} \qquad (2)$$

Example $C(e) = a e^2$ with $a > 0$

$$J = \int_{-\infty}^{\infty} a (x - f_{n,h})^2 g_{c,h}(x)\, dx \qquad (3)$$

Define $M_h = E_c[X_{n+h} \mid I_n] = \int_{-\infty}^{\infty} x\, g_{c,h}(x)\, dx$

After some algebraic manipulation (3) can be written as

$$J = a (M_h - f_{n,h})^2 + a \int_{-\infty}^{\infty} (x - M_h)^2 g_{c,h}(x)\, dx$$

Then

$$f_{n,h} = M_h = E_c\{X_{n+h} \mid I_n\} \qquad (4)$$

minimizes $J$.

Question: Is the predictor given by (4) is optimal for a wider class of cost functions than just the quadratic function?

Answer:

Theorem: The optimal predictor of $X_{n+h}$ given $I_n$ is $f_{n,h} = M_h$ if

(i) $C(e)$ is symmetric about $e=0$, the derivative $C'(e)$ exists almost everywhere and is strictly monotonically increasing on the whole range $-\infty < e < \infty$, an also $g_{c,h}(x)$ is symmetric about $x = M_h$

or

(ii) $C(e)$ is symmetric about $e=0$ and $g_{c,h}(x)$, symmetric about $x = M_h$, is continuous and unimodal

We could accept that $g_{c,h}(x)$ is symmetric around $x = M_h$, but what is harder to accept is that $C(e)$ is symmetric (Think of some examples of non-symmetric cost functions)

## Optimal Prediction under Asymmetric Cost Functions

**Theorem** : If the conditional distribution is assumed to be normal, then the optimal predictor is given by

$$\boxed{F_{n,h} = M_h + \alpha}$$

where $\alpha$ depends only on the cost function being used and not on $I_n$.

Example : LIN-LIN cost function

$$C(e) = ae, \quad e > 0, \quad a > 0$$
$$= 0 \qquad e = 0$$
$$= be, \quad e < 0, \quad b < 0$$

Then from (2), the expected cost with predictor $f \equiv f_{n,h}$ is

$$J = E_c\{C(X_{n+h} - f) \mid I_n\} = a\int_f^{\infty} (x-f)\, g_{c,h}(x)\, dx + b\int_{-\infty}^{f} (x-f)\, g_{c,h}(x)\, dx$$

Differentiating this with respect to f and equating to zero to find the minimum expected cost gives

$$G_{c,h}(f) = \frac{a}{a-b}$$

where $G_{c,h}(x)$ is the conditional cumulative distribution function of $X_{n+h} \mid I_n$, i.e.

$$G_{c,h}(x') = \int_{-\infty}^{x'} g_{c,h}(x)\, dx$$

In the symmetric case $a = -b$, and so the optimal predictor $f \equiv f_{n,h}$ will be given by $G_{c,h}(f) = \frac{1}{2}$, so that f is the median of $g_{c,h}(x)$.

# The Evaluation of Forecasts

An objective evaluation of a set of forecasts might seek to answer one or more of the following three questions:

(a) Is one set of forecasts better than its competitors?

(b) How "good", in some sense, is a particular set of forecasts?

(c) Can the forecast-generating mechanism be modified in some way so as to yield improved forecast performance?

We will only focus in (a).

$X_t, \quad t = 1, 2, \ldots N$

$f_t, \quad t = 1, 2, \ldots N$

$e_t = X_t - f_t \qquad t = 1, 2, \ldots N$

$$D_N^2 = \frac{1}{N} \sum_{t=1}^{N} e_t^2$$ (expected squared forecast error estimated)

Suppose now two competing forecasting procedures that produce errors $e_t^{(1)}$, $e_t^{(2)}$, $t=1,\ldots N$

Question: How can we determine whether one procedure performed significantly better than the other under the usual criteria of statistical significance?

Suppose that $(e_t^{(1)}, e_t^{(2)})$, $t=1,2,\ldots N$ constitutes a random sample from a bivariate normal distribution with means zero, variances $\sigma_1^2$ and $\sigma^2$, and correlation coefficient $\rho$. In particular it is assumed that $E(e_t^{(1)}) = E(e_t^{(2)}) = 0$ and

$$\mathrm{Cov}(e_t^{(1)} e_{t-j}^{(1)}) = 0 = \mathrm{Cov}(e_t^{(2)} e_{t-j}^{(2)}) \quad j \geq 0$$

$H_0: \sigma_1^2 = \sigma_2^2$

$H_a: \sigma_1^2 \neq \sigma_2^2$

Construct the pair of random variables $\boxed{e^{(1)} + e^{(2)}}$ and $\boxed{e^{(1)} - e^{(2)}}$. Now

$$E\left[(e^{(1)} + e^{(2)})(e^{(1)} - e^{(2)})\right] = \sigma_1^2 - \sigma_2^2$$

So how would you test the hypothesis??

## Comparison of forecasts with more general assumptions

("Comparing Predictive Accuracy" by Diebold and Mariano, JBES 1995)

This paper allows for forecast errors that are potentially non-gaussian, nonzero mean, serially correlated, and contemporaneously correlated.

$\{y_t\}_{t=1}^{T}$ Observed time series

$\{\hat{y}_t^1\}_{t=1}^{T}$ and $\{\hat{y}_t^2\}_{t=1}^{T}$ two different forecasts of $\{y_t\}_{t=1}^{T}$

$\{\hat{e}_t^1\}_{t=1}^{T}$ and $\{\hat{e}_t^2\}_{t=1}^{T}$ associated forecast errors

$g(y_t, \hat{y}_t^i) = g(\hat{e}_t^i)$ Cost function or loss function

$d_t = g(\hat{e}_t^1) - g(\hat{e}_t^2)$

$H_0$: Equal forecast accuracy

$H_1$: Different forecast accuracy

$$H_0: \quad E[d_t] = E\left[g(\hat{e}_t^1) - g(\hat{e}_t^2)\right] = 0$$

vs

$$H_1: \quad E[d_t] \neq 0$$

## Test statistic

If $\{d_t\}_{t=1}^{T}$ is assumed to be covariance stationary and short memory then

$$\sqrt{T}(\bar{d} - \mu) \xrightarrow{d} N(0, 2\pi f_d(0))$$

$2\pi f_d(0)$ LRV$_d$

where $\bar{d} = \frac{1}{T}\sum_{t=1}^{T}\left(g(\hat{e}_t^1) - g(\hat{e}_t^2)\right)$

$$f_d(0) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\gamma_d(\tau)$$

$$S = \frac{\bar{d}}{\sqrt{\frac{2\pi f_d(0)}{T}}} \sim N(0,1)$$

So we reject $H_0$ at 5% if

$$|S| > 1.96$$

We could also do one sided tests.

LRV = Long run variance $\left(\lim_{n\to\infty} Var\left(\sqrt{n}\bar{y}\right)\right)$ (if it exists)

Let $y_t = \mu + \psi(L)\varepsilon_t$, $\varepsilon_t \sim iid(0, \sigma^2)$

LRV is the variance of the A.D of

$$\sqrt{T}(\bar{y} - \mu) \xrightarrow{d} A.D(0, LRV)$$

with $\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t$

Non-parametric estimator

$$LRV = \sigma^2 \psi^2(1) = \sum_{-\infty}^{\infty} \gamma_k$$

$$= \gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k \quad \text{since } \gamma_k = \gamma_{-k}$$

Naive estimator

$$\widehat{LRV} = \hat{\gamma}_0 + 2\sum_{k=1}^{M_T} \hat{\gamma}_k \; ; \; M_T = \text{truncation point}$$

This estimator is inconsistent

Newey-West propose an estimator which downweights $\hat{\gamma}_k$ for large $k$

$$\widehat{LRV}_{NW} = \hat{\gamma}_0 + 2 \sum_{j=1}^{M_T} \left[1 - \frac{j}{M_{T+1}}\right] \hat{\gamma}_j$$

with $M_T \to \infty$ as $T \to \infty$

$$\frac{M_T}{T} \to 0$$

Bartlett's Window

Example $M_T = 4 \left(\frac{T}{100}\right)^{1/4}$

Then if $T = 100$ $M_T = 4$

and

$$\widehat{LRV}_{NW} = \hat{\gamma}_0 + 2 \sum_{j=1}^{4} \left[1 - \frac{j}{5}\right] \hat{\gamma}_j$$

$$= \hat{\gamma}_0 + 2 \left[\frac{4}{5} \hat{\gamma}_1 + \frac{3}{5} \hat{\gamma}_2 - \cdots\right.$$

Alternative Estimator (based on the AR representation)

Assume $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$

$\varepsilon_t \sim iid(0, \sigma^2)$

① Estimate $\hat{\phi}_1, \hat{\phi}_2, \ldots \hat{\phi}_p, \hat{\sigma}^2$

② $$\hat{\psi}(1) = \frac{1}{1 - \hat{\phi}_1 - \hat{\phi}_2 - \cdots \hat{\phi}_p}$$

③ $$\widehat{LRV}_{AR} = \hat{\sigma}^2 \left(\hat{\psi}(1)\right)^2$$

# More on LRV estimation

[From "Dynamic Nonlinear Econometric Analysis" by Pötscher and Prucha (1996)]

$$\sqrt{T}(\bar{Y}-\mu) \xrightarrow{d} AD(0, LRV)$$

$$LRV = \sum_{-\infty}^{\infty} \gamma_k = \gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k$$

Define $v_t = y_t - \mu$, then

$$LRV = E(v_t^2) + 2\sum_{j=1}^{\infty} E(v_t v_{t+j})$$

## How do we estimate it?

Some candidates:

(a) $$\frac{1}{n}\sum_{1}^{n} v_t^2 + 2\sum_{j=1}^{n-1} \frac{1}{n-j} \sum_{t=1}^{n-j} v_t v_{t+j}$$

This "naive" estimator is well-known to be inconsistent.

Intuitively speaking, the reason for the inconsistency is that the estimator is a sum of $(n-j)$ terms, each with a variance roughly the order $O\left(\frac{1}{n-j}\right)$. The variance of the estimator is now roughly "n" times as large ($O(1)$).

$$(\bullet\bullet)\quad \frac{1}{n}\sum_{1}^{\hat{}} v_t^2 + 2\sum_{j=1}^{n-1}\frac{1}{n}\sum_{t=1}^{n-j} v_t v_{t+j}$$

The reason of this estimator is that one possibility to obtain consistent estimators of the LRV is to reduce the variance of the estimator by excluding some of the sample moments

$$\frac{1}{n-j}\sum_{t=1}^{n-j} v_t v_{t+j}$$

from the formula of the "naive" estimator. It seems natural to exclude or down-weigh the sample moments corresponding to lags $\underline{j}$ close to $\underline{n}$.

This achieved by introducing weights into the formula for the "naive" estimator.

Down-weighing of the sample covariance has the effect of reducing the variance of the estimator at the expenses of introducing a bias.

The above estimator $(\bullet\bullet)$ is still inconsistent although some moderate down-weighing of the sample moments

$$\frac{1}{n-j}\sum_{t=1}^{n-j} v_t v_{t+j} \text{ takes place}$$

$$(\bullet\bullet\bullet)\quad \frac{1}{n}\sum_{t=1}^{n} v_t^2 + 2\sum_{j=1}^{M}\frac{1}{n}\sum_{t=1}^{n-j} v_t v_{t+j}$$

$$\downarrow P$$

$$E(v_t^2) + 2\sum_{t=1}^{M} E(v_t v_{t+j})$$

that is equal to LRV only if $E(v_t v_{t+j}) = 0$ for $j > M$ (for example if $v_t$ is m-dependent with $m \leq M$)

Because we are always assuming $E(v_t v_{t+j}) \to 0$ for $j \to \infty$, then it is clear that the bias will be smaller the larger $\underline{\underline{M}}$ is.

So a possible solution is to make $\underline{\underline{M}}$ sample size depend $\underline{\underline{M_n}}$ such that $M_n \to \infty$ to avoid bias, but slowly enough such that the variance still goes to zero as $n \to \infty$.

$$(\bullet\bullet\bullet\bullet)\quad \frac{1}{n}\sum_{t=1}^{n} v_t^2 + 2\sum_{j=1}^{M_n}\frac{1}{n}\sum_{t=1}^{n-j} v_t v_{t+j}$$

And more general

(• • • • •)

$$LRV_n = \frac{1}{n}\sum V_t^2 + 2\sum_{j=1}^{n-1} W(j,n)\frac{1}{n}\sum_{t=1}^{n-j} V_t V_{t+j}$$

for appropiated weights

$$W(j,n) \in \mathbb{R}$$

Clearly (• • • •) is a special case because it can be obtained with

$$W(j,n) = \begin{cases} 1 & 0 \le j \le M_n \\ 0 & M_n < j \le n-1 \end{cases}$$

Another example is

$$W(j,n) = \begin{cases} 1 - \frac{j}{M_n} & 0 \le j \le M_n \\ 0 & M_n < j \le n-1 \end{cases}$$

(Bartlett kernel)

# Combination of Forecasts

[Following Bates and Granger (1969)]

Two unbiased one-step ahead forecasts

$f_n^{(1)}$ $f_n^{(2)}$

Forecast errors $e_n^{(j)} = X_n - f_n^{(j)}$ $(j = 1, 2)$

$$E[e_n^{(j)}] = 0$$

$$E[e_n^{(j)}]^2 = \sigma_j^2 \quad j = 1, 2$$

and

$$E(e_n^{(1)} e_n^{(1)}) = \rho \sigma_1 \sigma_2$$

Consider now a combined forecast

$$\boxed{C_n = k f_n^{(1)} + (1-k) f_n^{(2)}}$$

$$e_n^{(c)} = X_n - C_n = k e_n^{(1)} + (1-k) e_n^{(2)}$$

Hence the error variance is

$$\boxed{\sigma_c^2 = k^2 \sigma_1^2 + (1-k)^2 \sigma_2^2 + 2k(1-k)\rho \sigma_1 \sigma_2} \quad (*)$$

This expression is minimized for the value of $k$ given by

$$\boxed{k_0 = \frac{\sigma_2^2 - \rho \sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho \sigma_1 \sigma_2}} \quad (**)$$

Substituting in the error variance expression:

$$\sigma_{c,0}^2 = \frac{\sigma_1^2 \sigma_2^2 (1-\rho^2)}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}$$

Note that $\sigma_{c,0}^2 < \min(\sigma_1^2, \sigma_2^2)$ unless $\rho = \frac{\sigma_1}{\sigma_2}$ or $\frac{\sigma_2}{\sigma_1}$

If either equality holds, then the variance of the combined forecast is equal tho the smaller of the two error variances. In most practical situations the best available combined forecast will outperform the better individual forecast.

From (**) one can obtain two "extreme" interesting results:

$$k_0 \gtreqless 0 \quad \text{iff} \quad \sigma_2/\sigma_1 \gtreqless \rho$$

If $f_n^{(2)}$ is the optimal forecast ($k_0 = 0$) based on a particular information set, any other forecast $f_n^{(1)}$ based on the same information set must be such that $\rho = \frac{\sigma_2}{\sigma_1}$, exactly.

The case $k_0 < 0$ is also interesting, think why?

The second "extreme" result is when in (*) $\rho \longrightarrow \begin{matrix} -1 \\ \text{or} \\ +1 \end{matrix}$. In the former case $\sigma^2_{c,o} \longrightarrow 0$, implying a perfect forecast

As $\rho \longrightarrow 1$, $\sigma^2_{c,o}$ also $\longrightarrow 0$ except when $\sigma_1 = \sigma_2$ in which case its limit is $\sigma_1^2$. Try to interpret this result.

## Some extra comments

(1) Expression (*,*) is not very useful because the involved parameters are unknown.

Think on how to estimate the weights in $C_n$

(2) Time varying weights.