

1. Forecasting basics – notation and population results

Suppose that you are interested in forecasting the value of y_{t+h} using a vector of variables x_t , where x_t might (and usually does) contain current and lagged values of y_t as well as current and lagged values of other variables. Let $f_{t+h/t}$ denote the forecast.

The forecast error is $e_{t+h/t} = y_{t+h} - f_{t+h/t}$.

Let $L(e_{t+h/t})$ denote the loss associated with the forecast error $e_{t+h/t}$. A typical loss function is $L(e_{t+h/t}) = e_{t+h/t}^2$ so that the loss is quadratic.

Optimal Forecasts

If the loss is quadratic, then $f_{t+h|t} = \mu_y(x_t) = E(y_{t+h} | x_t)$ is the forecast that minimizes

$$E(L(e_{t+h|t}) | x_t)$$

To prove this, let $f(x_t) = \mu_y(x_t) + g(x_t)$ and our goal is to show that the optimal value of $g(x_t)$ is 0. Note

$$(y_{t+h} - f(x_t))^2 = (y_{t+h} - \mu_y(x_t))^2 - 2g(x_t)(y_{t+h} - \mu_y(x_t)) + g(x_t)^2$$

and $g(x_t)$ appears in the last two terms on the rhs. Since (i)

$$E_{Y|X}[g(x_t)(y_{t+h} - \mu_y(x_t))] = g(x_t)\{E_{Y|X} y_{t+h} - \mu_y(x_t)\} = 0, \text{ and (ii) } g(x_t)^2 \geq 0, \text{ then the}$$

optimal $g(x_t) = 0$.

If the loss is $L(e_{t+h|t}) = |e_{t+h|t}|$, then $f_{t+h|t} = \text{med}(y_{t+h} | x_t)$ is the forecast that minimizes $E(L(e_{t+h|t}))$. To simplify notation, drop the t subscripts and let $h(y|x)$ denote the pdf of y given x . Let f denote the forecast. The goal is to find f to minimize $\int |y - f| h(y|x) dy = \int_{-\infty}^f (f - y) h(y|x) dy + \int_f^{\infty} (y - f) h(y|x) dy$. The first order conditions are $\int_{-\infty}^f h(y|x) dy = \int_f^{\infty} h(y|x) dy$ which yields the result.

There are a few general results for more general loss functions (Diebold, F.X. and P. Christoffersen (1997), "Optimal Prediction Under Asymmetric Loss," *Econometric Theory*, 13m 808-817 has a nice discussion.)

I will concentrate on results for quadratic loss

Some properties of optimal forecasts:

$E(e_{t+h/t} | x_t) = 0$ which implies

$$E(e_{t+h/t} x_t) = 0.$$

$e_{t+h/t}$ follows an $MA(h-1)$ process if x_t includes past values of y .

$$\sigma_y^2 = \sigma_f^2 + \sigma_e^2 \text{ which implies } \sigma_y^2 \geq \sigma_f^2.$$

Forecast combining.

Suppose $f_{t+h/t}^1$ and $f_{t+h/t}^2$ are two forecasts of y_{t+h} . Then an improved forecast can be constructed as $f_{t+h/t}^c = \beta_1 f_{t+h/t}^1 + \beta_2 f_{t+h/t}^2$ where β_1 and β_2 are the (population) linear regression coefficients from the regression of y_{t+h} onto $f_{t+h/t}^1$ and $f_{t+h/t}^2$.

Suppose that $f_{t+h/t}^1$ is an optimal forecast and $f_{t+h/t}^2$ is any forecast that uses information available in the construction of $f_{t+h/t}^1$. Then $\beta_1 = 1$ and $\beta_2 = 0$.

Evaluating Forecasts

Consider two competing forecasts, f_1 and f_2 with associated forecast errors e_1 and

e_2 . (I have suppressed the time subscripts for notational convenience.) Let $\sigma_1^2 = E(e_1^2)$

and $\sigma_2^2 = E(e_2^2)$

Loss Function Tests

Loss functions tests compare $E(L(e_1))$ and $E(L(e_2))$. For quadratic loss, consider the null of equal forecasting ability $H_0 : \sigma_1^2 = \sigma_2^2$ versus a one-sided or two sided alternative. Suppose that the stochastic process for e_1 and e_2 is sufficiently well-behaved so that

$$\frac{1}{\sqrt{T}} \sum \begin{pmatrix} e_{1,t}^2 - \sigma_1^2 \\ e_{2,t}^2 - \sigma_2^2 \end{pmatrix} \xrightarrow{d} N(0, V)$$

and write

$$e_{1,t}^2 - e_{2,t}^2 = \mu + a_t$$

where $\mu = \sigma_1^2 - \sigma_2^2$ and

$$a_t = (e_{1,t}^2 - \sigma_1^2) - (e_{2,t}^2 - \sigma_2^2).$$

Then

$$\sqrt{T}(\hat{\mu} - \mu) \xrightarrow{d} N(0, v),$$

where $v = V_{11} + V_{22} - 2V_{12}$, so that

$$\hat{\mu} \overset{a}{\sim} N\left(\mu, \frac{\hat{v}}{T}\right)$$

where \hat{v} is a consistent estimator of v . This approximation can be used to test the null and to construct confidence intervals for the difference in the risks or, using the delta-method, in the relative risk of the two forecasts.

Forecast Combining Tests

Suppose that f_1 dominates f_2 in the sense that $E(y | f_1, f_2) = f_1$, then when f_1 and f_2 are combined, the weight on f_1 will equal 1 and the weight on f_2 will equal 0.

Consider the combining regression described above, but with $\beta_1 + \beta_2 = 1$ imposed.

This can be written as

$$y_{t+h} = (1-\alpha)f_{1,t+h|t} + \alpha f_{2,t+h|t} + e_{c,t+h|t}$$

Rearranging, using the definition of the forecast errors, and simplifying notation by dropping the time subscripts yields

$$e_1 = \alpha(f_2 - e_1) + e_c.$$

When f_1 dominates f_2 then $E(y | f_1, f_2) = f_1$, so that $E(e_1 | f_1, f_2) = 0$. Let

$a = f_2 - f_1$. (Alternatively, by adding and subtracting y , $a = e_1 - e_2$.) Note that

$E(e_1 a) = 0$, follows from $E(e_1 | f_1, f_2) = 0$. Thus, the combining regression, under the

null that f_1 is optimal, becomes

$$e_1 = \alpha(a) + e_c$$

where $\alpha = 0$ (because a and e_1 are uncorrelated), so that $e_c = e_1$. Suppose that

$$\frac{1}{T} \sum a_t^2 \xrightarrow{p} \sigma_a^2 > 0$$

and

$$\frac{1}{\sqrt{T}} \sum e_{1,t} a_t \xrightarrow{d} N(0, q)$$

Then

$$\sqrt{T}(\hat{\alpha} - \alpha) \xrightarrow{d} N\left(0, \frac{q}{\sigma_a^4}\right)$$

so that

$$\hat{\alpha}^a \sim N\left(\alpha, \frac{1}{T} \frac{\hat{q}}{(\hat{\sigma}_a^2)^2}\right)$$

where \hat{q} and $\hat{\sigma}_a^2$ are consistent estimators. This approximation can be used to test the null hypothesis that $\alpha = 0$.

Forecasting using a sample

Often, forecasts are constructed from parametric models using estimated values of parameter. Let $\hat{f}_{t+h/t}$ denote the forecast constructed using these estimated parameters and let $\hat{e}_{t+h/t}$ denote the resulting forecast error. Some jargon:

If $\hat{f}_{t+h/t}$ is constructed using parameters estimated using data from time period 1 through t , then they are called “recursive” forecasts, and similarly for the forecast errors

If $\hat{f}_{t+h/t}$ is constructed using parameters estimated using data from time period $t-r$ through t , then they are called “rolling” forecasts, and similarly for the forecast errors.

Out-of-sample forecasting tests using estimated models

The loss-function and forecast combining tests developed above can be used using recursive and rolling forecast errors. In some cases, the asymptotic distributions need to be adjusted for sampling error in parameters of the estimated models. Let me work this out in the context of one specific example. (More detailed analyses are contained in West, K.D. (1996) "Asymptotic Inference About Predictive Ability," *Econometrica* 64 (1996), 1067-1084 and and McCracken, Michael (1999), "Asymptotics for Out-of-Sample Tests of Granger Causality"

http://www.missouri.edu/mwmd4f/MSE_all_new.pdf)

Suppose that forecast 1 is constructed using the regression model

$$y_{t+h} = x_t' \beta + e_{1,t+h}$$

so that $f_{1,t+h/t} = x_t' \beta$. Similarly, suppose that forecast 2 is constructed using the model

$$y_{t+h} = z_t' \gamma + e_{2,t+h}$$

so that $f_{2,t+h/t} = z_t' \gamma$.

A researcher has $T + R$ observations. He uses the first T observations to estimate β and γ by OLS. Then, using these estimators, he constructs forecasts over the remaining R observations.

Suppose that stochastic processes are sufficiently well behaved so that

$$\frac{1}{\sqrt{T}} \sum_{t=1}^T \begin{pmatrix} x_t e_{1,t+h} \\ z_t e_{2,t+h} \end{pmatrix} d \longrightarrow N(0, V)$$

and

$$(S_{xx}, S_{xz}, S_{zz}) p \longrightarrow (\Sigma_{xx}, \Sigma_{xz}, \Sigma_{zz})$$

where $S_{xx} = T^{-1} \sum_{t=1}^T x_t x_t'$ and so forth.

Let $\hat{e}_1 = y - x' \hat{\beta}$ and $\hat{e}_2 = y - z' \hat{\gamma}$, where superscripts have been suppressed. We now consider the properties of the loss function and encompassing tests constructed using these errors.

Loss Function Tests

The asymptotic distribution of these tests relied on

$$\frac{1}{\sqrt{T}} \sum \begin{pmatrix} e_{1,t}^2 - \sigma_1^2 \\ e_{2,t}^2 - \sigma_2^2 \end{pmatrix} \xrightarrow{d} N(0, V).$$

The tests using the estimated errors will have the same distribution if

$$\frac{1}{\sqrt{R}} \sum_{t=T+1}^{T+R} (e_{1,t}^2 - \sigma_1^2) - \frac{1}{\sqrt{R}} \sum_{t=T+1}^{T+R} (\hat{e}_{1,t}^2 - \sigma_1^2) p \longrightarrow 0$$

and similarly for e_2 . Now write

$$\hat{e}_1 = e_1 + x'(\beta - \hat{\beta})$$

so that

$$\begin{aligned} \frac{1}{\sqrt{R}} \sum_{t=T+1}^{T+R} (\hat{e}_{1,t}^2 - \sigma_1^2) &= \frac{1}{\sqrt{R}} \sum_{t=T+1}^{T+R} (e_{1,t}^2 - \sigma_1^2) \\ &\quad - 2(\hat{\beta} - \beta)' \frac{1}{\sqrt{R}} \sum x_t e_{1,t+h} \\ &\quad + (\hat{\beta} - \beta)' \frac{1}{\sqrt{R}} \sum x_t x_t' (\hat{\beta} - \beta) \end{aligned}$$

Now

$$(\hat{\beta} - \beta) \sim O_p(T^{-1/2}),$$

$$\frac{1}{\sqrt{R}} \sum x_t e_{1,t+h} \sim O_p(1)$$

$$\frac{1}{R} \sum x_t x_t' \sim O_p(1)$$

so that the second term vanishes by Slutsky's theorem when T is large. If $\frac{\sqrt{R}}{T} \longrightarrow 0$,

then the third term also vanishes. The results for e_2 are similar.

Thus, the test constructed using the estimated error terms behaves like the test

constructed using the true errors when the sample size is large.

Combining Tests

Combining tests were based on the regression

$$e_1 = \alpha(a) + e_c$$

where $a = e_1 - e_2$. Thus we need to consider the behavior of

$$\frac{1}{R} \sum (\hat{a}_t^2 - a_t^2)$$

and

$$\frac{1}{\sqrt{R}} \sum (\hat{e}_{1,t} \hat{a}_t - e_{1,t} a_t).$$

Write

$$\begin{aligned} \hat{a}_{t+h}^2 &= a_{t+h}^2 \\ &+ 2a_{t+h} z_t' (\hat{\gamma} - \gamma) + 2a_{t+h} x_t' (\hat{\beta} - \beta) \\ &+ 2(\hat{\gamma} - \gamma)' z_t x_t' (\hat{\beta} - \beta) \\ &+ (\hat{\gamma} - \gamma)' z_t z_t' (\hat{\gamma} - \gamma) \\ &+ (\hat{\beta} - \beta)' x_t x_t' (\hat{\beta} - \beta) \end{aligned}$$

Thus

$$\frac{1}{R} \sum (\hat{a}_t^2 - a_t^2) p \longrightarrow 0$$

if

$$\frac{1}{R} \sum a_{t+h} z_t' p \longrightarrow \Sigma_{az}$$

and

$$\frac{1}{R} \sum a_{t+h} x_t' p \longrightarrow \Sigma_{ax}$$

which will obtain under standard conditions (stationarity, ergodicity, existence of second moments.)

Now

$$\begin{aligned}\hat{e}_{1,t+h}\hat{a}_{t+h} &= e_{1,t+h}a_{t+h} \\ &+ e_{1,t+h}z'_t(\hat{\gamma}-\gamma) + e_{1,t+h}x'_t(\hat{\beta}-\beta) \\ &+ (\hat{\beta}-\beta)'x_t z'_t(\hat{\gamma}-\gamma) + (\hat{\beta}-\beta)'x_t x'_t(\hat{\beta}-\beta) \\ &+ (\hat{\beta}-\beta)'x_t a_{t+h}\end{aligned}$$

Thus, we can write

$$\begin{aligned}\frac{1}{\sqrt{R}}\sum(\hat{e}_{1,t}\hat{a}_t - e_{1,t}a_t) &= \left[\frac{1}{\sqrt{R}}\sum e_{1,t+h}z'_t\right](\hat{\gamma}-\gamma) + \left[\frac{1}{\sqrt{R}}\sum e_{1,t+h}x'_t\right](\hat{\beta}-\beta) \\ &+ \frac{\sqrt{R}}{T}[\sqrt{T}(\hat{\beta}-\beta)]'\left[\frac{1}{R}\sum x_t z'_t\right][\sqrt{T}(\hat{\gamma}-\gamma)] \\ &+ \frac{\sqrt{R}}{T}[\sqrt{T}(\hat{\beta}-\beta)]'\left[\frac{1}{R}\sum x_t x'_t\right][\sqrt{T}(\hat{\beta}-\beta)] \\ &+ \frac{\sqrt{R}}{T}[\sqrt{T}(\hat{\beta}-\beta)]'\left[\frac{1}{R}\sum x_t a_{t+h}\right]\end{aligned}$$

When forecast 1 is optimal, $E(Y | z, x) = x'\beta$, so that $E(e_1 | x, z) = 0$. Thus, $e_{1,t}z'_t$ and $e_{1,t}x'_t$ are mds, and it is reasonable to assume that

$$\frac{1}{\sqrt{R}}\sum\begin{pmatrix} e_{1,t+h}z'_t \\ e_{1,t+h}x'_t \end{pmatrix} \sim O_p(1)$$

(indeed that this random vector converges to a normal random variable.) Thus, the first term vanishes in probability. The second and third terms vanish when $\frac{\sqrt{R}}{T} \rightarrow 0$.

Now for the last term.

$$\frac{1}{R}\sum x_t a_{t+h} = \frac{1}{R}\sum x_t e_{1,t+h} - \frac{1}{R}\sum x_t e_{2,t+h}$$

The first term vanishes as x_t and $e_{1,t+h}$ are uncorrelated. The second does not. To see this, write $e_{2,t+h} = y_{t+h} - z'_t\hat{\gamma}$. Thus, using the optimal forecast for y , we have

$$e_{2,t+h} = e_{1,t+h} + [x_t - \pi z'_t]'\beta$$

where $\pi = \Sigma_{xz}\Sigma_{zz}^{-1}$. Or, writing $w_t = [x_t - \pi z'_t]'$, then

$$e_{2,t+h} = e_{1,t+h} + w_t' \beta$$

Noting that

$$x_t = \pi z_t + w_t$$

where the two terms on the rhs are uncorrelated, we have

$$E(x_t w_t') = E(w_t w_t') = \Sigma_{ww} = \Sigma_{xx} - \Sigma_{xz} \Sigma_{zz}^{-1} \Sigma_{zx}. \text{ Thus,}$$

$$\frac{1}{R} \sum x_t e_{2,t+h} p \longrightarrow \Sigma_{ww} \beta$$

Putting the pieces together we have

$$\frac{1}{\sqrt{R}} \sum \hat{e}_{1,t} \hat{a}_t = A_1 + A_2 + o_p(1)$$

where

$$A_1 = \frac{1}{\sqrt{R}} \sum e_{1,t} a_t$$

and

$$A_2 = -\sqrt{\frac{R}{T}} [\sqrt{T}(\hat{\beta} - \beta)] \left[\frac{1}{R} \sum x_t e_{2,t+h} \right].$$

So that the limit depends on $\frac{R}{T}$. When $\frac{R}{T} \longrightarrow \rho > 0$, we have

$$A_1 + A_2 d \longrightarrow N(0, V_1 + V_2)$$

where V_1 is the limiting distribution of A_1 and $V_2 = \rho \text{ trace}[\Sigma_{ww} V_{\hat{\beta}} \Sigma_{ww}]$, where $V_{\hat{\beta}}$ is

the limiting variance of $\sqrt{T}(\hat{\beta} - \beta)$. Note A_1 and A_2 are asymptotically independent.

Thus, sampling error in $\sqrt{T}(\hat{\beta} - \beta)$ means that the limiting distribution of the test statistic must be modified to include V_2 .

Nested and Non-nested Models

Suppose that x_t is a subset of the regressors in z_t . In this case, the asymptotic distributions of the tests discussed above change in an important way. To see this, recognize that in this case the null hypothesis in the loss function test means that $e_1 = e_2$. This is also implied by the null hypothesis in the forecasting combining test. Thus, the randomness in the test statistics depends entirely on the sampling error in the estimates of β and γ .

Thus, for example, in the loss function test

$$\begin{aligned} \sum (\hat{e}_{1,t+h}^2 - \hat{e}_{2,t+h}^2) &= -2(\hat{\beta} - \beta)' \sum x_t e_{t+h} \\ &\quad + (\hat{\beta} - \beta)' \sum x_t x_t' (\hat{\beta} - \beta) \\ &\quad - 2(\hat{\gamma} - \gamma)' \sum z_t e_{t+h} \\ &\quad + (\hat{\gamma} - \gamma)' \sum z_t z_t' (\hat{\gamma} - \gamma) \end{aligned}$$

where e_t is the common value of $e_{1,t}$ and $e_{2,t}$. Notice the sum is $O_p(1)$ when

$\frac{R}{T} \rightarrow p > 0$. The paper by McCracken covers this in detail.