Forecasting

Jesus Gonzalo

November 9, 2020

Some Useful References:

- Granger, C.W.J. and Newbold, P., Forecasting Economic Time Series, second edition. New York: Academic Press, 1986.
- Elliott, G., A. Timmermann, Economic Forecasting. Princeton University, 2016.
- "Comparing Predictive Accuracy" , Francis X. Diebold and Roberto S. Mariano, JBES 1995.
- "Asymptotic Inference about Predictive Ability" , Kenneth West, Econometrica 1996.

Outline

Outline

1. Basic concepts
   - Information set
   - Conditional Variables
   - Cost function
   - Linear forecast
2. Generalized cost functions
3. The evaluation of forecasts
4. Combination of forecasts (Bates and Granger)

- Let $\{X_t\}$ be some discrete time, stationary and stochastic process.
- We are at $t = n$ and wish to forecast h periods ahead, so we want to forecast $X_t + h$.
- $\mathcal{I}_n$: information available at time n.
- Univariate information set:
  - A sample of previous values of the series: X_{n-j} for $j = 0, 1, \ldots, n$.
  - Some properties of the stochastic process $\{X_t\}$. e.g. $E(x_t)$, stationarity, etc.
- Multivariate information set:
  - A sample of previous values of the series: X_{n-j}, Y_{n-j}, Z_{n-j}, $\ldots$ for $j = 0, 1, \ldots, n$.
  - Some properties of the stochastic process X_{n-j}, Y_{n-j}, Z_{n-j}, $\ldots$

- X_{n+h} is a random variable.
- It can be fully characterized by a probability density.
- Since $\mathcal{I}_n$ has to be used, we need a conditional density function, i.e.

$$\mathrm{Prob}(x < X_{n+h} \leq x + \mathrm{d}x | \mathcal{I}_n) = g_{c,h}(x)\mathrm{d}x \tag{1}$$

- If $g_c(x)$ is known, other properties of X_{n+h} can be immediately determined.
- If not, we attempt to find:
  - A point forecast: $\hat{X}_{n+h}$.
  - A confidence interval:

$$\mathrm{Prob}(x_1 < X_{n+h} \leq x_2) = \alpha \tag{2}$$

($X_{n+h} \in [x_1, x_2]$ with probability α).

- Criteria to find the best point forecast? Consider the forecast $f_{n,h}$
- Cost (Risk) function:

$$C(e) \quad with \quad C(0) = 0 \quad \text{and} \quad e_{n,h} = X_{n+h} - f_{n,h} \tag{3}$$

- The best forecast is given by:

$$f^*_{n,h} = \underset{\{f_{n,h}\}}{\arg\min}\, E\left[C(e_{n,h})|\mathcal{I}_n\right] \tag{4}$$

Example

$C(e) = ae^2$. In this case the best forecast is $f_{n,h} = E(X_{n+h}|\mathcal{I}_n)$ (prove it!!!)

- We will rarely get to know the conditional density function sufficiently well to find a complete solution of (4).
- We will impose restrictions.
- In particular we focus on linear models: $f_{n,h}$ is a linear function of the data available in $\mathcal{I}_n$.

Outline

Generalized cost functions

- A generalized cost function $C(e)$ satisfy:
  1. $C(0) = 0$
  2. Monotonic non decreasing for $e > 0$:

$$e_1 > e_2 > 0 \implies C(e_1) \geq C(e_2) \tag{5}$$

  3. Monotonic non increasing for $e < 0$
- Notice that doesn't need to be symmetrical.
- The optimal forecast $f_{n,h}$ is the one that minimizes:

$$J = \int_{-\infty}^{+\infty} C(x - f_{n,h}) g_{c,h}(x) \mathrm{d}x \tag{6}$$

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

We have to minimize:

$$J = \int_{-\infty}^{+\infty} a(x - f_{n,h})^2 g_{c,h}(x)\mathrm{d}x \tag{7}$$

Define $M_h = \mathbb{E}\left[X_{n+h}|\mathcal{I}_n\right] = \int_{-\infty}^{+\infty} x g_{c,h}(x)\mathrm{d}x$. After some manipulation, (7) can be written as:

$$J = a(M_h - f_{n,h})^2 + a\int_{-\infty}^{+\infty} (x - M_h)^2 g_{c,h}(x)\mathrm{d}x \tag{8}$$

Then:

$$f_{n,h} = M_h = \mathbb{E}\left[X_{n+h}|\mathcal{I}_n\right] \tag{9}$$

is the optimal forecast.

- Is $f_{n,h} = M_h$ the optimal predictor for a wider class of cost functions?

Theorem (Conditional expectation as best forecast)

$f_{n,h} = M_h$ is the optimal predictor if:

1. *$C(e)$ is symmetric around e=0.*
2. *$C'(e)$ exists almost everywhere and is strictly increasing for $-\infty < e < \infty$.*
3. *$g_{c,h}(x)$ is symmetric around $x = M_h$.*

or if:

1. *$C(e)$ is symmetric around e=0.*
2. *$g_{c,h}(x)$ is symmetric around $x = M_h$, is continuous and unimodal.*

- $C(e)$ is not always symmetric... Think on some examples...

- What if the cost function is not symmetric?

Theorem (Conditional expectation as best forecast)

If the conditional distribution $g_{c,h}(x)$ is assumed to be normal, then the optimal predictor is given by:

$$f_{n,h} = M_h + \alpha \tag{10}$$

where α only depends on the cost function $C(e)$ (not on $\mathcal{I}_n$). (Christoffersen and Diebold 1997).

Example

Let:

$$C(e) = \begin{cases} ae & \text{if} e > 0 \\ 0 & \text{if} e = 0 \\ be & \text{if} e < 0 \end{cases} \quad ; \quad a > 0, b < 0 \tag{11}$$

The expected cost is:

$$J = \mathbb{E}\left[C(X_{n+h} - f)|\mathcal{I}_n\right] = a\int_f^{\infty}(x-f)g_{c,h}(x)\mathrm{d}x + b\int_{-\infty}^{f}(x-f)g_{c,h}(x)\mathrm{d}x \tag{12}$$

First order condition:

$$G_{c,h}(f) = \frac{a}{a-b} \tag{13}$$

where $G_{c,h}$ is the conditional cumulative distribution. For the symmetric case ($a = -b$):

$$G_{c,h}(f) = 1/2 \tag{14}$$

so that f optimal is the median of $g_{c,h}$.

Outline

1. Basic concepts
   - Information set
   - Conditional Variables
   - Cost function
   - Linear forecast
2. Generalized cost functions
3. The evaluation of forecasts
4. Combination of forecasts (Bates and Granger)

The evaluation of forecast

- An objective evaluation of a set of forecast might seek to answer:
  1. Is one set of forecast better than its competitors?
  2. How 'good', in some sense, is a particular set of forecasts?
  3. Can the forecast-generation mechanism be modified in some way so as to yield improved forecast performance?

Example

Let:

$$X_t, f_t \qquad t = 0, 1, 2, \ldots, N$$
$$e_t = X_t - f_t \qquad t = 0, 1, 2, \ldots, N$$

The expected squared forecast error estimated is:

$$D_N^2 = \frac{1}{N} \sum_{t=1}^{N} e_t^2 \tag{15}$$

The evaluation of forecast

- Suppose that are two forecasting competing procedures that produce errors: $\{e_t^{(1)}, e_t^{(2)}\}_{t=1}^{N}$
- Additionally, suppose $\{e_t^{(1)}, e_t^{(2)}\}_{t=1}^{N}$ is a random sample of a bivariate distribution with zero mean, variance σ^2, correlation coefficient ρ and $\text{Cov}(e_t^{(i)}, e_{t-j}^{(i)}) = 0$.
- Test:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

- Construct a pair of random variables $e^{(1)} + e^{(2)}$ and $e^{(1)} - e^{(2)}$ and compute the covariance:

$$\mathbb{E}\left[(e^{(1)} + e^{(2)})(e^{(1)} - e^{(2)})\right] = \sigma_1^2 - \sigma_2^2 \tag{16}$$

- Then regress $e^{(1)} + e^{(2)}$ on $e^{(1)} - e^{(2)}$ by OLS, and find the regression coefficient.
- If the coefficient is zero, then they are uncorrelated, i.e. $\sigma_1^2 = \sigma_2^2$.

The evaluation of forecast

'Comparing predictive accuracy' by Diebold and Mariano, 1995

- This paper allows for forecast errors that are potentially non-Gaussian, nonzero mean, serially correlated and contemporaneously correlated, that is, under general assumptions.
- Let:
  - $\{y_t\}_{t=1}^T$ 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 $\{\hat{e}_t^{(2)}\}_{t=1}^T$ the associated forecast errors.
  - $g(y_t, \hat{y}_t^{(i)}) = g(\hat{e}_t^{(i)})$ cost function or loss function.
  - $d_t = g(\hat{e}_t^{(1)}) - g(\hat{e}_t^{(2)})$
- Hypothesis:

$$H_0 : \mathbb{E}[d_t] = 0$$
$$H_1 : \mathbb{E}[d_t] \neq 0$$

The evaluation of forecast

'Comparing predictive accuracy' by Diebold and Mariano, 1995

- **Test statistic.** If $\{d_t\}_{t=1}^{T}$ is assumed to be covariance stationary and short memory, then:

$$\sqrt{T}(\overline{d} - \mu) \xrightarrow{d} N(0, 2\pi f_d(0)) \tag{17}$$

- where $2\pi f_d(0)$ is the Long-run Variance.

$$\overline{d} = \frac{1}{T}\sum_{t=1}^{T}\left[g(\hat{e}_t^{(1)}) - g(\hat{e}_t^{(2)})\right] \quad ; \quad f_d(0) = \frac{1}{2\pi}\sum_{\tau=-\infty}^{\infty}\gamma_d(\tau) \tag{18}$$

Then:

$$S = \frac{\overline{d}}{\sqrt{\frac{2\pi \hat{f}_d(0)}{T}}} \sim N(0,1) \tag{19}$$

So we reject H_0 at 5 % if $|S| > 1.96$.

- No parameter uncertainty. Note that $\hat{e}$ (forecasting error) is calculated with the true parameters and no with their estimates.
- Think on how to carry on this test via a simple linear regression:

$$d_t = \alpha + u_t \tag{20}$$

Then.......?

The evaluation of forecast

'Comparing predictive accuracy' by Diebold and Mariano, 1995

- **Long run variance ($\lim_{n\to\infty} \mathbb{V}(\sqrt{T}\bar{y})$ if it exists).** Let $y_t = \mu + \Psi(L)\epsilon_t$, $\epsilon_t \sim iid(0, \sigma^2)$. LRV is the variance of the asymptotic distribution of

$$\sqrt{T}\,(\bar{y} - \mu) \overset{d}{\sim} AD(0, LRV) \tag{21}$$

- **Non-parametric estimator**

$$\begin{aligned} LRV = \sigma^2\Psi^2(1) &= \sum_{-\infty}^{\infty} \gamma_k \\ &= \gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k \end{aligned}$$

Naïve estimator

$$L\hat{R}V = \hat{\gamma}_0 + 2\sum_{k=1}^{M_T} \hat{\gamma}_k \tag{22}$$

M_T = truncation point. This estimator is inconsistent.

The evaluation of forecast

'Comparing predictive accuracy' by Diebold and Mariano, 1995

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

$$L\hat{R}V_{NW} = \hat{\gamma}_0 + 2\sum_{j=1}^{M_T} \underbrace{\left[1 - \frac{j}{M_{T+1}}\right]}_{\text{Bartlett's Window}} \hat{\gamma}_j \tag{23}$$

with $M_T \to \infty$ and $\frac{M_T}{T} \to 0$ as $T \to \infty$.

Example

$M_T = 4\left(\frac{T}{100}\right)^{\frac{1}{4}}$. Then if $T = 100$, $M_T = 4$ and

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

$$\hat{\gamma}_0 + 2\left[\frac{4}{5}\hat{\gamma}_1 + \frac{3}{5}\hat{\gamma}_2 + \dots\right]$$

- **Alternative estimator** (based on the AR representation). Assume $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$, $\epsilon_t \sim iid(0, \sigma^2)$.
  1. Estimate $\hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p, \hat{\sigma}^2$
  2. $\hat{\Psi}(1) = \frac{1}{1-\hat{\phi}_1-\hat{\phi}_2-\ldots\hat{\phi}_p}$
  3. $L\hat{R}V_{AR} = \hat{\sigma}^2 \left(\hat{\Psi}(1)\right)^2$

More on LRV estimation

'Dynamic Nonlinear Econometric Analysis' by Pötscher and Prucha, 1996

$$\sqrt{T}\,(\bar{y}-\mu) \overset{d}{\sim} AD(0, LRV) \tag{24}$$

$$LRV = \sum_{\infty}^{\infty} \gamma_k = \gamma_0 + 2\sum_{k=1}^{\infty} \gamma_k \tag{25}$$

Define $v_t = y_t - \mu$. Then,

$$LRV = \mathbb{E}\left(v_t^2\right) + 2\sum_{j=1}^{\infty} \mathbb{E}\left(v_t v_{t+j}\right) \tag{26}$$

How do we estimate it? We will check some candidates.

First candidate:

$$\frac{1}{n}\sum_{1}^{n} v_t^2 + 2\sum_{j=1}^{n-j} v_t v_{t+j} \tag{27}$$

- This "naïve" 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 then roughly "n" times as large $(O(1))$.

More on LRV estimation

'Dynamic Nonlinear Econometric Analysis' by Pötscher and Prucha, 1996

Second candidate:

$$\frac{1}{n}\sum_{1}^{n} v_t^2 + 2\sum_{j=1}^{n-1}\frac{1}{n}\sum_{t=1}^{n-j} v_t v_{t+j} \tag{28}$$

- 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 naïve estimator. It seems natural to exclude or down-weigh the sample moments corresponding to lags j close to n.
- This is achieved by introducing weights into the formula for the naïve estimator.
- Down-weighing of the sample covariance has the effect of reducing the variance of the estimator at the expense of introducing a bias.
- This candidate is still inconsistent although some moderate down-weighing of the sample moments $\frac{1}{n-j}\sum_{t=1}^{n-j} v_t v_{t+j}$ takes place.

Third candidate:

$$\frac{1}{n}\sum_{1}^{n} v_t^2 + 2\sum_{j=1}^{M}\frac{1}{n}\sum_{t=1}^{n-j} v_t v_{t+j} \xrightarrow{p} \mathbb{E}\left(v_t^2\right) + 2\sum_{t=1}^{M}\mathbb{E}\left(v_t v_{t+j}\right) \tag{29}$$

- The probability limit is equal to LRV only if $\mathbb{E}\left(v_t v_{t+j}\right) = 0$ for $j > M$ (for example if v_t is m-dependent with $m \leq M$).
- Because we are always assuming that $\mathbb{E}\left(v_t v_{t+j}\right) \to 0$ for $j \to \infty$, then it is clear that the bias will be smaller the larger M is.
- So a possible solution is to make M sample size dependent (M_n) such that $M_n \to \infty$ to avoid bias, but slowly enough so that the variance still goes to zero as $n \to \infty$.

More on LRV estimation

'Dynamic Nonlinear Econometric Analysis' by Pötscher and Prucha, 1996

Fourth candidate:

$$\frac{1}{n}\sum_{1}^{n} v_t^2 + 2\sum_{j=1}^{M_n}\frac{1}{n}\sum_{t=1}^{n-j} v_t v_{t+j} \tag{30}$$

or more general, fifth candidate:

$$LRV_n = \frac{1}{n}\sum_{1}^{n} v_t^2 + 2\sum_{j=1}^{n-1} w\left(j,n\right)\sum_{t=1}^{n-j} v_t v_{t+j} \tag{31}$$

for appropriate weights $w\left(j,n\right) \in \mathbb{R}$

- Clearly, (30) is a special case because it can be obtained with

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

- Another example is

$$w(j,n) = \begin{cases} 1 - \frac{j}{M_n} & 0 \le j \le M_n \\ 0 & M_n \le j \le n-1 \end{cases} \quad \text{(Bartlett Kernel)}$$

Outline

Combination of forecasts (an ecological principle)

Bates & Granger, 1969

- Two unbiased one-step-ahead forecasts $f_n^{(1)}$ and $f_n^{(2)}$
- Forecast errors

$$
\begin{aligned}
e_n^{(j)} &= x_n - f_n^{(j)} \quad j = 1, 2 \\
\mathbb{E}\left[e_n^{(j)}\right] &= 0 \\
\mathbb{E}\left[\left(e_n^{(j)}\right)^2\right] &= \sigma_j^2 \\
\mathbb{E}\left[e_n^{(1)} e_n^{(2)}\right] &= \rho\sigma_1\sigma_2
\end{aligned}
$$

Combination of forecasts

Bates & Granger, 1969

- Consider now a combined forecast

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

- Hence the error variance

$$\sigma_C^2 = k^2 \sigma_1^2 + (1-k)^2\, \sigma_2^2 + 2k\,(1-k)\,\rho\sigma_1\sigma_2 \tag{32}$$

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

$$k_0 = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{33}$$

- Substituting in the error variance expression

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

- Notice that $\sigma_C^2 < min\left(\sigma_1^2, \sigma_2^2\right)$ unless $\rho = \frac{\sigma_1}{\sigma_2}$ or $\rho = \frac{\sigma_2}{\sigma_1}$.
- If either equality holds, then the variance of the combined forecast error is equal to the smaller of the two error variances. In most practical situations, the best available combined forecast will outperform the better individual forecast.

- From expression (33) one can obtain two "extreme" interesting results
- The first one is

$$k_0 \geq (\leq) 0 \iff \frac{\sigma_2}{\sigma_1} \geq (\leq) \rho$$

- 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 one is when in (32) $\rho \to -1$ or $\rho \to 1$.
- When $\rho \to -1$, then $\sigma_{C,0} \to 0$, implying a perfect forecast.
- When $\rho \to 1$, then $\sigma_{C,0} \to 0$ except when $\sigma_1 = \sigma_2$ in which case the limit is σ_1^2. Try to interpret this result.

1. Expression (33) is not very useful because the parameters are unknown.
2. Two ways of obtaining the weights in C_n.
   1. Either plug in sample moments into population moments
   2. or use

      $$X_{n+1} = kf_n^{(1)} + (1-k)\, f_n^{(2)} + e_n,$$

      to regress X_{n+1} on $f_n^{(1)}$ and $f_n^{(2)}$ and obtain the coefficients by OLS, imposing that they should add up to one.
3. Time varying weights.

Some final (??) comments

1. Do not forget to read Wei,W., Time Series Analysis: Univariate and Multivariate Methods(1990) for standard Box-Jenkins univariate $ARMA(p, q)$ prediction.
2. Think on how to carry on MULTI-step forecast: direct versus plug-in.
3. Think on how to forecast with Non-Linear Models.
4. Think on how to forecast "events" that never occurred before.