EC4304: Economic and Financial Forecasting

J.Y. Liew

October 2023

1 Introduction to Forecasting

In general, we wish to forecast the distribution of a random variable $F(y) = P(Y \leq y)$. However, a point forecast $\hat{Y}$ is more desirable because it is our best guess for the unknown value. $\hat{Y}$ is a function of the predictive distribution $F(y)$. The function depends on the **loss function**.

1.1 Forecast Error, Loss Functions, Risk

A forecast error is given by

$$e = Y - \hat{Y}$$

The cost of a forecast error is determined by the loss function $L(e)$.

Properties of loss functions:

- $L(0) = 0$
- $L(e) \geq 0$ for all e
- $e_2 < e_1 < 0 \Rightarrow L(e_1) \leq L(e_2)$; $e_2 > e_1 > 0 \Rightarrow L(e_1) \leq L(e_2)$

Types of forecasting:

- squared loss: $L(e) = e^2$
- absolute loss: $L(e) = |e|$
- level-dependent loss: $L(e, Y) = |\frac{e}{Y}|$
- linex loss: $L(e) = b(e^{ae} - ae - 1), a \neq 0, b > 0$ etc.

The **risk** of a forecast is its **expected loss**.

$$R(\hat{Y}) = E[L(e)] = E[L(Y - \hat{Y})]$$

Under square loss, $R(\hat{Y}) = E[(Y - \hat{Y})^2] = MSE$.
Under absolute loss, $R(\hat{Y}) = E(|Y - \hat{Y}|) = MAE$.

Note: different loss functions gives different optimal forecast, we can't compare the risks across different loss functions

1.2 Forecast Interval

A forecast interval is defined by

$$C = [\hat{Y}_L, \hat{Y}_U]$$

A $(100 \times X)\%$ forecast interval satisfies:

$$P(Y \in C) = X$$

Popular choices of X are $\{0.9, 0.8, 0.68, 0.5\}$.$\hat{Y}_L = \frac{1-X}{2}$ quantiles of $F(y)$, $\hat{Y}_U = \frac{1-(1-X)}{2}$ quantiles of $F(y)$. The α^{th} quantile of Y is the number $q(\alpha)$ that satisfies $\alpha = F(q(\alpha))$.

1.2.1 Monotonicity Rule

Suppose Y is a random variable with the α^{th} quantile of $q_Y(\alpha)$. Let $m = g(Y)$ be an increasing transformation of Y, e.g.

$$m = a + bY$$

$$m = ln(Y)$$

$$m = e^Y$$

Then the α^{th} quantile of m is $g(q_Y(\alpha))$:

$$q_m(\alpha) = a + bq_Y(\alpha)$$

$$q_m(\alpha) = ln(q_Y(\alpha))$$

$$q_m(\alpha) = e^{q_Y(\alpha)}$$

We can use this rule to form forecast intervals.

1.2.2 Normal Rule

Suppose $Y \sim N(\mu, \sigma^2)$, then $\hat{Y} = \mu$. The linear function $c = \frac{Y-\mu}{\sigma} \sim N(0,1)$. The $100 \times (1-\alpha)\%$ forecast interval is $[\mu - \sigma z_{\frac{\alpha}{2}}, \mu + \sigma z_{\frac{\alpha}{2}}]$ using the monotonicity rule.
The procedure is:

1. find Z's α^{th} quantile, using Z-table
2. transform back to Y from Z $(\mu + \sigma z)$

Basically, use **Increasing Transformation** to transform Y to Z. Since $Z \sim N(0,1)$, we can apply **normal rule** $([-Z_{\frac{\alpha}{2}}, Z_{\frac{\alpha}{2}}])$, then **transform back** to Y, the value we obtained, by monotonicity rule, is Y's α^{th} quantile.

$$E(Y_{t+h}|\Omega_t) = T_t + S_t + C_t$$

2 Modelling and Forecasting Trend

If trend is not modelled, we are likely to suffer from unit root problem. Another problem is the MSFE diverges as T increases.
Common trend specifications:

- linear: $T_t = \beta_0 + \beta_1 t$
- quadratic: $T_t = \beta_0 + \beta_1 t + \beta_2 t^2$
- exponential trend: $T_t = \beta_0 e^{\beta_1 t}$
- log-linear trend: $ln(T_t) = ln(\beta_0) + \beta_1 t$

2.1 Model Selection

BIC and AIC provide good measures for us to choose the model that performs best. However, we need to note that we cannot compare models with different dependent variables and different sample sizes. Pick the model with lowest AIC/BIC value.

2.2 Point Forecast Construction

$$Y_t = \beta_0 + \beta_1 t + \epsilon_t$$
$$Y_{T+h} = \beta_0 + \beta_1 (T+h) + \epsilon_{T+h}$$
$$E_T(Y_{T+h}) = \beta_0 + \beta_1 (T+h)$$
$$\hat{Y}_{T+h} = \beta_0 + \beta_1 (T+h)$$

2.3 RMSFE

$$Y_{T+1} - \hat{Y}_{T+1|T} = \underbrace{\epsilon_{T+1}}_{\text{irreducible error}} - \underbrace{[(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)(T+1)]}_{\text{estimation uncertainty}}$$

$$MSFE = \sigma_\epsilon^2 + var(\hat{\beta}_0) + (T+1)^2 var(\hat{\beta}_1) + 2(T+1) cov(\hat{\beta}_0, \hat{\beta}_1)$$

Estimation uncertainty can be negligible if we have true parameters, large sample size, and less estimators.
How to check if we capture the trend?
Plot the residuals, if the residuals show a little trend, then we basically doing well.

2.4 Breaking Trend

If the slope of the trend changing in the data at some point, we can model it using interaction between trend variable and a dummy.

if $t < \tau$, $Y_t = \beta_0 + \beta_1 t + u_t$
if $t \geq \tau$, $Y_t = \alpha_0 + \alpha_1 t + u_t$

$$\begin{aligned} Y_t &= (\beta_0 + \beta_1 t)I(t < \tau) + (\alpha_0 + \alpha_1 t)I(t \geq \tau) + u_t \\ &= (\beta_0 + \beta_1 t) + ((\alpha_0 + \beta_0) + (\alpha_1 - \beta_1)t)I(t \geq \tau) + u_t \\ &= \beta_0 + \beta_1 t + \beta_2 d_t + \beta_3 t d_t + u_t \end{aligned}$$

Note that we use pre-trend parameters to subtract post-trend parameters because we want to remove the part that ovelapped.

2.4.1 Continuous Break

We can use **spline** technique to smooth the breaking trend.

$$\begin{aligned} Y_t &= \gamma_0 + \gamma_1 t + \gamma_2 (t - \tau)I(t \geq \tau) + u_t \\ &= \gamma_0 + \gamma_1 t + \gamma_2 t^* + u_t \text{ , where } t^* = (t - \tau)I(t \geq \tau) \end{aligned}$$

t^* is 0 before the break τ and smoothly increasing trend afterward.

3 Modelling and Forecasting Seasonality

The seasonal component is a repetitive pattern over a **calendar year**.

$$S_t = \sum_{i=1}^{S} \gamma_i D_{it} = \begin{cases} \gamma_1 & \text{if } t = \text{ period 1} \\ \gamma_2 & \text{if } t = \text{ period 2} \\ . & \\ . & \\ \gamma_s & \text{if } t = \text{ period s} \end{cases}$$

$$Y_t = \alpha + \sum_{i=1}^{S-1} \beta_i D_{it} + \epsilon_t$$

How to check the seasonal pattern has been captured by our model?

Plot the residuals out. If it shows no season, then we have removed the seasonal pattern from our data.

3.1 Deseasonalization

We can perform a simple seasonal adjustment by subtracting the seasonal component from the original series. This means that we can **add the mean of the series to the residuals** because residuals are the remaining part that isn't captured in the seasonal model. Note that residuals aren't the deseasonalized series because it has zero mean but the time series mean is not zero.

The intuition behind it is in the seasonal dummy model, we let the mean vary with seasons. Once this variation is taken care of, the series becomes a constant mean plus noise.

4 Characterizing and Modeling Cycles

Cycle is whatever persistent dynamics that remain after accounting for trend and seasonality.

4.1 Stationarity

Y_t is mean stationary if $E(Y_t) = \mu \quad \forall\, t$
Y_t is mean stationary if $Var(Y_t) = \sigma^2 \quad \forall\, t$

4.1.1 Autocovariance

The autocovariance is the covariance structure of Y_t with itself at different time displacemments. The k^{th} order autocovariance of Y_t is

$$\gamma(t,k) = Cov(Y_t, Y_{t-k}) = E[(Y_t - \mu)(Y_{t-k} - \mu)]$$

Notice the role of mean stationarity here.

4.1.2 Autocorrelation

The k^{th} order autocorrelation of Y_t is

$$\rho(t,k) = \frac{Cov(Y_t, Y_{t-k})}{\sqrt{Var(Y_t)Var(Y_{t-k})}} = \frac{Cov(Y_t, Y_{t-k})}{Var(Y_t)}$$

Notice the role fo variance stationarity here.

$\rho(1) < 0 \implies Y_t$ changes direction in adjacent periods

4.1.3 Covariance Stationarity

A time series process Y_t is covariance stationary if its

1. mean constant over time : $E(Y_t) = \mu \quad \forall\, t$
2. variance constant over time : $Var(Y_t) = \sigma^2 \quad \forall\, t$
3. covariance constant over displacements : $\gamma(t,k) = \gamma(k) \quad \forall\, t, k$
4. 2^{nd} moment is finite : $E(Y_t^2) < \infty$

Some important results based on covariance stationarity:

1. $\gamma(k) = \gamma(-k)$
2. $\gamma(0) = Var(Y_t) = \sigma^2$
3. $|\gamma(k)| \leq \gamma(0) \quad \forall\, k$
4. $\rho(k) = \frac{\gamma(k)}{\gamma(0)}$
5. $\rho(0) = 1$
6. $\rho(k) = \rho(-k)$

Strong stationary is different from covariance stationary because it assumes all moments of Y_t to be constant over time. However, strong stationary $\not\Rightarrow$ covariance stationary because it doesn't say anything about finite variance.

4.2 White Noise

$$Y_t \sim WN(\mu, \sigma^2)$$

A white noise process has zero autocorrelations $\rho(k) = 0$ for $k > 0$. It is serially uncorrelated, so linearly **unforecastable**. However, it is not necessarily iid. Although white noise is mean and variance stationary, it doesn't imply conditional variance is constant. Serially uncorrelated $\neq$ serially independent.

4.3 Ergodicity

We think of the time series expectation as an ensemble average

$$E(Y_t) = plim_{I\to\infty} \frac{1}{I} \sum_{i=1}^{I} Y_t^{(i)}$$

In human words, the above expression means we run I times of simulation for Y at time t. So for time t, we will have I simulated Ys. Then take the average of these $Y_t^{(i)}$, it will converge to a value. This value is called the ensemble average. This concept is similar to the β in regression: we are using sample $\hat{\beta}$ to estimate the true unknown parameter β.Hence, ergodicity ensures LLN works.

A process is **ergodic** if the time average converges to ensemble average as T grows large. This happens when $|\rho(k)| \to 0$ as $k \to \infty$. This means, if Y_t is ergodic, the **long-horizon forecast converges to the unconditional mean**: $\hat{Y}_{T+h|T} \approx E(Y_t)$.

4.4 Estimations

Population mean:

$$\mu = E(Y_t)$$

Sample mean:

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^{T} Y_t$$

Population autocovariance:

$$E[(Y_t - \mu)(Y_{t-k} - \mu)]$$

Sample autocovariance:

$$\hat{\gamma}(k) = \frac{1}{T} \sum_{t=k+1}^{T} (Y_t - \hat{\mu})(Y_{t-k} - \hat{\mu})$$

Population autocorrelation:

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)}$$

Sample autocorrelation:

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}$$

If Y_t is independent white noise, then the **confidence bands for autocorrelations** is,

$$Var(\hat{\rho}) \approx \frac{1}{T}$$

Bartlett's formula: Assuming Y_t is $MA(q)$ (i.e. $\rho(q) = 0 \; for \; k > q$,

$$Var(\hat{\rho}(k)) \approx \frac{1}{T}(1 + 2\sum_{i=1}^{q} \rho(i)^2), \; k > q$$

If sample autocorrelations are all within the Bartlett's band, then Y_t is likely white noise.

4.5 Joint Test for Autocorrelations

$$H_0 : \rho(1) = \rho(2) = ... = \rho(m) = 0$$
$$H_1 : \exists k \in \{t_i\}_{i=1}^{T} \ s.t. \ \rho(k) \neq 0$$

We can use 2 statistics to do the test

- Box-Pierce Q-statistics:

$$Q_{BP} = T \sum_{i=1}^{m} \hat{\rho}^2(i) \sim \chi^2_m$$

- Ljung-Box Q-statistics:

$$Q_{LB} = T(T+2) \sum_{i=1}^{m} \frac{1}{T-i} \hat{\rho}^2(i) \sim \chi^2_m$$

If the null hypothesis is rejected, then we accept that Y_t is not white noise. Now the question is **how to choose** m? We can use $m = \sqrt{T}$.

4.6 Wold's Theorem

For simplicity's sake, let Y_t be a mean-zero covariance stationary process without trend or seasonality, then Y_t can be expressed as:

$$Y_t = B(L)\epsilon_t = \sum_{i=0}^{\infty} b_i \epsilon_{t-i} \ , \ \epsilon_t \sim WN(0,^2)$$

$$\sum_{i=0}^{\infty} b_i^2 < \infty$$

$$\epsilon_t = Y_t - \hat{E}(Y_t | Y_{t-s} \ , \ s \geq 1)$$

Any stationary process can be approximated by the **general linear process** above. Wold's theorem also holds for non-mean zero, trend and seasonal time series. What we need to do is detrend and deseasonalize it.

A stationary time series process is constructed as a linear function of **innovations**, ϵ_t, which are white noise.

$$E(Y_t) = 0$$
$$Var(Y_t) = E(Y_t^2) = \sigma^2 \sum_{i=0}^{\infty} b_i^2 < \infty$$
$$E(Y_t | \Omega_{t-1}) = \sum_{i=0}^{\infty} b_i \epsilon_{t-i}$$
$$Var(Y_t | \Omega_{t-1}) = \sigma^2$$

What's so useful for these moments? We will see later.

The entire backbone of modelling cycles is

$$B(L) \approx \frac{\Theta(L)}{\Phi(L)} = \frac{\sum_{i=0}^{q} \theta_i L^i}{\sum_{j=0}^{p} \phi_j L^j}$$

which means, if Y_t is stationary, we can expressed it as a general linear process, which can be approximated by this rational series. The goal of our forecasting is, thus, to find a good approximation for $B(L)$ (AR, MA, ARMA).

4.7 MA Process

MA models are linear functions of stochastic errors. The Wold representation is an example of MA(∞). Simplest MA(1) model is

$$Y_t = \epsilon_t + \theta\epsilon_{t-1} = (1+\theta L)\epsilon_t \, , \, \epsilon_t \sim WN(0, \sigma^2)$$

θ controls the degree of serial correlation. This shows that the innovations have impact in future Y.

4.7.1 Unconditional Moments

$$E(Y_t) = 0$$
$$Var(Y_t) = \sigma^2(1+\theta^2)$$

The variance depends on θ: if Y_t is more serially correlated, $Var(Y_t)$ will be larger.

4.7.2 Conditional Moments

$$E(Y_t|\Omega_{t-1}) = \theta\epsilon_{t-1}$$
$$Var(Y_t|\Omega_{t-1}) = \sigma^2$$

Intuition of conditional mean: optimal conditional mean is unforecastable (because all are random errors), but if we have the information of the past random error, we know it affects today's Y at some level, so we can use that to be our forecast.

4.7.3 Autocovariance and Autocorrelation

$$\gamma(1) = \theta\sigma^2$$
$$\gamma(k) = 0 \ , \ for \ k > 1$$
$$\rho(1) = \frac{\theta}{1+\theta^2}$$
$$\rho(k) = 0$$

Intuition: The error in MA(q) process will only affect Y up to q periods, after that, it will have no effect on it. This makes Y forecastable.
Because of that, θ must be in (-1, 1) and $\rho(1)$ must be in (-0.5, 0.5) to guarantee **invertibility**.

4.7.4 Stationarity and Invertibility of MA

If we can express the MA process in an **autoregressive representation**, then we say that MA process is **invertible**.

By doing rearranging, lag, and combine repeatedly on MA(1), we wil obtain

$$Y_t = -\sum_{i=1}^{\infty}(-\theta)^i Y_{t-i} + \epsilon_t$$

This series converges if $|\theta| < 1$. We can also write it as

$$(1 + \theta L)^{-1} Y_t = \epsilon_t$$

This implies that we can express Y_t as a function of only present and past values of Y.

4.7.5 MA(q) Process

In MA(q), Y_t is invertible if

$$Y_t = (1 + \theta_1 L + \theta_2 L^2 + ... + \theta_q L^q)\epsilon_t = \Theta(L)\epsilon_t$$

and all the q roots of the polynomial are outside of the unit circle.

Remeber our goal to find a good approximation for $B(L)$? MA(q) is possibly one simple approximation as

$$B(L) \approx \frac{\Theta(L)}{\Phi(L)} = \frac{\sum_{i=0}^{q} \theta_i L^i}{1} = \Theta(L)$$

4.8 AR Process

An AR(1) process is given by

$$Y_t = \phi Y_{t-1} + \epsilon_t \ , \ \epsilon_t \sim WN(0, \sigma^2)$$

which can also be written as

$$(1 - \phi L) Y_t = \epsilon_t$$

4.8.1 Inversion of AR(1)

We can also rewrite it as an MA(∞):

$$Y_t = \sum_{i=0}^{\infty} \phi^i \epsilon_{t-i}$$

We then need $|\phi| < 1$ in order to make this series convergent.

4.8.2 Unconditional Moments

$$E(Y_t) = 0$$

$$Var(Y_t) = \frac{\sigma^2}{1 - \phi^2}$$

Note that if $\phi = 1$, $Var(Y_t) = \infty$, which contradicts with covariance stationarity. Additionally,

$$Var(Y_t) = Var(Y_{t-1}) + \sigma^2 > Var(Y_{t-1})$$

which means $Var(Y_t)$ depends on time. Therefore, $|\phi| < 1$ is necessary for covariance stationarity.

4.8.3 Random Walk / Unit Root

An AR(1) with $\phi = 1$ is a random walk / unit root process.

$$Y_t = Y_{t-1} + \epsilon_t$$

$$Y_t = Y_0 + \sum_{i=0}^{t-1} \epsilon_{t-i}$$

Intuition : the shocks have permanent effects.

4.8.4 Conditional Moments

$$E(Y_t|\Omega_{t-1}) = \phi Y_{t-1}$$
$$Var(Y_t|\Omega_{t-1}) = \sigma^2$$

4.8.5 Autocovariance of AR(1)

To obtain this result, we just need to multiply both sides of Y_t by Y_{t-1} and take the expectation. This is called the **Yule-Walker Equation**.

$$\gamma(k) = \phi\gamma(k-1)$$

We just need to know $\gamma(0)$ and we are already able to work out for other autocovariances.

$$\gamma(k) = \phi^k \frac{\sigma^2}{1-\phi^2} = \phi^k \gamma(0)$$
$$\rho(k) = \phi^k$$

This shows how fast/slow the autocorrelation decays in AR(1).

4.8.6 AR(p) Process

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t \ , \ \epsilon_t \sim WN(0, \sigma^2)$$
$$\Phi(L)Y_t = (1 - \phi_1 L - \phi_2 L^2 + ... + \phi_p L^p)Y_t = \epsilon_t \ , \ \epsilon_t \sim WN(0, \sigma^2)$$

Back to our goal, AR(p) process is possibly an approximation of Wold representation

$$B(L) \approx \frac{\Theta(L)}{\Phi(L)} = \frac{1}{\sum_{i=0}^{p} \phi_i L^i} = \frac{1}{\Phi(L)}$$

Note: In MA(q), it is always stationary, we concern about its invertibility. In AR(p), it is always invertible, we concern about its stationarity.

4.9 ARMA Process

Simplest example: ARMA(1,1)

$$Y_t = \phi Y_{t-1} + \epsilon_t + \theta \epsilon_{t-1} \ , \ \epsilon_t \sim WN(0, \sigma^2)$$

We need $|\phi| < 1$ for stationarity (AR) and $|\theta| < 1$ for invertibility (MA).

Generalization:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + ... + \theta_q \epsilon_{t-q}$$

$$\Phi(L) Y_t = \Theta(L) \epsilon_t$$

$$Y_t = \frac{\Phi(L)}{\Theta(L)} \epsilon_t$$

Another approximation for Wold representation!!

4.10 Summary

- We think of C_t as a covariance stationary and ergodic time series process.
- We can approximate any stationary process by a general linear process (Wold's Theorem).
- MA, AR, ARMA naturally arise to achieve this goal, each has different conditions to fulfill.

5 Estimating and Forecasting Cycles

From last section, we concluded that it makes sense to model C_t as a covarinace stationary process. We can then apply Wold's theorem to model it. What question we are asking now is **what order to choose for AR, MA, ARMA**?

5.1 Box-Jenkins Methodology

For MA(q), only the first q autocorrelations are nonzero, so the **ACF** should cut off after lag q.
For AR(p), the autocorrelations may decline gradually, but the **PACF** should cut off after lag p.
If neither ACF/PACF shows clear cut-off, we may try an ARMA model, but starts with low order.
We can also examine residuals for **no remaining serial correlation** (Ljung-Box Q Test) and **normally distributed** (density plot).

Why do we use PACF to decide AR(p)?
AR is a process that the past value of Y_t has effect on present Y_t, directly. Thus, it is reasonable to control for other lags of Y_t and look at the marginal effect of Y_{t-q} on Y_t.

Why do we use ACF to decide MA(q)?
MA is a process that the past errors affect the value of Y_t. In that case, we cannot observe the marginal impact of each period's ϵ on Y_t because we can only observe the value of Y_t. However, the effect is there, so we can only look at the ACF to decide.

5.2 MA(q) Forecast Generalization

The optimal forecasts are:

$$E_T(Y_{T+1}) = E_T(\epsilon_{T+1} + \theta_1\epsilon_T + ... + \theta_q\epsilon_{T-q+1}) = \theta_1\epsilon_T + ... + \theta_q\epsilon_{T-q+1}$$

$$E_T(Y_{T+2}) = E_T(\epsilon_{T+2} + \theta_1\epsilon_{T+1} + \theta_2\epsilon_T + ... + \theta_q\epsilon_{T-q+2}) = \theta_2\epsilon_T + ... + \theta_q\epsilon_{T-q+2}$$
$$E_T(Y_{T+q}) = E_T(\epsilon_{T+q} + \theta_1\epsilon_{T+q-1} + \theta_2\epsilon_{T+q-2} + ... + \theta_q\epsilon_T) = \theta_q\epsilon_T$$
$$E_T(Y_{T+h}) = E_T(\epsilon_{T+h} + \theta_1\epsilon_{T+h-1} + \theta_2\epsilon_{T+h-2} + ... + \theta_q\epsilon_{T+h-q}) = 0 \quad h > q$$

Notice that the longer the forecast horizon, the less we can predict because we have less info from the past errors. When we reach h-period ahead, we can't use past Y or past ϵ to do prediction because all the effects of ϵ from the past are dead.

The forecast errors are:

$$e_{T+1|T} = Y_{T+1} - E_T(Y_{T+1}) = \epsilon_{T+1}$$

$$e_{T+2|T} = Y_{T+2} - E_T(Y_{T+2}) = \epsilon_{T+2} + \theta_1\epsilon_{T+1}$$
$$e_{T+q|T} = Y_{T+q} - E_T(Y_{T+q}) = \epsilon_{T+q} + \theta_1\epsilon_{T+q-1} + \theta_2\epsilon_{T+q-2} + ... + \theta_{q-1}\epsilon_{T+1}$$
$$e_{T+h|T} = Y_{T+h} - E_T(Y_{T+h}) = \epsilon_{T+h} + \theta_1\epsilon_{T+h-1} + \theta_2\epsilon_{T+h-2} + ... + \theta_q\epsilon_{T+h-q} \ , \ h > q$$

Notice that forecast errors are MA(h-1) for $h \leq q$. Beyond that, the error is exactly MA(q). Therefore, MA(q) forecastable up to q periods out of sample. The reason is the same as above, we have no info to do the forecast, so we will get back what we have, which is a bunch of errors.

The forecast error variances are:

$$Var(e_{T+1|T}) = Var(\epsilon_{T+1}) = \sigma^2$$

$$Var(e_{T+2|T}) = Var(\epsilon_{T+2} + \theta_1\epsilon_{T+1}) = \sigma^2(1 + \theta_1^2)$$
$$Var(e_{T+3|T}) = Var(\epsilon_{T+3} + \theta_1\epsilon_{T+3} + \theta_2\epsilon_{T+1}) = \sigma^2(1 + \theta_1^2 + \theta_2^2)$$
$$Var(e_{T+q|T}) = Var(\epsilon_{T+q}+\theta_1\epsilon_{T+q-1}+\theta_2\epsilon_{T+q-2}+...+\theta_{q-1}\epsilon_{T+1}) = \sigma^2(1+\theta_1^2+\theta_2^2+...+\theta_{q-1}^2)$$
$$Var(e_{T+h|T}) = Var(\epsilon_{T+h}+\theta_1\epsilon_{T+h-1}+\theta_2\epsilon_{T+h-2}+...+\theta_q\epsilon_{T+h-q}) = \sigma^2(1+\theta_1^2+\theta_2^2+...+\theta_q^2) \ , \ h > q$$

Notice that for $h < q$, $Var(e_{T+h}) < Var(Y_T)$; $h = q$, $Var(e_{T+h}) = Var(Y_T)$; $h > q$, $Var(e_{T+h}) > Var(Y_T)$, where Y_T is MA(q) process. This is also why the forecast interval fans out as h increases.

Since dynamics wash out quickly, pure MA(q) models not used often in forecasting typically persistent economic data.

5.3 AR(1) with Intercept

$$Y_t = \alpha + \phi Y_{t-1} + \epsilon_t$$
$$E(Y_t) = \frac{\alpha}{1-\phi}$$
$$E_T(Y_T) = \alpha + \phi Y_T$$
$$e_{T+1|T} = \epsilon_{T+1}$$
$$Var(e_{T+1|T}) = \sigma^2$$

Alternatively,

$$\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\hat{\epsilon}_t^2$$

This is just an introduction to AR(1) with intercept. All computations are simple (take expectation and do some cancellation). More on AR will be discussed after this.

6 Forecasting with AR(p) Models

6.1 h-step Ahead and the Chain Rule

Consider AR(1) without intercept,

$$\hat{Y}_{T+1|T} = \phi Y_T$$

$$\hat{Y}_{T+2|T} = \phi^2 Y_T = \phi Y_{T+1|T}$$
$$\hat{Y}_{T+h|T} = \phi^h Y_T = \phi Y_{T+h-1|T}$$

Chain rule shows that in order to do h-step ahead forecast, we can multiply the (h-1)-step ahead forecast by the coefficient. Hence, we only need the first forecast to obtain further forecasts.

6.2 2-step Ahead Forecasting

6.2.1 Plug-in Method

$$\begin{aligned} Y_t &= \alpha + \phi Y_{t-1} + \epsilon_t \\ &= \alpha + \phi(\alpha + \phi Y_{t-2} + \epsilon_{t-1}) + \epsilon_t \\ &= (1+\phi)\alpha + \phi^2 Y_{t-2} + \epsilon_t + \phi\epsilon_{t-1} \end{aligned}$$

Then 2-period ahead forecast is

$$\hat{Y}_{T+2|T} = (1+\phi)\alpha + \phi^2 Y_t$$

Then we plug in the estimates $\hat{\phi}$, $\hat{\alpha}$ (from running a usual AR regression) to get $\hat{Y}_{T+2|T}$. This is a simple method but can be cumbersome.

6.2.2 Iterated Method

We know how to build a 1-step forecast, so it's is easy to build

$$\hat{Y}_{T+1|T} = \hat{\alpha} + \hat{\phi} Y_T$$
$$\hat{Y}_{T+2|T} = \hat{\alpha} + \hat{\phi} \hat{Y}_{T+1|T}$$

This method applies the chain rule to get further forecasts.

6.2.3 Direct Method

We can rewrite the Plug-in Method's equation as

$$Y_t = \alpha^* + \phi^* Y_{t-2} + u_t$$

This can be estimated directly by OLS. The error term is not white noise here and is different from the iterated estimator.

6.3 2-step Forecast Errors

From the direct method above, $u_t = \epsilon_t + \phi\epsilon_{t-1}$, where U is not white noise. Thus, the variance of the error term is

$$Var(u_t) = Var(\epsilon_t + \phi\epsilon_{t-1} = (1 + \phi^2)\sigma^2$$

We need it to construct forecast intervals.

6.3.1 Plug-in Variance Estimation

Same logic as before, use estimates to replace unknowns:

$$\hat{\sigma}_u = \sqrt{(1 + \hat{\phi}^2)\hat{\sigma}^2}$$

We have to run AR(1) regression to get the estimates. It is also hard to generalize beyond AR(1).

6.3.2 Direct Variance Estimation

Just use the regression RMSE;

$$\hat{\sigma}_u = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \hat{u}_t^2}$$

6.3.3 Iterated Forecast Variance Estimation

This is a numerical computation option, I will not include the details here hehehe XP

All p-step ahead forecast is similar to the approach above, just plug in and do the math.

7 Putting the Component Model Together

7.1 Linear Trend + Cycle Model

$$T_t = \mu_1 + \mu_2 t + C_t$$
$$C_t = \phi C_{t-1} + \epsilon_t$$

Lag the first equation, multiply by ϕ and subtract, rearranging we can get:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 t + \epsilon_t$$

If there is trend but we omit it in our model, then it is likely that we get $\hat{\beta} \approx 1$, which is nearly unit root process.

7.2 Seasonal + Cycle Model

$$Y_t = S_t + C_t$$
$$C_t = \phi C_{t-1} + \epsilon_t$$

Lag the first equation, multiply by ϕ and subtract, rearranging we can get:

$$Y_t = \phi Y_{t-1} + S_t - \phi S_{t-1} + \epsilon_t$$

However, lagged dummies are meaningless, so we can omit it. The regression model is thus:

$$Y_t = \alpha_0 + \sum_{i=1}^{s-1} \alpha_i D_{it} + \beta Y_{t-1} + \epsilon_t$$

7.3 Trend + Seasonal + Cycle Model

$$Y_t = T_t + S_t + C_t$$
$$T_t = \mu_1 + \mu_2 t + C_t$$
$$S_t = \sum_{i=1}^{S} \alpha_i D_{it}$$
$$C_t = \phi_1 C_{t-1} + ... + \phi_p C_{t-p} + \epsilon_t$$

This implies the regression model:

$$Y_t = \sum_{i=1}^{S} \alpha_i D_{it} + \gamma t + \beta_1 Y_{t-1} + ... + \beta_p Y_{t-p} + \epsilon_t$$

7.4 Pseudo Out-Of-Sample Forecasting

Only use N observations out of T available for model estimation. Reserve P observations at the end of the sample for evaluation. Produce a series of h-step forecasts for any fixed horizon h and update estimates with additional data as you move through time. We can assess how well can the model forecast the data that was not used in estimation.

7.4.1 Fixed Estimation Window

- Includes only the first N observations to estimate model parameters once and for all.
- Usually used when estimation costs are very high.
- Not desirable when model instability over time is suspected.

7.4.2 Expanding Estimation Window

- Includes the first N observations to estimate model and produce the first forecast.
- Next period, the same information plus an extra observation is used to update the model.
- Needs DGP to be stationary, otherwise, it produces biased estimates. But we can have more data to outweigh the bias.

7.4.3 Rolling Estimation Window

- Includes most recent N observation sot estimate the mdoel.
- Next period, the observation 1 is dropped and observation N+1 is added to update the mdoel etc.
- Used when we are not sure that the DGP is stationary and don't want to contaminate forecasts with older potentially irrelevant data.
- Gives higher parameter uncertainty.

8 Standard Errors of OLS Estimates in TS Regression

8.1 Classical Variance and Standard Error

$$\hat{\beta} = \frac{\sum_{t=1}^{T} Y_t X_t}{\sum_{t=1}^{T} X_t^2} = \beta + \frac{\frac{1}{T}\sum_{t=1}^{T} X_t e_t}{\frac{1}{T}\sum_{t=1}^{T} X_t^2}$$

$$\hat{\beta} = \beta + \frac{\sum_{t=1}^{T} v_t}{TVar(X_t)}$$

$$AVar(\hat{\beta}) = \frac{Var(\sum_{t=1}^{T} v_t)}{T^2 Var(X_t)^2}$$

$$Var(\sum_{t=1}^{T} v_t) = \sum_{t=1}^{T} Var(v_t) + \sum_{j\neq i}^{T} Cov(v_t, v_j)$$

If the observations are independent, then $\sum_{j\neq i}^{T} Cov(v_t, v_j) = 0$, then :

$$AVar(\hat{\beta}) = \frac{Var(X_t e_t)}{TVar(X_t)^2}$$

This means under independence assumption, $X_t e_t$ are uncorrelated with $X_j e_j$, this happens in 1-step-ahead regressions, when errors are unforecastable. Therefore, OLS estimator is consistent when doing 1-step-ahead forecast.

The standard error based on the formula we derived is :

$$SE(\hat{\beta}) = \sqrt{\frac{\hat{\sigma_e}^2}{T\hat{\sigma}_X^2}}$$

Robust standard error is :

$$SE(\hat{\beta}) = \sqrt{\frac{\hat{\sigma_{v_t}}^2}{T\hat{\sigma}_X^2}}$$

In summary, in 1-step-ahead time series models where errors are unforecastable, usign robust errors is appropriate.

8.2 Regression with Correlated Errors

In some models, we would expect errors to be correlated. Then classical and robust SE estimates will not be appropriate.

Define the adjustment factor:

$$f_T = \frac{Var(\sum_{t=1}^{T} v_t)}{TVar(v_t)}$$

The general formula becomes:

$$AVar(\hat{\beta}) = \frac{Var(v_t)}{TVar(X_t)^2} f_T$$

If v_t are uncorrelated, then $f_T = 1$. Otherwise it adjusts for serial correlation.

Intuition: If v_t correlated, $Var(\sum_{t=1}^{T} v_t) > TVar(v_t)$ becasue of covariance term. Thus, $f_T > 1 \Rightarrow AVar(\hat{\beta})$ increases.

$$f_T = \frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{T} \rho(t-s) = 1 + 2 \sum_{s=1}^{T-1} (\frac{T-s}{T}) \rho(s)$$

As T grows large, we have:

$$f_T = 1 + 2 \sum_{s=1}^{T-1} (\frac{T-s}{T}) \rho(s) \longrightarrow 1 + 2 \sum_{s=1}^{\infty} \rho(s) = f$$

$$AVar(\hat{\beta}) = \frac{Var(v_t)}{TVar(X_t)^2} f$$

8.3 HAC Standard Errors

We will multiply the usual variance estimate by f. We can use sample autocorrelations to estimate but for longer lags, estimates get progressively worse. Thus, we can truncate the infinite sum at some lag m.

8.3.1 Unweighted HAC Estimator

Choose a truncation parameter m, then compute:

$$\hat{f} = 1 + 2\sum_{s=1}^{m} \hat{\rho}(s)$$

However, this will sometimes deliver negative variance estimates (since $\hat{\rho}(s)$ can be negative.

8.3.2 Weighted HAC Estimator (Newey-West)

Choose a truncation parameter m, but smooth the sum using weights:

$$\hat{f} = 1 + 2\sum_{s=1}^{m} (\frac{m-s}{m})\hat{\rho}(s)$$

The estimator always stays non-negative.

How to choose m?

- Schwert's max lag: $m = 12(\frac{T}{100})^{\frac{1}{4}}$
- Stock and Watson default: $m = 0.75T^{\frac{1}{3}}$
- Trend/Seasonal default: $m = 1.4T^{\frac{1}{3}}$

To summarize,

- errors are serially uncorrelated, use robust errors
- errors are correlated, use HAC errors
  - for dynamic regression (cycle), use SW default
  - for pure trend/seasonal regression, use trend/seasonal default

8.4 h-step-ahead forecast regression

$$Y_t = \alpha + \beta Y_{t-h} + e_t$$

If the model correctly specified, the forecast error e_t is going to be MA(h-1). We need to adjust for serial correlation - use HAC errors, set the truncation parameter to forecast horizon minus one: $m = h - 1$.

8.5 Joint Hypothesis Tests

We need to use the appropriate standard error when performing joint hypothesis tests (F-test).

9 Model Selection

There are tradeoff in model selection: estimation error vs. model misspecification. More variables = more estimation error; fewer variables = more chance to miss important preidctors.

$adjusted - R^2$ gives not enough penalty.

9.1 Sequential Tests

We can use sequential t-tests or sequential F-tests to choose variables. But it is not designed to select best forecast model and can perform badly. Usually, F-tests are preferred to t-tests in presence of high correlation among regressors (imperfect multicollinearity).

9.2 BIC

$$P(M_1|D) \propto exp(-\frac{BIC}{2})$$

$$BIC = Tln(\frac{SSR}{T}) + kln(T)$$

The model with the smallest BIC is the model that is most likely to be true.

BIC models have consistency property. However, BIC selection is not specifically designed to produce a good forecast.

BIC assumes conditional homoskedasticity.

9.3 AIC

$$AIC = Tln(\frac{SSR}{T}) + 2k$$

We can see that BIC put a harsher penalty on model size.

AIC is an approximately unbiased estimate of the MSFE. Hence, AIC is designed to makes the best forecast. Unlike BIC, AIC is not consistent, but it will asymptotically pick the best forecasting model.

AIC assumes conditional homoskedasticity.

9.4 Predictive Least Squares

Compute true out-of-sample forecasts, the assocaited forecast errors, and pick the model with the smallest value of the loss function. This is similar to the validation approach in machine learning.

The out-of-sample forecast errors are:

$$\tilde{e} = T_t - \hat{Y}_t$$

$$PLS = \sqrt{\frac{1}{P}\sum_{t=M+1}^{T} \tilde{e}_t^2}$$

We then select the model with the smallest PLS.

Disadvantages:

- tends to overestimate true MSFE.
- tends to over-parsimonious.
- very sensitive to the choice of P

10 Forecasting with Regression Models

If the conditional mean of Y_t depends on present period X_t, then we need also an h-period ahead forecast for X.

10.1 Plug-in Method

Suppose the model for Y_t is:

$$\hat{Y}_{T+h|T} = \alpha + \beta X_{T+h}$$

and we have a model for X_t:

$$\hat{X}_{T+h|T} = \hat{\gamma} + \hat{\phi} X_T$$

Then we can plug in $\hat{X}_T$ to predict Y.

10.2 Direct Method

$$E(Y_t|\Omega_{t-h}) = \alpha + \beta E(X_t|\Omega_{t-h}) = \alpha + \beta(\gamma + \phi X_{t-h}) = \mu + \theta X_{t-h}$$
$$Y_t = \mu + \theta X_{t-h} + \epsilon_t$$

We can obtain forecast from the h-step ahead regression directly.

10.3 ADL Models

We can add lags of X and Y together to improve the performance of our forecasts.

$$Y_t = \mu + \alpha_1 Y_{t-1} + ... + \alpha_q Y_{t-q} + \beta_1 X_{t-1} + ... + \beta_k X_{t-k} + e_t$$

10.4 Granger Causality in Mean

If X helps predicting Y, then we can say that X Granger-causes Y. This doesn't mean that X causes Y, only means that one helps predicting another.

H_0: $\beta_1 = ... = \beta_k = 0$ (i.e. X doesn't cause Y)
H_1: at least one $\beta \neq 0$
The test uses F-statistics by using the HAC errors with SW default)

Granger-causality may not work out-of-sample and non-stable over time.

Even though the no-causality hypothesis is not rejected, it does not necessarily mean X will not at all help with forecasting Y. We need to use AIC to select the best ADL models even if the test is not rejected.

It is a common practice to use economic leading indicators in our ADL models.

11 Forecast Combination

Forecast combination means we take different forecast values and combine them together. One of the most simple approaches is taking the simple average of all forecasts. In machine learning language, this is called ensemble learning.

Suppose we have two forecasts of Y, f_1 and f_2, the weighted forecast would be:

$$f = wf_1 + (1 - w)f_2$$

The variance of the averaged forecast would be:

$$Var(f) = w^2\sigma_1^2 + (1 - w)^2\sigma_2^2$$

By taking the FOC to minimize $Var(f)$ wrt w, we have:

$$w^8 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

If we have multiple forecasts, then:

$$w_j^* = \frac{\sigma_j^{-2}}{\sigma_1^{-2} + ... + \sigma_N^{-2}}$$

We can see that if the variance of the forecast is lareg, its respective weight will be smaller.

11.1 Bates-Granger Combination

Assumes uncorrelated forecasts, but can work well in practice.

$$w_j^* = \frac{\hat{\sigma}_j^{-2}}{\hat{\sigma}_1^{-2} + ... + \hat{\sigma}_N^{-2}}$$

To implement:

1. produce a series of POOS forecasts and associated forecast errors
2. compute forecast variances and invert them
3. normalize by their sum to get weights for all models

11.2 Granger-Ramanathan Combination

GR combination allows for correlation between forecasts.

$$Y_t = \beta_1 f_{1t} + ... + \beta_N f_{Nt} + e_t$$

We need to use a constrained regression:

1. omit intercept (can include intercept when there is bias)
2. force nonnegative betas
3. make coefficients sum to 1

11.3 Bayesian Model Averaging

$$P(M|D) \propto exp(-\frac{BIC}{2})$$

These probabilities can be used to form forecast weights.

$$w_m^* = exp(-\frac{BIC_m}{2})$$

$$w_m = \frac{w_m^*}{\sum_{m=1}^M w_m^*}$$

If one model is more likely to be the true model, then its forecast receives more weights.

However, using BIC to compute weights can lead to extremely large values. It is common to adjust it for computation. Let $\Delta BIC_m = (BIC_m - BIC^*)$ be the BIC difference between that of model m and the best one. Then

$$w_m^* = exp(-\frac{\Delta BIC_m}{2})$$

$$w_m = \frac{w_m^*}{\sum_{m=1}^M w_m^*}$$

This will be more numerically stable.

11.4 Weighted AIC

$$w_m \propto exp(-\frac{AIC}{2})$$

There is no strong theory for this but it is common in practice. We can do the same thing in BMA here

$$w_m^* = exp(-\frac{\Delta AIC_m}{2})$$

$$w_m = \frac{w_m^*}{\sum_{m=1}^M w_m^*}$$

If criteria differ, BMA and WAIC will give different weights to different models. Weights will go to zero for models sufficiently different from the best (AIC/BIC difference of 10 and above).

12 Forecast Evaluation

Recall that **Optimal Forecast is unbiased**, which means $E(e_{t+h|t}) = 0$. We can regress the forecast errors on a constant.

$$\hat{e}_t = \alpha + u_t$$

We should find that $\alpha = 0$.

If serial correlation is present, we can fit an appropriate MA model to soak up the seiral correlation, then test for the constant

$$\hat{e}_t = \alpha + b_1\hat{e}_{t-1} + ... + b_2\hat{e}_{t-h} + u_t$$

Recall that **Optimal Forecast 1-step ahead forecast error is white noise**. We can examine the ACF and Ljung-Box statistics for joint tests of autocorrelations.

$$\hat{e}_t = \alpha + u_t$$

Do Q-test after this regression.

Recall that **Optimal Forecast h-step errors are at most MA(h-1)**. We can plot the estimated ACF of forecast errors, examine whether autocorrelatons beyond lag (h-1) are significant.

Recall that **Optimal Forecast have h-step ahead errors with variances nondecreasing in h, and converging to the unconditional variance of the process**. The key property of optimal forecast errors is unforecastability. The information set at time t, and the optimal forecasts themselves, should not help predicting optimal forecast errors (i.e. nothing left in the error for us to do forecast, we extracted all info).

$$e_{t+h|t} = \alpha + \beta \hat{Y}_{t+h|t} + \epsilon_{t+h}$$
$$\alpha = 0,\ \beta = 0$$

12.1 Mincer-Zarnowitz Regression

$$Y_{t+h} = \alpha + \beta \hat{Y}_{t+h|t} + u_t$$

Perform a joint test of $H_0:\ \alpha = 0$ and $\beta = 1$. If reject, there is systematic bias in the forecast. Forecast errors are correlate to forecast. We need to use the correct standard errors to do the test.

12.2 Meese-Rogoff Puzzle

More complex forecasts may not do even better than a simple naive forecast.

12.3 Statistical Comparison of Forecast Risk

When examining accuracy measures of different model forecasts, we may not be sure whether the differences are significant enough to declare one model superior. We thus may wish to test for risk equality for models a and b: $E(L(e^a_{t+h|t})) = E(L(e^b_{t+h|t}))$. This can be written as $E(d_t) = E(L(e^a_{t+h|t}) - L(e^b_{t+h|t}))$

12.4 Diebold Mariano Test

$$d_t = L(e^{(1)}_{t+h|t}) - L(e^{(2)}_{t+h|t})$$

Assumption for DM: d_t is covariance stationary. Under this assumption,

$$\bar{d} = \frac{1}{P}\sum_{t=1}^{P} d_t \xrightarrow{\text{P}} E(d_t)$$

$$\sqrt{P}(\bar{d} - E(d_t)) \xrightarrow{\text{d}} N(0, \sigma^2_{\bar{d}})$$

The test statistics:

$$DM_{12} = \frac{\bar{d}}{\hat{\sigma}_{\bar{d}}/\sqrt{P}} \sim N(0,1)$$

We are basically doing a t-test of a mean being equal to zero.

12.4.1 Issues with the DM Test

- Need large sample size
- neglect estimation errors
- cannot be applied to nested models with expanding window scheme

12.4.2 Some Fixes

When we have finite sample, we can do a correction:

$$t_{HLN} = (1 + P^{-1}(1-2h) + P^{-2}h(h-1))^{1/2} t_{DM}$$

where h is the forecast horizon, P is the number of POOS forecasts.

13 Volatility Modelling and Forecasting

Take a mean-zero series, assume that the error term is unforecastable. The conditional variance of Y_t is then:

$$Var(Y_t|\Omega_{t-1}) = E([Y_t - E(Y_t|\Omega_{t-1})]^2|\Omega_{t-1}) = E(\epsilon_t^2|\Omega_{t-1})$$

The variance of Y_t, squared error, could be potentially be forecastable. If the squared white noise term is forecastable, then the conditional variance is time varying and serially correlated. Although we cannot predict the sign of the change in Y_t, we can predict the magnitude.

13.1 ARCH

Volatility can look serially correlated and could be an AR process. This may suggest white noise with autoregressive dynamics in the conditional variance.

Consider a simple example:

$$Y_t = \mu + \epsilon_t,\ \epsilon_t|\Omega_{t-1} \sim N(0, \sigma_t^2)$$

$$\sigma_t^2 = Var(\epsilon_t|\Omega_{t-1}) = \omega + \alpha\epsilon_{t-1}^2,\ \omega > 0,\ 0 \leq \alpha < 1$$

$$\sigma^2 = \frac{\omega}{1-\alpha}$$

$\alpha > 0$ means conditional variance is high when the lag of squared error is high. $\alpha = 0$ returns back to usual constant variance case. $\omega > 0$ to keep positive variance. $\alpha \neq 0$ to keep finite variance.

We can rewrite and plug in back to the ARCH equation:

$$\sigma_t^2 = \sigma^2 + \alpha(\epsilon_{t-1}^2 - \sigma^2)$$

The conditional variance is a combination of the unconditional variance and the deviation of the squared error term from its average value.

13.1.1 Variance Forecast

Given parameter estimates, the estimated conditional variance for period t is:

$$\hat{\sigma}_t^2 = \hat{\omega} + \hat{\alpha}\hat{\epsilon}_{t-1}^2 = \hat{\omega} + \hat{\alpha}(Y_{t-1} - \hat{\mu})^2$$

We can use the conditional standard deviation to obtain more realistic forecast intervals for the conditional mean forecast:

$$\hat{Y}_{t+1|t} \pm Z_{\alpha/2}\hat{\sigma}_{t+1|t}$$

The forecast interval will be narrower in the calm period, and wider in the volatile period.

13.1.2 Detecting ARCH Effects

After modelling the conditional mean of Y_t, check for serial correlation in squared residuals through ACF/Ljung-Box stats.

We can also use Engle's LM test: test m coefficients jointly zero in the regression

$$\epsilon_i^2 = \beta_0 + \beta_1\epsilon_{i-1}^2 + ... + \beta_m\epsilon_{i-m}^2 + u_i$$

This is an F-test for β. If H_0 is rejected, then there is ARCH effect.

13.1.3 ARCH(p) Order Selection

We can check the PACF of squared residuals from the mean model. We can also use AIC/BIC to make the final model selection.

13.2 GARCH

Consider GARCH(1,1) model:

$$Y_t = \epsilon_t,\ \epsilon_t|\Omega_{t-1} \sim N(0, \sigma_t^2)$$

$$\sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

$$\omega > 0,\ \alpha \geq 0,\ \beta \geq 0,\ \alpha + \beta < 1$$

In practice, we never consider GARCH order higher than GARCH(2,2).

We can work out that:

$$\sigma^2 = \frac{\omega}{1 - \alpha - \beta}$$

This implies that ϵ_t^2 is ARMA(1,1):

$$\epsilon_t^2 = \omega + (\alpha + \beta)\epsilon_{t-1}^2 - \beta v_{t-1} + v_t$$

GARCH model can capture volatility clustering and leptokurtosis observed in actual data. But it cannot capture the asymmetric effect on volatility of positive vs negative returns.

13.3 TGARCH

$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\epsilon_{t-1}^2 + \gamma\epsilon_{t-1}^2 I(\epsilon_{t-1} < 0)$$

The last term equals to 1 if the last period shock was negative. This makes the ARCH effect asymmetric: if error is positive, the effect is α, if error is negative, the effect is $\alpha + \gamma$.

13.4 Exponential GARCH

$$ln(\sigma_t^2) = \omega + \beta ln(\sigma_{t-1}^2) + \alpha\Big|\frac{\epsilon_{t-1}}{\sigma_{t-1}}\Big| + \gamma\frac{\epsilon_{t-1}}{\sigma_{t-1}}$$

Volatility depends on both absolute magnitude and sign of shocks. Allows for asymmetric response throught the last term.

13.5 GARCH-in-Mean

$$Y_t = \beta_0 + \beta_1\sigma_t^2 + \epsilon_t$$
$$\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha\epsilon_{t-1}^2$$

The mean is correlated to the volatility.

13.6 Realized Variance

Unlike mean, we don't observe volatility, so we have to estimate it somehow. The volatility process is often modeled as follows:

$$IV_t = \int_{t-1}^{t} \sigma_s^2 ds$$

The IV above is the integrated volatility over, say, a day.

However, IV is still not observable, but a consistent estimator is realized variance:

$$RV_t = \sum_{i=1}^{M} r_{t,i}^2 = \sum_{i=1}^{M}(ln(P_{t-1+i\Delta}) - ln(P_{t-1+(i-1)\Delta}))^2 \xrightarrow{P} IV_t$$

Typically, RV is computed using 5-minute returns.

13.7 QLIKE Loss Function

QLIKE loss function is applicale only for strictly positive random variables.

$$OLIKE = \frac{Y}{\hat{Y}} - log(\frac{Y}{\hat{Y}}) - 1$$

Y is the proxy for true variance, $\hat{Y}$ is the variance forecast.

13.8 Robust Regression

In high frequency variance context, the data is prone to "noise" in the form of outliers and sampling errors. To mitigate these errors, we can use robust regression. Intuitively, the procedure weighs observations using a recursive procedure so that leverage ponts receive small or even zero weights thus not affecting estimates as much.

13.9 HAR Model

AR models fitted directly to RV can be as successful as GARCH but are quite cumbersome. Hence, we can have a more parsimonious model for RV.

Define the multi-period realized variance as follows:

$$RV_{t,t+h} = \frac{1}{h}(RV_{t,t+1} + ... + RV_{t,t+h})$$

Then, the HAR models daily RV using first lags of daily, weekly, and monthly RV:

$$RV_{t+1} = \alpha + \beta_D RV_t + \beta_W RV_{t-5,t} + \beta_M RV_{t-22,t} + \epsilon_{t+1}$$

Intuition: This model summarize high persistence in volatility through considering past daily, weekly and monthly averages, which avoids long lags of the daily AR model.

HAR model can be at least as good to GARCH and often much better, while the model is very simple.

14 Forecasting in Presence of Unit Roots and Cointegration

14.1 Unit Roots

Consider an AR process:

$$\Phi(L)Y_t = \epsilon_t$$

It is said to have a unit root if:

$$\Phi(1) = 0$$

Example:

$$Y_t = Y_{t-1} + \epsilon_t$$

The presence of unit roots means the optimal forecast for h-period later is today's value (i.e. useless forecast).

14.2 Random Walk with Drift

AR(1) with intercept and unit root:

$$Y_t = \alpha + Y_{t-1} + \epsilon_t$$

This is identical to a trend plus random walk. We can rewrite as:

$$Y_t = Y_0 + \alpha_t + \sum_{j=1}^{t} \epsilon_j$$

The presence of random walk with drift means the optimal forecast for h-period later is today's value plus trend (i.e. useless forecast).

If Y is a random walk process, it is not stationary.

14.2.1 General Procedure

If Y_t has a unit root, transform by differencing:

$$Z_t = \Delta Y_t = Y_t - Y_{t-1}$$

The transformation eliminates the unit root, so Z_t is stationary. Then, build a forecasting model for Z_t.

14.3 ARIMA(p,d,q) Models

$$Phi(L)(1-L)^d Y_t = \alpha + \Theta(L)\epsilon_t$$
$$Z_t = (1-L)^d Y_t$$

d is called the order of integration. The series Y_t becomes stationary after being differenced d times.

14.4 Spurious Regression

Caused by unit roots. Take two independent random walks, they seem to have a high correlation. We can either first-difference the series if unit root is suspected, or include lags of dependent variable.

14.5 Cointegration

Sometimes two or more integreated series move together in a regular way over the long run as if they have a common trend. If taking a linear combination of such series reduces the order of integration, the series are said to be cointegrated.

Y_t and X_t are cointegrated if $Z_t = Y_t - \theta X_t$ is stationary.

If Y and X are both I(1) and cointegrated, then the optimal regression for modelling Y is:

$$\Delta Y_t = \alpha + \rho Z_{t-1} + \beta_1 \Delta Y_{t-1} + ... + \beta_p \Delta Y_{t-p} + \phi_1 \Delta X_{t-1} + ... + \phi_q \Delta X_{t-q} + \epsilon_t$$

$$Z_t = Y_t - \theta X_t$$

This is a dynamic regression in differences augmented with the error-correction term Z_{t-1}.