Logical Intuitions

Quantitative methods for intuitive understandings

ARIMAX model for stationary time series

Posted on **2010/09/21**

3 Votes

Prerequisite: Review of ordinary regression and preview of time series, Generalized Linear Regression

Why?

When we talk about time series, it means a series of process affected by its own past state. Either it's a Markov process which is determined only by its previous states, or not, we have to learn something more to build a correct estimation model for this, because we can't use covariance matrix method to construct it. That's where (partial) autocorrelation comes in. Fortunately, they are easy to compute. (^_^)

In this post, let's take a look at autoregressive integrated moving average with exogenous variables. For stationary time series, this one model covers everything to be considered.

Autocorrelation and partial autocorrelation

As mentioned above, time series requires something other than covariance matrix because it is assumed to be a combination of previous states, at least, partially. Well, you may remember a basic rule $a = \rho^*_{ab} b + \rho^*_{ac} c$ Thus, we have to compute partial autocorrelation of its previous states.

Autocorrelation computation procedures

*(1) Prepare a (n * k+1) matrix of $y_t,\ y_{t-1},\ \cdots,\ y_{t-k}$*

(2) Standardize this matrix by dividing each columns by standard deviations after subtracting means from each column. $x_{ij} = \frac{y_{ij} - \bar{y_j}}{\sigma_j}$ Let this standardized matrix X.

(3) Now, we can compute autocorrelation matrix. $\Gamma = \frac{1}{n-1} X^T X$

(4) This autocorrelation matrix will be a (k+1, k+1) matrix like;

$$\Gamma = \begin{pmatrix} 1 & \rho_{01} & \rho_{02} & \cdots & \rho_{0k} \\ \rho_{10} & 1 & \cdots & \cdots & \rho_{1k} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & \vdots \\ \rho_{k0} & \rho_{k1} & \cdots & \cdots & 1 \end{pmatrix}$$

Partial autocorrelation computation procedures

(1) Take the first column of Γ, leaving out 1 and let this vector P. It will be a (k, 1) vector like;

$$P = \begin{pmatrix} \rho_{10} \\ \rho_{20} \\ \vdots \\ \rho_{k0} \end{pmatrix}$$

(2) Take Γ, leaving out its last row and column. It will be a (k, k) matrix. Let this matrix Γ.*

(3) Now, we can compute partial autocorrelation vector. $\Phi = (\Gamma^*)^{-1} P$

(4) This partial autocorrelation vector will be a (k, 1) vector like; $\Phi = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_k \end{pmatrix}$

You may wonder why we need (partial) aucorrelation matrix, not covariance matrix. The reason will be provided in additional information below. For now, let's just go ahead with consensus of using it.

Applications

(1) Model identification

Once autocorrelations are computed, then we can graph them and look at its shape. It is called **correlogram.** Shape of correlogram tells us the property of time series, and thus what kind of model to be used. It is summarized below;

If correlogram is…

(1) Exponentially decaying to zero: AR(p)

(2) Bouncing positive and negative, converging to zero: AR(p)

(3) Essentially zero with 1 or more spikes: MA(q)

(4) Decaying after a few lags: ARMA(p, q)

(5) All (close to) zero: White noise

(6) Spikes in some intervals: Seasonality exists

(7) No decay to zero: Non-stationary data

There are some other methods for model selection such as AIC (Akaike's Information Criterion) and FPE (Akaike's Final Prediction Error). You can look into them if you want. I think it's enough, though. Well, you know, employing appropriate assumptions and variables is much more important for economists and analysts to build a sense-making model.

(2) Order identification for AR(p)

Once it's found AR(p) is good for the data, then we have to identify p. Graph partial autocorrelation and look at its shape. It is called **partial correlogram.** The *appropriate order p is 1 point before sudden and significant slowdown to zero.* For example, if partial correlogram reads 0.9, 0.8, 0.78, 0.21, 0.18, etc... over the lags, then the order p should be 3.

(3) Order identification for MA(q)

Order of moving average model is determined in the same way as AR(p), but using correlogram.

(4) Confidence interval for randomness test and MA(q) order

When correlogram is hanging around zero (case 5 above), we can make it sure whether it is random or not by seeing whether it's within confidence interval or not. **However, note here that all random economic variables don't necessarily stay within this range even if they are random or following moving average model;** that's why Hirotogu Akaike developed AIC and FPE to automatize model selection.

Confidence interval of autocorrelation for randomness: $\pm\frac{z_{1-0.5\alpha}}{\sqrt{n}}$ *where z is z-score for correspondent confidence interval.*

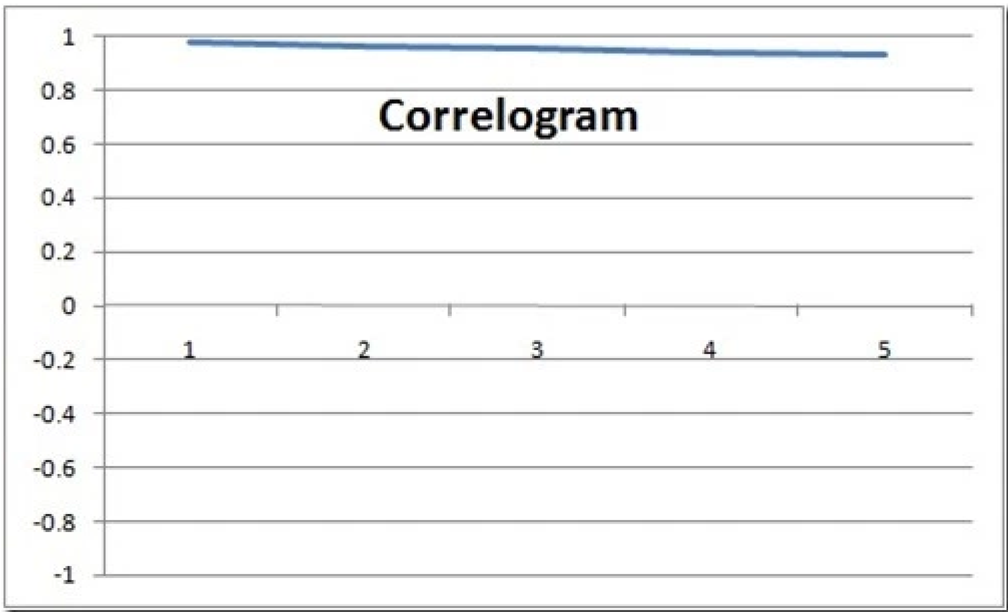
And we can do the same thing for order q of moving average model.

Confidence interval of autocorrelation for MA(q): $\pm z_{1-0.5\alpha}\sqrt{\frac{1}{n}(1+2\sum_{i=1}^{k}y_i^2}$

But, personally, I think having a look at (partial) correlogram is enough, though. And finally, here is 95% confidence interval of partial autocorrelation. If you really want to build a model to explain the data as much as possible, then you'll have to determine order p as the point just before slowdown into this confidence interval.

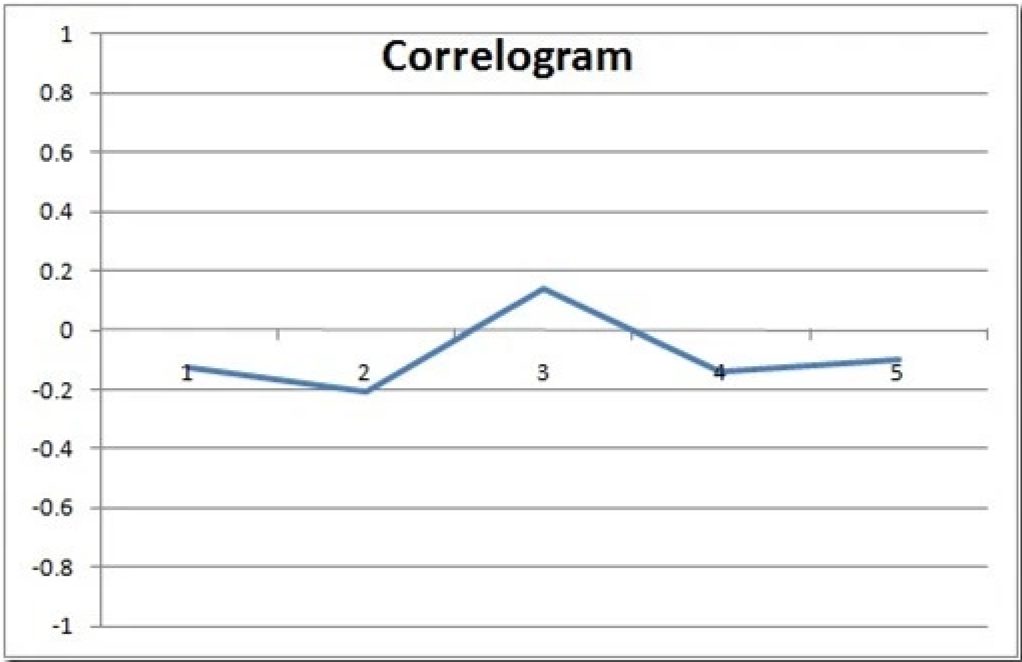
95% Confidence interval of partial autocorrelation: $\pm\frac{2}{\sqrt{n}}$

Example



It is a correlogram computed from daily stock **prices** of Microsoft over 2008. I computed it for only 5 lags but you can see it's obviously near 1 and very slowly decreasing. It's one of common patterns in non-stationary time series. No matter you know common patterns or not, we can conclude stock price is non-stationary, since correlogram looks obviously non-zero.

And here is another correlogram.



It is a correlogram computed from daily **returns** of stock of Microsoft over 2008. As mentioned in the article "Review of ordinary regression and preview of time series," *it's essentially stock price data integrated of 1 order.* You may see it's fluctuating around zero. Therefore, we can conclude returns of stock price is stationary and white noise.

Note: As daily return of stock price is a white noise, we can conclude stock price is a random walk process. Because, as you must know, today's price is sum of yesterday's price and today's daily return. It can be written as $y_t = y_{t-1} + \epsilon_t$ where ε is white noise. Therefore, we can say it is a random walk.

Anyway, once we have partial autocorrelation vector, then left part is extremely easy. Let's go into the model.

Concepts

Now, let's have a look at ***autoregressive integrated moving average model with exogenous variables.*** Don't worry for its long title. While very popular and commonly used to explain time series, *it's simply a combination of 3 models:* autoregressive model using previous states; moving average model using past residuals; and ordinay regression model using external variables, on integrated time series. It's usually written as **ARIMAX(p, d, q, b)** which is title of this post. Below are concepts of autoregressive and moving average model.

AR(p) model

Autoregressive model assumes Markov process; i.e. current state (or value) is totally determined only by past states. It is expressed as $y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$

MA(q) model

Note **moving average here is nothing to do with simple/exponential/weighted moving average, a technical indicator for stocks/bonds/commodities/currency rates, suggested by Joseph Granville.** Well, actually, I think it shouldn't be called "moving average" because it doesn't use means of data varying over time as a regression coefficient. Thus, its name "moving average" implies nothing about the model, only confusing students. Anyway, moving average model in econometrics assumes time series determined by residuals. So, it's written as $y_t = \theta_{t-1}\epsilon_{t-1} + \theta_{t-2}\epsilon_{t-2} + \cdots + \theta_{t-q}\epsilon_{t-q} + \epsilon_t$ As you may see, since it doesn't have any other variables, *it is automatically assuming that current change per step is determined by past changes per step.*

Finally, ARIMAX(p, d, q, b) is…

Usually, after integration of I(d), AR(p) is computed and then MA(q) follows. Finally, b external variables are introduced and ordinary linear regression is computed. That's ARIMAX(p, d, q, b) model. Sometimes we need to adjust data for seasonality; it will be introduced as additional information in this article.

Computation

Step 1: Compute autocorrelation of data and graph correlogram to determine the model. If needed, integrate it of some order. In this case, let's say correlogram is decaying after a few lags (case 4) for data integrated of 1 order. Therefore, we need ARMA(p, q) model for data of I(1); which is ARIMA(p, d, q).

Step 2: Compute partial autocorrelation and graph partial correlogram to determine order p and q.

Step 3: Construct AR(p). Coefficients are, thankfully, partial autocorrelations themselves!!!

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

Step 4: Pick up residuals and compute its autocorrelation and partial autocorrelations. It should be little bit easier because we already know q.

Step 5: Construct MA(q). It's virtually computing AR(p) once again for residuals term for order q. Of course, coefficients are residuals' partial autocorrelations.

$$\epsilon_t = \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \nu_t$$

Step 6: Now we have ARIMA(p, d, q). Finally, compute ordinary linear regression for ν

$$\nu_t = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_b x_b + \omega_t$$

Step 7: Combine AR(p), MA(q), and OLR. Then, it's ARIMAX(p, d, q, b).

$$y_t = \sum_{i=1}^{p}(\phi_i y_{t-i}) + \sum_{j=1}^{q}(\theta_j \epsilon_{t-j}) + \sum_{h=1}^{b}(\beta_h x_h) + \omega_t$$

Test

After building ARIMAX model, we need to test whether our model is good enough, leaving out any unnecessary variables. For this purpose, we have to check *randomness in* ***residuals.*** If residual is random, that means it is white noise; i.e. our model works a nice job. Test procedure is as below;

Ljung-Box test

Ljung-Box test is implemented to test randomness in a series.

Step 1: Compute ARIMAX(p, d, q, b) and autocorrelations of residual term ε in ARIMAX model.

Step 2: Compute Q statistic; $Q = \frac{n(n+2)}{n-k}\sum_{i=1}^{k}\rho^2$

Step 3: Compare Q with Chi-square; $Q \sim \chi^2_k$ *If Q exceeds Chi-square, then our model is not enough.*

Additional information

(1) Reason for (partial) autocorrelation matrix

Now, let's look into why for autocorrelation matrix. It's simple. *Well, you may remember, for WLS, we had to divide residual term by sample standard deviation; it was done to correct heteroskedasticity.* By having a detail look, you may see we are doing the same thing for time series by deriving autocorrelation matrix; what we did above is essentially $\frac{1}{n-1}\sum_{i=1}^{n}\sum_{j,h=0}^{p}(\frac{x_{ij}-\bar{x_j}}{\sigma_j}\frac{x_{ih}-\bar{x_h}}{\sigma_h})$ Therefore, we are dividing sample variance or covariance by standard deviation (or standard co-deviation).

In other words, ***we are correcting heteroskedasticity in time series. We have to always do it, because autovariance in time series is always heteroskedastic;*** it would be close to variance for 1 or 2 lags, but very volatile in random for more lags. Therefore, we have to correct heteroskedasticity in the same way as we did for White Least Square on volatile variance in random.

(2) Conditions of AR(p) and MA(q)

As ordinary or generalized linear regression did, time series models also have their conditions. It is as below;

AR: $\phi_1 + \phi_2 + \cdots + \phi_p = 1$

MA: $|\theta_i| < 1$

(3) Detrending

Trending time series are often non-stationary. Thus, we have to detrend it in these cases *by taking difference of orders: 1 order for linear trend; 2 orders for quadratic trend.* Taking difference is a way of making data stationary and can be done in many manners such as ***taking logarithms, shifts in percent, difference in value itself.***

(4) Seasonality adjustment

If time series has seasonality, we have 2 choices. *Adjust data for seasonality, or include seasonal term in the model.* Whatever choose, we have to first, if any, find out seasonality, because it is sometimes the case we don't know the length of cycle. Here is how to do it.

Detecting seasonality

(1) If correlogram shows spikes at some specified intervals, then that interval is the length of cycle; *e.g. when we have yearly-cyclical time series which is monthly data, then correlogram will spikes at lags of 12, 24, 36…*

(2) Draw a run chart; *compute sample mean and subtract it from each observation. If there is any cycles, then it will show cyclical updowns.*

Dealing with seasonality

(1) Include seasonal term in the model; e.g. when cycle is 1 year and we have monthly data, include y_{t-12} in the model. It is called **SARIMA** model.

(2) Make a new time series by adjusting seasonality and build a model for this new data. For adjustment, see below.

Seasonal adjustment

Step 1: Compute simple moving average in same manner as a technical indicator in financial markets, suggested by Joseph Granville. It is sometimes written as SMA(g). For example, if you have quarterly data showing yearly seasonality (g = 4), then compute average of 4 observations. Setting this average m, then it will be like the followings;
$m_1 = \frac{1}{4}(y_1+y_2+y_3+y_4),\ m_2 = \frac{1}{4}(y_2+y_3+y_4+y_5),\ m_3 = \frac{1}{4}(y_3+y_4+y_5+y_6),\ \cdots$

Step 2: Pair this moving average with **central observation of cycle length.** For example, when g = 5, then $m_1 \to y_3,\ m_2 \to y_4,\ m_3 \to y_5,\ \cdots$ If g = 9, then $m_1 \to y_5,\ m_2 \to y_6,\ \cdots$

Note: You may see if g is odds, SMA will directly correspond to each observations from 0.5(1 + g)-th observation. If g is an even, then we have to adjust SMA; e.g. if g = 4, then first SMA(4) should come to 2.5th observation which doesn't exist. Therefore, in the case g is even, we should take average of neighboring SMAs; say, when g = 8, then it will be like as the following;
$m_1^c = 0.5(m_1 + m_2) \to y_5,\ m_2^c = 0.5(m_2 + m_3) \to y_6,\ m_3^c = 0.5(m_3 + m_4) \to y_7,\ \cdots$

Step 3: Once SMA is paired with observations, we have 2 choices. Subtract SMA from data if constant seasonality is assumed, or divide data by SMA if proportional seasonality is assumed. Then, we can obtain **seasonal index.**

Step 4: Averaging seasonal index cyclically will give **seasonality;** i.e. if g = 6, then derive 6 averages as below;
$E(SMA_1) = E(m_1 + m_7 + \cdots),\ E(SMA_2) = E(m_2 + m_8 + \cdots), \cdots, E(SMA_6) = E(m_6 + m_{12} + \cdots)$

Step 5: Finally, subtract seasonality from data if constant seasonality is assumed, or divide data by seasonality if proportional seasonality is assumed. Then, we will obtain seaonal-adjusted data.

Reference

Wikipedia. *"自己回帰移動平均モデル",* Available from

http://ja.wikipedia.org/wiki/自己回帰移動平均モデル Accessed 19 September 2010

Wikipedia. *"Correlogram",* Available from

http://en.wikipedia.org/wiki/Correlogram Accessed 19 September 2010

Wikipedia. *"Box-Jenkins",* Available from

http://en.wikipedia.org/wiki/Box-Jenkins Accessed 19 September 2010

NIST., *"6.4.4.6.3. Partial Autocorrelation Plot",* Available from

http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4463.htm Accessed 19 September 2010

NIST., *"1.3.3.27. Spectral Plot"*, Available from

http://www.itl.nist.gov/div898/handbook/eda/section3/spectrum.htm Accessed 19 September 2010

Kurita. Takio., *"自己回帰モデル"*, Available from

http://home.hiroshima-u.ac.jp/tkurita/lecture/statimage/node13.html Accessed 19 September 2010

Swift. L. and Spiezia. F., *Quantitative Methods for Business, Management, and Finance,* 4th ed. 2008.

RELATED

Ordinary Linear Regression
In "Econometrics"

Review of ordinary regression and preview of time series
In "Econometrics"

Feasible Generalized Least Squares
In "FGLS"

This entry was posted in **Econometrics** and tagged **AR model**, **ARIMA model**, **ARIMAX model**, **Autocorrelation**, **Correlogram**, **MA model**, **Partial autocorrelation**, **Partial correlogram**, **Run chart**, **SARIMAX model**, **Seasonality**, **Time series** by **Len**. Bookmark the **permalink [https://blacklen.wordpress.com/2010/09/21/arimax-model-for-stationary-time-series/]** .

2 THOUGHTS ON "ARIMAX MODEL FOR STATIONARY TIME SERIES"

Pingback: ARCH and GARCH model for volatility estimation « Logical Intuitions

Anonymous
on **2015/10/04 at 6:26 AM** said:

0 0 Rate This

Very nice description

:)