บทที่ 3
ระเบียบวิธีวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อที่จะพยากรณ์ภายใต้การวิเคราะห์หลายตัวแปร โดยจะทำการวิเคราะห์ตัวแปร ผลตอบแทน (Y), มูลค่าซื้อขายหลักทรัพย์สุทธิของนักลงทุนชาวต่างชาติในตลาดหลักทรัพย์แห่งประเทศไทย (X1), สัดส่วนการลงทุนในหลักทรัพย์ของนักลงทุนต่างชาวชาติกับการลงทุนในหลักทรัพย์ทั้งหมด (X2) และ Interaction term (X3=X1*X2) หรือก็คือปฏิกิริยาร่วมของ X1 และ X2 ร่วมกันในช่วงระยะเวลาเดียวกันเพื่อพยากรณ์ถึงผลตอบแทนของหลักทรัพย์ในตลาดหลักทรัพย์แห่งประเทศไทย และ ARFIMAX (p,d,q,X) model ได้ถูกนำมาใช้ในการพยากรณ์ผลตอบแทนหลักทรัพย์

3.1 วิธีการทดสอบ Unit root.

3.1.1 DF-Test, ADF Test (1979)

DF-Test ใช้ 3 สมการเพื่อทดสอบ unit root ใน Y_t และ Y_t คือ ข้อมูลอนุกรมเวลา

$DY_t = \alpha Y_{t-1} + U_t$ --------- (1B) [ไม่มีการสกัดกั้น]

$DY_t = \beta_t + \alpha Y_{t-1} + U_t$ --------- (2B) [มีตัวสกัดกั้น]

$DY_t = \beta_1 + \beta_t + \alpha Y_{t-1} + U_t$ --------- (3B) [การสกัดกั้น + แนวโน้ม]

โดยที่

$\alpha = (\rho - 1)$: สมมติฐานหลัก คือ $\alpha = (\rho - 1) = 0$ (ข้อมูลที่ไม่หยุดนิ่ง ($\rho = 1$))

ถ้า α > ค่าสถิติ Mackinnon สรุปได้ว่า ข้อมูลอนุกรมเวลานั้นมีการหยุดนิ่ง หรือ I(d) = I(0) มิฉะนั้น จะปฏิเสธ สมมติฐานหลัก ที่ว่า $\alpha = (\rho - 1) = 0$ หรือ [$\rho = 1$] เพราะถ้าหากว่า α มีนัยยะสำคัญทางสถิติ ในทุกๆระดับ ทำให้ $\alpha \neq 0$ ($\rho \neq 1$).

ถ้า α < ค่าสถิติ Mackinnon สรุปได้ว่า ข้อมูลอนุกรมเวลานั้นไม่หยุดนิ่ง หรือ I(d) = I(d) และยอมรับสมมติฐานหลัก ที่ว่า $\alpha = (\rho - 1) = 0$ or [$\rho = 1$] เพราะถ้าหากว่า α ไม่มีนัยสำคัญทางสถิติ ณ ระดับใดๆ ทำให้ $\alpha = 0$ ($\rho = 1$).

ADF-Test ใช้สำหรับ การทดสอบ unit root เมื่อพบว่า ปัญหาสหสัมพันธ์ของตัวรบกวนสูงขึ้น ในข้อมูลอนุกรมเวลา โดยก่อนที่จะใช้ ADF-Test, dw จะต้องถูกตรวจสอบด้วย ค่าสถิติจากสมการ DF-Test (2B) และ (3B)

$$D\,Y_t = \beta_1 + \beta_t + \alpha Y_{t-1} + ß_i \Sigma^m{}_{i=1} \Delta Y_{t-i} + \varepsilon_t \text{ ------- (4B)}$$

เมื่อแทนค่า ($ß_i \Sigma^m{}_{i=1} \Delta Y_{t-i}$) ในสมการ (4B) จากนั้น ค่า t-statistics ของ α ก่อน Y_{t-1} จะเปลี่ยนแปลง และ ค่า t-statistics จะเปลี่ยนแปลงด้วย ดังนั้น ADF-Test จะเป็นจริง สำหรับ higher order serial correlation โดยการเพิ่ม lagged differenced terms ในด้านขวา การทดสอบสมมติฐานสำหรับ unit root ในข้อมูลอนุกรมเวลา โดยใช้ วิธีการ ADF-Test สำหรับ วิธีการ DF-test และข้อสรุปเดียวกัน ของข้อมูลอนุกรมเวลา คือ หยุดนิ่ง หรือไม่หยุดนิ่ง

3.1.2 Phillips-Perron Test (PP-Test:1987,1988)

กระบวนการทดสอบ unit root นี้ พัฒนาโดย Phillips and Perron (1988) ซึ่งได้นำเสนอ วิธีการ nonparametric สำหรับ การควบคุม higher-order serial correlation ในข้อมูลอนุกรมเวลา

$$D\,Y_t = \alpha + \beta_t\, Y_{t-1} + \varepsilon_t \text{ ------- (5B)}$$

PP-test ทำให้เกิดความถูกต้องของค่า t-statistic สำหรับค่า γ coefficient ของ AR(1) regression เพื่ออธิบาย serial correlation ในสมการ(5B) การตรวจสอบจะเป็น nonparametric เมื่อใช้ในการประมาณค่า spectrum ของสมการ(5B) ที่ความถี่เท่ากับศูนย์ ซึ่งตรงกับ heteroskedasticity และ autocorrelation ของรูปแบบที่ไม่ระบุ

$$\gamma_j = (1/T)\, \Sigma^T{}_{t=j+1}\, \varepsilon^*{}_t\, \varepsilon^*{}_{t-j} \text{ --------- (6B)}$$

$$W^2 = \gamma_0 + 2 \sum^{q}_{j=1} [\ 1 - j/(q+1)\ ]\ \gamma_j \text{ ------ (7B)}$$

โดย

W^2 = ตัวประมาณค่า Newey-west heteroskedasticity autocorrelation consistent

γ_j = ค่า coefficient จาก AR(1) ในสมการ(5B)

$\varepsilon^*_t \varepsilon^*_{t-j}$ = ค่าความคลาดเคลื่อน ที่ได้จากสมการ(5B)

q = floor(4 (T/100)$^{2/9)}$, [q คือ truncation lag]

และการทำ PP-Test (t_{pp}) มีค่า t-statistic คำนวนตามสมการ (8B) เหมือนกันกับที่ t_b, s_b ซึ่งก็คือค่า t-statistics และ ความคลาดเคลื่อนมาตรฐาน (β_t) ได้จากการถดถอยในสมการ (5B) และ s* คือค่าความคลาดเคลื่อนมาตรฐาน ที่ได้จากการถดถอย ในสมการเดียวกัน

โดย

$$\text{PP-Test } (t_{pp}) = [(\gamma_0^{1/2} t_{b)} / (W)] - [(W^2 - \gamma_0)\ T\ s_b / (2\ W\ s^*)] \text{ --(8B)}$$

การแจกแจงเชิงเส้นกำกับ สำหรับ PP-Test (t_{pp}) เหมือนกันกับ ADF-Test และมีการทดสอบสมมติฐาน ดังนี้

H_0 : สมมติฐานหลัก โดย ข้อมูลอนุกรมเวลา ไม่หยุดนิ่ง

H_1 : อนุกรมเวลา หยุดนิ่ง

ถ้า PP-Test (t_{pp}) > ค่าสถิติ Mackinnon สรุปได้ว่า ข้อมูลอนุกรมเวลา มีการหยุดนิ่ง มิฉะนั้น ปฏิเสธ สมมติฐานหลัก ที่เป็นข้อมูลไม่หยุดนิ่ง

ถ้า PP-Test (t_{pp}) < ค่าสถิติ Mackinnon สรุปได้ว่า ข้อมูลอนุกรมเวลา ไม่หยุดนิ่ง พร้อมกันกับ ยอมรับสมมติฐานหลัก

3.2 Long Memory Test

เป็นการทดสอบว่าตัวแปรนั้น ๆ มี Long Memory หรือไม่ ซึ่งตัวแปรที่มี Long Memory คือตัวแปรที่จะได้รับผลกระทบได้ในระยะยาว แต่ในระยะสั้นแล้วจะไม่ได้รับผลกระทบนั้นเอง

โดยมีสมมุติฐานคือ Ho : d = 0 (ไม่มี Long Memory)

Ha : d = 0 (มี Long Memory)

3.2.1 Test for Long Memory : การทดสอบ R/S

การทดสอบ R/S test ถูกพัฒนาโดย Harold Edwin Hurst ในช่วง 1960 และ Mandelbrot & Wallis(1969) ใช้ในการคำนวณค่า พารามิเตอร์ *H*, ที่ใช้วัด ความหนาแน่นของ long range dependenceในอนุกรมเวลา

อนุกรมเวลาของช่วง T แบ่งออกเป็น *n* sub-series ของช่วง *m* และสำหรับทุกๆ sub-series โดยแต่ละ sub-series จะมีค่า m = 1,…,n, to เพื่อหาค่ามัธยฐาน (E_m) และส่วนเบี่ยงเบนมาตรฐาน (S_m) และขจัด ค่ามัธยฐานตัวอย่าง ที่ $Z_{i,m} = X_{i,m} – E_m$ สำหรับ i = 1,…..,m

หลังจากนั้นจึง สร้างอนุกรมเวลาโดยใช้รูปแบบของ $W_{i,m} = \sum_{j=1}^{i} Z_{j,m}$ โดยที่ i = 1,..,m และ เพื่อหา ระยะของ $R_m = \max\{W_{1,m,\ldots}, W_{n,m}\} - \min\{W_{1,m,\ldots}, W_{n,m}\}$

การกำหนด rescaled range ของ R_m โดยใช้ $\frac{Rm}{Sm}$ แบบเดียวกับในกรณีของอนุกรมเวลา สามารถหาค่า *R*, *S* และ *H* ตามข้อกำหนดต่อไปนี้

- กำหนดให้ R คือ ระยะ ที่อยู่ในตัวแปร , *k* คือ ค่าคงที่ และ T คือ ช่วงความยาว ของเวลา $R = k \times T^{0.5}$
- กำหนดให้ R/S คือ rescaled range, m จำนวนครั้งของกาสำรวจ, k คือ ค่าคงที่ และ *H* คือ Hurst exponent จะสามารถนำมาใช้ในอนุกรมเวลาขนาดใหญ่ได้

$$\frac{R}{S} = k \times m^H$$

- ค่า Hurst exponent หาได้จาก :

$$\log(R/S)m = \log k + H \log m$$

และมีข้อกำหนดว่า :

- ถ้า H value = 0.5 อนุกรมเวลาจะเป็นเคลื่อนไปอย่างสุ่ม และ เป็นอิสระ
- ถ้า H value =(0, 0.5) อนุกรมเวลาจะเป็นแบบไม่คงตัว กระบวนการจะคลอบคลุมเพียงแค่วงแคบ เมื่อเทียบกับกรณีการเคลื่อนไปอย่างสุ่ม
- ถ้า H value =(0.5, 1) อนุกรมเวลา จะเป็นชุดข้อมูลที่คงตัว กระบวนการจะคลอบคลุมเป็นวงกว้าง เมื่อเทียบกับกรณีการเคลื่อนไปอย่างสุ่ม

3.2.2 Test for Long Memory : การทดสอบ Modified R/S

การทดสอบ modified R/S พัฒนามาจาก การทดสอบ classical R/S ที่เสนอโดย Hurst(1951) ในขณะที่กำลังศึกษา ข้อมูลอนุกรมเวลาทางอุทกวิทยาของแม่น้ำไนล์ สำหรับชุดคำตอบ $\{ x_1, x_2, \ldots, x_T \}$ นั้น Lo (1991) ปรับปรุง ทดสอบ แบบ classical โดยการให้นิยาม (ดูสมการ (1))

$$Q_T = \hat{R} / \hat{\sigma}_T(q) \text{ -------- (1)}$$

โดยที่

$$\hat{R} = \max_{0<i\le T} \sum_{t=1}^{i} (x_i - \overline{X}) - \min_{0<i\le T} \sum_{t=1}^{i} (X_i - \overline{X}),$$

$$\hat{\sigma}_T^2(q) = \hat{\sigma}^2 + 2\sum_{j=1}^{q} w_j(q)\hat{\gamma}_j,$$

และกำหนดให้ :

$w_j(q) = 1 - | j / q |,$

$\hat{\sigma}^2$ = การเปลี่ยนแปลงของข้อมูลตามปกติ

$\overline{X}$ = ค่ามัธยฐานของข้อมูล

$\hat{\gamma}_j$ = lag – j autocovariance ของข้อมูลและช่วงที่ต้องการของ lag q หาได้จากสมการที่ 2

$$q = \text{int} \left[((3T)/2)^{1/3} ((2\hat{\rho})/1 - \hat{\rho}^2)^{2/3} \right] \text{-------- (2)}$$

โดยที่ $\hat{\rho}$ คือ sample autocorrelation coefficient อันดับแรก และ int [ ] คือ ฟังก์ชันจำนวนเต็ม ภายใต้สมมติฐานหลัก ที่ไม่มีความทรงจำระยะยาว หรือไม่มี rang dependence ระยะยาว ซึ่ง Lo (1991) เสนอว่า การกระจายที่มีขอบเขต ของค่าสถิติ Q_T ในสมการ (1) ได้มาจากฟังก์ชันการกระจาย ของความแตกต่างระหว่าง ค่าสูงสุดและต่ำสุดของ Brownian bridge บนช่วงของแต่ละหน่วย เพราะฉะนั้น เป็นการง่ายที่จะทำการทดสอบให้ได้ค่า p-value

3.2.3 Test for Long Memory : GPH Test

กระบวนการ GPH Test ถูกพัฒนาโดย Geweke, J. และ S. Porter-Hudak(1983) เพื่อแสดงถึง การประมาณค่า OLS estimator ของ d จากสมการถดถอย : (equation 3)

$$\ln[I(\xi)] = a - \hat{d}\ln[\sin^2(\frac{\xi_\lambda}{2})] + e_\lambda \quad \lambda = 1,\ldots,v \qquad \text{---- (3)}$$

โดยที่

$$I(\xi) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} e^{it\xi}(x_t - \bar{x})\right|^2 \qquad \text{--------- (4)}$$

และสมการที่ 4 คือ Periodogram (การประมาณค่าความหนาแน่น ของ spectral) ของค่า x ที่ requency (ξ) เหมือนกันกับ ค่า bandwidth v ถูกเลือกไว้สำหรับ $T \to \infty$, $v \to \infty$ แต่ $\frac{v}{T} \to 0$ แนวคิดของ Geweke and Porter-Hudak พิจารณาว่า อิทธิพลของ T จะอยู่ระหว่าง (0.5,0.6) และสมมติฐานหลักของกระบวนการความทรงจำระยะยาว ความชันของสมการถดถอย d เท่ากับศูนย์ และ ค่า t-statistics สามารถใช้ในการแสดงผลการทดสอบได้

3.3 Run ARFIMAX(p,d,q)

สามารถแสดงได้ดังนี้ (ดูสมการที่ 1.1J)

$$y_t = c_t y_{t-1} + c_2 y_{t-2} + \ldots + c_k y_{t-k} + \varepsilon_t + d_1\varepsilon_{t-1} + d_2\varepsilon_{t-2} + \ldots + d_l\varepsilon_{t-l}, \qquad \text{------ (1.1J)}$$

หรือ

$$\left(1 - \sum_{i=1}^{k} C_i L^i\right) y_t = \left(1 + \sum_{i=l}^{l} d_i L^i\right)\varepsilon_t$$

และ L คือ the lag operator { $\Sigma^{3}_{i=1}(L^{i})y_t = y_{t-1} + y_{t-2} + y_{t-3}$ } เหมือนกันกับ ARMA โดยมี ตัวแปรภายนอก หรือ ARMAX (k,l) : (ดูสมการที่ 1.2J)

$$c(L)(y_t - X_t^{/}\beta) = D(L)\varepsilon_t, \quad \text{----------- (1.2J)}$$

โดย

$$c(L) = \left(1 - \sum_{i=1}^{k} C_i L^i\right)$$

$$D(L) = \left(1 + \sum_{i=l}^{l} d_i L^i\right)\varepsilon_t$$

แบบจำลอง ARFIMAX (p, d^*,q ,X) {p=k, d^*=Fractional differencing operator, q=l} สามารถเขียนอยู่ในรูปสมการ 1.3J ดังนี้

$$c(L)(1-L)^{d^*}(y_t - X_t^{/}\beta) = D(L)\varepsilon_t; \text{------ (1.3J)}$$

โดยที่ $(1\text{-}L)^{d^*}$ คือ การทำงานของ fractional differencing และ d* $\in$ (-0.5, 0.5) คือ พารามิเตอร์ของ fractional differencing

โดย Y คือ Return ของหลักทรัพย์ PTT, PTTEP, SCC, KBANK และ CPALL

X1 คือ มูลค่าซื้อขายหลักทรัพย์สุทธิของนักลงทุนชาวต่างชาติในตลาดหลักทรัพย์แห่งประเทศไทย

X2 คือ สัดส่วนการลงทุนในหลักทรัพย์ของนักลงทุนต่างชาวชาติกับการลงทุนในหลักทรัพย์ทั้งหมด

X3 คือ ปฏิกิริยาร่วมของ X1 และ X2 (Interaction term = X1 * X2)

3.4 เลือก Best Model

- คัดเลือกแบบจำลอง

ที่มีค่าของ Autoregressive และ moving average ที่มีนัยสำคัญทางสถิติ ณ ระดับ 1%, 5% และ 10% มาพิจารณา

- พิจารณาค่า d parameter

โดยคำนวณจาก A+ (1.96 * S.E) โดย A คือค่าสัมประสิทธิ์ของ d parameter ในแบบจำลองนั้น ๆ และค่า S.E. คือ ค่าความคลาดเคลื่อนของ d parameter ในแบบจำลองนั้น ๆ หากตัวเลขที่คำนวณออกมาได้นั้นอยู่ในช่วง -0.5 ถึง 0.5 จะแสดงว่า d parameter นั้น Stationary

- พิจารณา Bayesian information criterion

พิจารณาค่าของ Bayesian information criterion (BIC) โดย BIC จะถูกใช้ในการเลือกแบบจำลองที่ดีที่สุดของ ARFIMAX (p,d,q,X) เนื่องจาก BIC เหมาะสมสำหรับการพิจารณาข้อมูลที่มีจำนวนมาก ๆ และ ในการพยากรณ์ผลตอบแทนหลักทรัพย์ แบบจำลองที่ดีที่สุดนั้น จะต้องมีค่าของ Bayesian information criterion (BIC) ของแบบจำลองน้อยที่สุด

BIC อยู่ภายใต้สมมติฐานว่าการกระจายข้อมูลที่มีลักษณะดังนี้

- x = ข้อมูล
- n = จำนวนข้อมูล x
- k = จำนวน พารามิเตอร์
- $p(x \mid k)$ = ความน่าจะเป็นของการข้อมูลที่กำหนดค่าพารามิเตอร์
- L = ค่า maximized value of the likelihood function ของการประมาณ

BIC สูตรสำหรับเป็นดังนี้

$$-2 \cdot \ln p(x|k) \approx \mathrm{BIC} = -2 \cdot \ln L + k \ln(n).$$

ภายใต้สมมติฐานของโมเดลที่ผิดพลาดหรือตัวรบกวน เป็นการกระจายแบบปกติทำให้กลายมาเป็น

$$\mathrm{BIC} = \ln(\sigma_\varepsilon^2) + \frac{k}{n}\ln(n).$$

โดยที่ σ_ε^2 เป็นความแปรปรวนข้อผิดพลาด

3.5 ดูความคลาดเคลื่อนระหว่างค่าที่แท้จริงกับค่าที่พยากรณ์ออกมาได้

3.5.1 The Mean Absolute Error (MAE)

ค่ามัธยฐานของความคลาดเคลื่อนที่เป็นค่าสัมบูรณ์ (MAE) ในทางสถิติ ค่ามัธยฐานของความคลาดเคลื่อนที่เป็นค่าสัมบูรณ์ (MAE) เป็นค่าปริมาณ ที่ใช้แสดงความใกล้เคียงของการทำนายและการพยากรณ์ ต่อผลผลิตขั้นสุดท้าย ค่ามัธยฐานของความคลาดเคลื่อนที่เป็นค่าสัมบูรณ์ (MAE) แสดงในสมการ (1X)

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| . \quad \text{-------- (1X)}$$

จากการแนะนำตามชื่อ ค่ามัธยฐานของความคลาดเคลื่อนที่เป็นค่าสัมบูรณ์ คือค่าเฉลี่ยของความคลาดเคลื่อนที่เป็นค่าสัมบูรณ์ $e_i = f_i - y_i$, โดย f_i คือค่าของตัวพยากรณ์ และ y_i คือค่าที่แท้จริง สังเกตว่า การกำหนดทางเลือก จะรวมถึง ความถี่สัมพัทธ์ เป็นค่าถ่วงน้ำหนัก ค่ามัธยฐานของความคลาดเคลื่อนที่ค่าสัมบูรณ์ เป็นตัวตรวจวัดปกติ สำหรับความคลาดเคลื่อนของการทำนายในการวิเคราะห์อนุกรมเวลา และ บทความนี้ใช้ ค่ามัธยฐานของความคลาดเคลื่อนค่าสัมบูรณ์ (MAE)ตรวจวัดความคลาดเคลื่อนของผลตอบแทนของหลักทรัพย์ โดยมีพื้นฐานจากแนวคิดวิธีการพยากรณ์แบบ ARFIMA

3.5.2 The Mean Absolute Percentage Error (MAPE)

ค่าร้อยละของมัธยฐานของความคลาดเคลื่อนที่เป็นค่าสัมบูรณ์ (MAPE) ในทางสถิติ ค่ามัธยฐานของความคลาดเคลื่อนที่เป็นค่าสัมบูรณ์ (MAE) เป็นตัววัดความแม่นยำในค่าอนุกรมเวลาที่เหมาะสมในทางสถิติ โดยเฉพาะแนวโน้ม ซึ่งโดยเฉพาะกับ การแสดงความแม่นยำเป็นร้อยละ และข้อกำหนดของ MAPE สามารถแสดงได้ตามสมการ (2X)

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \quad \text{---------- (2X)}$$

โดย A_t คือค่าที่แท้จริง และ F_t คือค่าที่ได้จากการพยากรณ์

ความแตกต่างระหว่าง A_t และ F_t จากนั้นจึงหารด้วย A_t อีกครั้ง ค่าสัมบูรณ์ของการคำนวณนี้ เป็นผลรวมของทุกๆจดที่เหมาะสม หรือจุดที่พยากรณ์ และหารอีกครั้งด้วย จุดที่เหมาะสม n ทำให้เป็นร้อยละของความคลาดเคลื่อน ที่สามารถเปรียบเทียบกับ ความคลาดเคลื่อนของอนุกรมเวลาที่เหมาะสม ที่แตกต่างกันในส่วนของระดับ และในบทความนี้ใช้ MAPE ตรวจวัดความแม่นยำใน ผลตอบแทนของหลักทรัพย์ โดยมีพื้นฐานจากแนวคิด วิธีการพยากรณ์ ARFIMA

ประโยชน์ของ ค่า MAPE คือความสามารถในการเปรียบเทียบระหว่างความแตกต่างของแบบจำลองการพยากรณ์ และมีความชัดเจนในการแปรความหมาย (Fretchling, 1996) ตัวชี้นำของการแปลความหมายMAPE’s เป็นดังนี้

- ถ้าหากค่า MAPE น้อยกว่า 10% การทำนายจะมี“ความแม่นยำสูงมาก”
- ถ้าหากค่า MAPE อยู่ระหว่าง 10%-20% การทำนายจะมี“ความแม่นยำสูง”
- ถ้าหากค่า MAPE อยู่ระหว่าง20-50% การทำนายจะมี“ความแม่นยำปานกลาง”
- ถ้าหากค่า MAPE สูงกว่า 50% การทำนายจะ “ไม่มีความแม่นยำ” (Lewis, 1982)

ดังนั้น แบบจำลองที่ดีที่สุดของ ARFIMA (p,d,q) models จะถูกใช้ในการพยากรณ์ผลตอบแทนของหลักทรัพย์ในตลาดหลักทรัพย์แห่งประเทศไทย