# Chapter 3

# Methodology

This study only focus on econometric models and the theoretical model this study will not explain more. The methodology employed in this study has three steps: First, adjust the each stock index data in form of the rate of return and use ADF and PP test to test is there any unit root in our data. Second，use linear and nonlinear models fit for China's Real Estate Sector Stock and Shenzhen Index respectively，use tsDyn package in R. Based on AIC, BIC and MAPE this study select the appropriate models. Last, according to the appropriate model, analysis the dependence relationship between returns in percentage of Real Estate Sector Stock and Shenzhen Index based on empirical copula approach. The details of the econometric models are as follows.

## 3.1 Data processing and testing

### 3.1.1 Rate of Return

Adjust information of China's real estate sector stock and Shenzhen Index in form of return terms follow as:

$$X_R = \ln(\frac{P_{r,t}}{P_{r,t-1}}) \tag{3.1}$$

$$X_S = \ln(\frac{P_{s,t}}{P_{s,t-1}}) \tag{3.2}$$

Where, $X_R$ is the return on China's real estate sector stock, $X_S$ is the return on Shenzhen Index. $P_{r,t}$ is the China's real estate sector stock at time t, $P_{r,t-1}$ is the China's real estate sector stock at time t-1, $P_{s,t}$ is the Shenzhen Index at time t, $P_{s,t-1}$ is the Shenzhen Index at time t-1.

### 3.1.2 Unit root test by using ADF test and PP test

1. ADF test, test the time series data for stationary follow as:

$$\Delta G_{R,t} = \alpha_1 + \beta_1 t + \delta_1 G_{R,t-1} + \sum_{i=1}^{m} s_i \Delta G_{R,t-1} + \varepsilon_{1t} \tag{3.3}$$

$$\Delta G_{S,t} = \alpha_2 + \beta_2 t + \delta_2 G_{S,t-1} + \sum_{i=1}^{m} s_i \Delta G_{S,t-1} + \varepsilon_{2t} \tag{3.4}$$

Where, $G_{R,t}$ is the China's real estate sector stock at time t, $G_{R,t-1}$ is the China's real estate sector stock at time t-1, $G_{S,t}$ is the Shenzhen Index at time t, $G_{S,t-1}$ is the Shenzhen Index at time t-1. $\alpha_1, \beta_1, \delta_1, s\ \alpha_2, \beta_2, \delta_2$ are the parameters, $\varepsilon_{1t}\ \varepsilon_{2t}$ are the error term, t is the trend.

The hypotheses for test are following as:

$H_0$: $\delta_i = 0$ (non-stationary)

$H_1$: $\delta_i < 0$ (stationary) when i=1, 2

If accept $H_0$ means these two stock indexes have unit root and non-stationary but if accept $H_1$ means that these two stock indexes have no unit root and stationary.

2. PP test, this test was developed by Phillips and Perron (1988). The model as follows:

$$\tilde{t} = t_p \left(\frac{w_0}{B}\right)^{1/2} - \frac{T(B - w_0)(s_p(\hat{\varphi}))}{2B^{1/2}e} \tag{3.5}$$

Where $\tilde{t}$ the ratio of p, B is is the residual of estimate, $w_0$ is the consistent estimate of error variance, e is the standard error from the regression test.

The hypotheses for test same as ADF test.

**3.2 Linear and Nonlinear method models and selection**

In the time-series literature reviews, despite linear methods often work well. It also provides a useful standard as a basis for analyses. However, there is no promising why real observations should all be linear. Moreover, stock market is a kind of complex nonlinear dynamic system; linear model could not accurate show stock index fluctuation. So using nonlinear models seems to be reasonable choice. And this study know linear model is one of the special nonlinear models.

**(1) Autoregressive-linear model (AR-linear Model)**

The basic linear model is AR model can be written in equation:

$$y_{rt+s} = \phi + \phi_0 y_{rt} + \phi_1 y_{1-d} + \ldots + \phi_m y_{rt-(m-1)d} + \varepsilon_{rt+s} \tag{3.6}$$

$$y_{st+s} = \phi + \phi_0 y_{st} + \phi_1 y_{1-d} + \ldots + \phi_m y_{st-(m-1)d} + \varepsilon_{st+s} \tag{3.7}$$

Where $y_{rt}$ is real estate sector stock data at time t, $\phi$ is parameter and coefficient of $y_{rt}$ in the model. In addition, $\varepsilon$ is error term of this equation.

Where $y_{st}$ is Shenzhen index data at time t, $\phi$ is parameter and coefficient of $y_{st}$ in the model. In addition, $\varepsilon$ is error term of this equation.

**(2) Self-Exciting Threshold Autoregressive Model (SETAR Model)**

The general Self-Exciting Threshold Autoregressive Model can be written in equation:

$$y_{rt+s} = \begin{cases} \phi_1 + \phi_{10} y_{rt} + \phi_{11} y_{rt-d} + \ldots + \phi_{1L} y_{rt-(L-1)d} + \varepsilon_{rt+s} & Z_{rt} \le th \\ \phi_2 + \phi_{20} y_{rt} + \phi_{21} y_{rt-d} + \ldots + \phi_{2H} y_{rt-(H-1)d} + \varepsilon_{rt+s} & Z_{rt} > th \end{cases} \tag{3.8}$$

$$y_{st+s} = \begin{cases} \phi_1 + \phi_{10} y_{st} + \phi_{11} y_{st-d} + \ldots + \phi_{1L} y_{st-(L-1)d} + \varepsilon_{st+s} & Z_{st} \le th \\ \phi_2 + \phi_{20} y_{st} + \phi_{21} y_{st-d} + \ldots + \phi_{2H} y_{st-(H-1)d} + \varepsilon_{st+s} & Z_{st} > th \end{cases} \tag{3.9}$$

Where $y_{rt}$ is real estate sector stock data at time t, $\phi$ is the parameter and coefficient of equation. In addition, $\varepsilon$ is error term of this equation and $Z_{rt}$ is a threshold variable in the model. The “L” is represented lower regime of model and “H” is represented the higher regime of the model.

Where $y_{st}$ is Shenzhen index data at time t, $\phi$ is the parameter and coefficient of equation. In addition, $\varepsilon$ is error term of this equation and $Z_{st}$ is a threshold variable in the model. The “L” is represented lower regime of model and “H” is represented the higher regime of the model.

**(3) Logistic Smooth Transition Autoregressive Model (LSTAR model)**

The general Logistic Smooth Transition Autoregressive Model model can be written in equation:

$$\begin{aligned} y_{rt+s} &= (\phi_1 + \phi_{10} y_{rt} + \phi_{11} y_{rt-d} + \ldots + \phi_{1L} y_{rt-(L-1)d})(1 - G(z_{rt}, \gamma, th)) + \\ &(\phi_2 + \phi_{20} y_{rt} + \phi_{21} y_{rt-d} + \ldots + \phi_{2H} y_{rt-(H-1)d}) G(z_{rt}, \gamma, th) + \varepsilon_{rt+s} \end{aligned} \quad (3.10)$$

$$\begin{aligned} y_{st+s} &= (\phi_1 + \phi_{10} y_{st} + \phi_{11} y_{t-sd} + \ldots + \phi_{1L} y_{st-(L-1)d})(1 - G(z_{st}, \gamma, th)) + \\ &(\phi_2 + \phi_{20} y_{st} + \phi_{21} y_{st-d} + \ldots + \phi_{2H} y_{st-(H-1)d}) G(z_{st}, \gamma, th) + \varepsilon_{st+s} \end{aligned} \quad (3.11)$$

Where $y_{rt}$ is the real estate sector stock data at time t, $\phi$ is the parameter and coefficient of equation. In addition, $\varepsilon$ is error term of this equation and $Z_{rt}$ is a threshold variable in the model. The “L” is represented lower regime of model and “H” is represented the higher regime of the model. Moreover, “G” is the logistic function and $\phi, \gamma$ , th are the parameters to be computed.

Where $y_{st}$ is the Shenzhen index data at time t, $\phi$ is the parameter and coefficient of equation. In addition, $\varepsilon$ is error term of this equation and $Z_{st}$ is a threshold variable in the model. The “L” is represented lower regime of model and “H” is represented the higher regime of the model. Moreover, “G” is the logistic function and $\phi, \gamma$ , th are the parameters to be computed.

**(4) Neural Network Models (NNT Model)**

From the Neural Network Model can be written in equation:

$$y_{rt+s} = \beta_0 + \sum_{j=1}^{D} \beta_j g(\gamma_{0j} + \sum_{i=1}^{m} \gamma_{ij} y_{rt-(i-1)d}) \tag{3.12}$$

$$y_{st+s} = \beta_0 + \sum_{j=1}^{D} \beta_j g(\gamma_{0j} + \sum_{i=1}^{m} \gamma_{ij} y_{st-(i-1)d}) \tag{3.13}$$

Where $y_{rt}$ is real estate sector stock data at time t, the $\beta_0$ is parameter of equation. In a hidden units and activation function g.

Where $y_{st}$ is Shenzhen index data at time t, the $\beta_0$ is parameter of equation. In a hidden units and activation function g .

**(5) Additive Autoregressive Model (AAR Model)**

The generalized non-parametric additive model (Generalized Additive Model) can be written in equation:

$$y_{rt+s} = \mu + \sum_{i=1}^{m} s_i(y_{rt-(i-1)d}) \tag{3.14}$$

$$y_{st+s} = \mu + \sum_{i=1}^{m} s_i(y_{st-(i-1)d}) \tag{3.15}$$

Where $y_{rt}$ is real estate sector stock data at time t. $S_i$ are smooth functions represented by penalized cubic regression.

Where $y_{st}$ is Shenzhen index data at time t. $S_i$ are smooth functions represented by penalized cubic regression.

**(6) Information Criteria: Akaike Information Criteria (AIC), Schwartz Information Criteria (SIC or BIC)** [1]

The Akaike (1974, 1976) and Schwarz (1978) information criteria for selecting the most parsimonious correct model are respectively.

Akaike: $$c_n(h) = \frac{-2 \cdot \ln(L_n(h))}{n} + \frac{2h}{n} \quad (3.16)$$

Schwarz: $$c_n(h) = \frac{-2 \cdot \ln(L_n(h))}{n} + \frac{h \cdot \ln(n)}{n} \quad (3.17)$$

Where $L_n(h)$ is the maximized value of the likelihood function for the estimated model, h is the number of parameters used, n is the sample size.

The value of both Akaike Information Criteria (AIC), Schwartz Information Criteria (SIC or BIC) using to select the appropriate nonlinear models for assessing the Real Estate Sector Stock and Shenzhen Index, respectively. The model has the lowest value of AIC and SIC that the best model.

**(7) The Mean Absolute Percentage Error (MAPE)**

The formula of MAPE follows as:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{S_t - R_t}{S_t}\right| \quad (3.18)$$

Where $S_t$ is the actual value, $R_t$ is the forecast value, n is the data of forecast.

If the MAPE value is less than 10%, it is highly accurate forecast. If the MAPE value is between 10% to 20%, it is good forecast. If the MAPE value is between 20% to 50%, it is reasonable forecast. If the MAPE value is greater than 50%, it is inaccurate forecast.

## 3.3 Copulas Theory

Copula theory not only captures nonlinear asymmetric and tail dependence, but also calculates easily the joint distribution of financial time series. Copula technique can be also used widely in studying the characteristics of financial markets，portfolio aggregation and risk analysis etc. The concept applies to high dimensional processes, but for simplicity this study focus on the bivariate case in this paper. Therefore we will base on empirical copula theory to analysis dependence measures between these two stock Indexes in People’s Republic of China.

### 3.3.1 The empirical copula

Empirical copula theory as introduced by Deheuvels [1979]. The advantage of empirical copula based on real data distribution, and it can accurate analysis the distribution information. So this theory is more reliable.

Suppose that the marginal distribution F is continuous, therefore the copula associated to F is unique.

Copula explained as follows:

$$C = \{(\frac{t_1}{T},...,\frac{t_N}{T});1 \le n \le N, t_n = 0,...,T\} \tag{3.19}$$

Through

$$\hat{C}(\frac{t_1}{T},...,\frac{t_N}{T}) = \frac{1}{T}\sum_{t=1}^{T}\prod_{n=1}^{N} 1_{[\gamma_n^t \le t_n]} \tag{3.20}$$

We will get the empirical copula.

Suppose $\{x_k, y_k\}_{k=1}^{n}$ stand for the bivariate time series observation distribution and length of the observation is n, the Empirical copula $C_n$ explained as follows:

$$C_n(\frac{i}{n},\frac{j}{n}) = \{\sum \text{ The sample which satisfied } x \le x_i \cap y \le y_j\}/\text{n} \tag{3.21}$$

### 3.3.2 Rank correlation

In order to capture the overall dependence measures between return in percentage of these two stock index, I use a nonlinear correlation method including Kendall.tau and Spearman.rho.

(1) Kendall.tau

Suppose $(x_1, y_1),(x_2, y_2)$ are i.i.d vector, $x_1, x_2 \in x, y_1, y_2 \in y$.

τ is between [-1, 1], suppose the copula function of $(x_1,y_1)$ is C (u, v), that the function τ explained as follows:

$$\tau = 4\iint_{0}^{1} C(u,v)dC(u,v) - 1 \tag{3.22}$$

(2) Spearmam.rho

H(x, y) is the joint distribution of (x, y).

The copula function C (u, v) is given, where u=F(x), v=G (y), that the function ρ explained as follows:

$$\rho = 12\iint_{0}^{1} C(u,v)dC(u,v) - 3 \tag{3.33}$$

u and v stand for uniform distribution of People’s Republic of China’s exchange rates and Thailand’s exchange rates.

**3.3.3 Pearson's linear correlation coefficient**

In order to compare with the nonlinear correlation，the Pearson’s linear correlation will compute.

Two variables Pearson's linear correlation can be written in equation：

$$\rho_{u,v} = \frac{E(UV) - E(U)E(V)}{\sqrt{E(U^2) - (E(U))^2}\sqrt{E(V^2) - (E(V)^2}} \tag{3.34}$$

The sample Pearson correlation coefficient can be written in equation：

$$\gamma_{uv} = \frac{\sum u_i v_i - \bar{u}\bar{v}}{(n-1)s_x s} = \frac{n\sum u_i v_i - \sum u_i \sum v_i}{\sqrt{n\sum u_i^2 - (\sum u_i)^2}\sqrt{n\sum v_i^2 - (\sum v_i)^2}} \tag{3.35}$$

Depending on the numbers observations involved in, sometimes the correlation will be unstable.

**3.3.4 Dependence Measures**

There are 4 properties explain the dependence measures shown below (Embrechts, Lindskog and McNeil (2003)):

1. $\delta(U,V) = \delta(U,V)$

2. $-1 \le \delta(U,V) \le 1$

3. $\delta(U,V) = 1$ If X and Y are comonotonic; as well as $\delta(U,V) = -1$ if U and V are comontonic.

4. If T is exactly monotonic, then $\delta(T(U),V) = \{\delta(U,V), T = \mathit{increasing}$ or $-\delta(U,V), T = \mathit{decreasing}\}$

Normally, the Pearson linear correlation only satisfied the first two properties but the rank correlation measures Spearman's rho and Kendall's tau which are nonlinear correlation fit all of the properties we showed. Therefore, the copulas use the Spearman's rho and Kendall's tau calculating the dependence measures between U and V which are random variables.