TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

─────── \* ───────

ĐỒ ÁN

**TỐT NGHIỆP ĐẠI HỌC**

NGÀNH KHOA HỌC MÁY TÍNH

**NGHIÊN CỨU MÔ HÌNH ARIMA DỰ BÁO CHUỖI THỜI GIAN**

Sinh viên thực hiện : **Dương Văn Chỉnh**

Lớp KHMT - K55

Giáo viên hướng dẫn: **TS. Đinh Viết Sang**

HÀ NỘI 5-2015

**PHIẾU GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP**

1. Thông tin về sinh viên

Họ và tên sinh viên: Dương Văn Chỉnh

Điện thoại liên lạc: 01644357605 Email: ctk41sphn@gmail.com

Lớp: KHMT K55 Hệ đào tạo: Đại học chính quy

Đồ án tốt nghiệp được thực hiện tại:

Phòng thí nghiệm Hệ thống máy tính, viện Công nghệ thông tin và Truyền thông, trường đại học Bách khoa Hà Nội.

Thời gian làm ĐATN: Từ ngày 06/02/2014 đến 30/05/2014.

2. Mục đích nội dung của ĐATN

Nghiên cứu mô hình ARIMA xử lý dữ liệu chuỗi thời gian

3. Các nhiệm vụ cụ thể của ĐATN

* Tìm hiểu chuỗi thời gian và mô hình ARIMA
* Áp dụng mô hình ARIMA vào bài toán dự báo giá cổ phiếu vndex

4. Lời cam đoan của sinh viên:

Tôi – *Dương Văn Chỉnh* - cam kết ĐATN là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của *TS. Đinh Viết Sang*.

Các kết quả nêu trong ĐATN là trung thực, không phải là sao chép toàn văn của bất kỳ công trình nào khác.

|  |  |
| --- | --- |
|  | *Hà Nội, ngày 30 tháng 05 năm 2015*  Tác giả ĐATN  *Dương Văn Chỉnh* |

5. Xác nhận của giáo viên hướng dẫn về mức độ hoàn thành của ĐATN và cho phép bảo vệ:

|  |  |
| --- | --- |
|  | *Hà Nội, ngày 30 tháng 05 năm 2015*  Giáo viên hướng dẫn  *TS. Đinh Viết Sang* |

**Chương I: Giới thiệu**

* 1. **Giới thiệu chung**

Ngày nay khi xã hội ngày càng phát triển thì lượng thông tin càng tăng lên với tốc dộ bùng nổ. Lượng dữ liệu khổng lồ ấy là một nguồn tài nguyên vô giá nếu như chúng ta biết cách phát hiện và khai thác những thông tin hữu ích có trong đó. Như vậy vấn đề đặt ra với dữ liệu của chúng ta là việc lưu trữ và khai thác chúng. Các phương pháp khai thác dữ liệu truyền thống ngày càng không đáp ứng được nhu cầu thực tế. Một khuynh hướng kỹ thuật mới ra đời đó là “Kỹ thuật khai phá dữ liệu và khám phá tri thức”(Knownledge Discovery and Data mining – KDD)

Công nghệ khai phá dữ liệu ra đời đã cho phép chùng ta khai thác những tri thức hữu dụng bằng việc trích xuất những thông tin trong mối quan hệ hoặc mối tương quan nhất định ncó mối quan hệ hoặc mối tương quan nhất định từ một kho dữ liệu lớn (cực lớn) mà bình thường không thể nhận diện được từ đó giải quyết các bài toán tìm kiếm, dự báo các xu thế, các hành vi trong tương lai, và nhiều tính năng thông minh khác. Ngày nay, các công nghệ data mining được ứng dụng rộng rãi trong hầu hết các lĩnh vực: phân tích dữ liệu, dự báo, …

Dự báo là một nhu cầu không thể thiếu cho những hoạt động của con người trong bối cảnh bùng nổ thông tin. Dự báo sẽ cung cấp những cơ sở cần thiết cho các hoạch định, và có thể nói rằng nếu không có khoa học dự báo thì những dự định tương lai của con người vạch ra sẽ không có sự thuyết phục đáng kể.

Trong công tác phân tích dự báo, vấn đề quan trọng hàng đầu cần đặt ra là việc năm bắt tối đa thông tin về lĩnh vực dự báo. Thông tin ở đây có thể hiểu một cách cụ thể gồm : (1) các số liệu quá khứ của lĩnh vực dự báo, (2) diễn biến tình hình hiện trạng cũng như động thái phát triển của lĩnh vực dự báo và (3) đánh giá một cách đầy đủ nhất các nhân tố ảnh hưởng cả về định lượng lẫn định tính

Căn cứ vào nội dung phương pháp và mục đích của dự báo, người ta chia dự báo thành hai loại: Phương pháp định tính và phương pháp định lượng. Phương pháp định tính thường phụ thuộc rất nhiều vào kinh nghiệm của một hay nhiều chuyên gia trong lĩnh vực liên quan. Phương pháp này thường được áp dụng, kết quả dự báo sẽ được các chuyên gian trong lĩnh vực liên quan nhận xét, đánh giá và đưa ra kết luận cuối. Phương pháp định lượng sử dụng những dữ liệu quá khứ theo thời gian, dựa trên dữ liệu lịch sử để phát hiện chiều hướng vận động của đối tượng phù hợp với một mô hình toán học nào đó và đồng thời sử dụng mô hình đó làm mô hình ước lượng. Tiếp cận định lượng dựa trên giả định rằng giá trị tương lai của biến số dự báo sẽ phụ thuộc vào xu thế vận động của đối tượng đó trong quá khứ. Phương pháp dự báo theo chuỗi thời gian là một phương pháp định lượng. Phương pháp chuỗi thời gian sẽ dựa trên việc phân tích chuỗi quan sát của một biến duy nhất theo biến số độc lập là thời gian. Giả định chủ yếu là biến số dự báo sẽ giữ nguyên chiều hướng phát triển đã xảy ra trong quá khứ và hiện tại

* 1. **Bài toán dự báo tài chính**

Một trong những vấn đề quan trọng nhất trong lĩnh vực tài chính hiện đại là tìm kiếm những cách thức hiệu quả để tóm tắt và hình dung dữ liệu thị trường chứng khoán để cung cấp cho các cá nhân hoặc tổ chức những thông tin hữu ích về các hành vi thị trường hỗ trợ việc ra các quyết định đầu tư. Số lượng lớn dữ liệu có giá trị được tạo ra bởi thị trường chứng khoán đã thu hút được các nhà nghiên cứu khám phá vấn đề này bằng cách sử dụng các phương pháp khác nhau.

Đối với Việt Nam, thị trường chứng khoán còn khá mới mẻ, song ai cũng biết được tiềm năng và lợi ích đáng kể của nó. Việc khai thác được thị trường này sẽ đem lại lợi ích kinh tế cao. Dự báo thị trường chứng khoán là một công việc khá quan trọng để khai thác lĩnh vực này. Chính vì vậy tôi đã chọn đề tài “***Nghiên cứu mô hình ARIMA dự báo chuỗi thời gian***” để làm luận văn tốt nghiệp với mục đích hiểu được công nghệ 3 data mining cũng như ứng dụng to lớn của nó trong việc dự báo, dự đoán xu hướng trong tương lai, đặc biệt là trong lĩnh vực thị trường tài chính, chứng khoán từ đó có những quyết định đầu tư, giao dịch phù hợp.

* 1. **Mục tiêu của đồ án và phương pháp thực hiện**
* **Mục tiêu:**
* Tìm hiểu kỹ thuật phân tích dữ liệu chuỗi thời gian trong khai phá dữ liệu áp dụng vào bài toán dự báo nói chung và dự báo trong thị trường chứng khoán nói riêng
* Tìm hiểu mô hình ARIMA (**Auto Regressive Integrate Moving Average**) với chức năng nhận dạng mô hình, ước lượng các tham số và đưa ra kết quả dự báo dựa trên các tham số ước lượng đã được lựa chọn một cách tối ưu. Thực nghiệm mô hình ARIMA trên dữ liệu thời gian thực, áp dụng với dữ liệu chứng khoán hướng tới việc dự báo chứng khoán.
* **Phương pháp thực hiện:**
* Tìm hiểu chuỗi thời gian, các đặc điểm của chuỗi thời gian
* Tìm hiểu cơ sở lý thuyết về mô hình ARIMA cho dữ liệu thời gian thực(time series) và cách áp dụng vào bài toán thực tế - dự báo sự lên xuống của thị trường chứng khoán
* Xây dựng và thi hành mô hình ARIMA và ứng dụng vào bài toán khai phá dữ liệu chuỗi thời gian trong dự báo tài chính, chứng khoán
* Tìm hiểu và phân tích dữ liệu chứng khoán đang có được
* Sử dụng ngôn ngữ thống kê **R** để làm phần mềm đơn giản đánh giá và dự báo dữ liệu giá cổ phiếu
* Đánh giá kết quả dự báo được

**Chương II: Dữ liệu chuỗi thời gian**

1. **. Khái niệm chuỗi thời gian**

Chuỗi thời gian(tiếng Anh: time series) trong thống kê, xử lý tín hiệu, kinh tế lượng và bài toán tài chính là một chuỗi các điểm dữ liệu, được đo theo từng khoành khắc thời gian liền nhau theo một tần suất thời gian thống nhất.

Phân tích chuỗi thời gian bao gồm các phương pháp để phân tích dữ liệu chuỗi thời gian, để từ đó trích xuất ra được các thuộc tính thống kê có ý nghĩa và các đặc điểm của dữ liệu. Dự đoán chuỗi thời gian là việc sử dụng mô hình để dự đoán các sự kiện thời gian dựa vào các sự kiện đã biết trong quá khứ để từ đó dự đoán các điểm dữ liệu trước khi nó xảy ra (hoặc được đo). Chuỗi thời gian thường được vẽ theo các đồ thị

1. **. Chuỗi thời gian thực**

Dữ liệu thời gian thực hay chuỗi thời gian là một chuỗi các giá trị của một đại lượng nào đó được ghi nhận là thời gian

Các giá trị của chuỗi thời gian của đại lượng X được kí hiều là X1, X2,…, Xt,…Xn với Xi là giá trị của đại lượng X tại thời điểm i

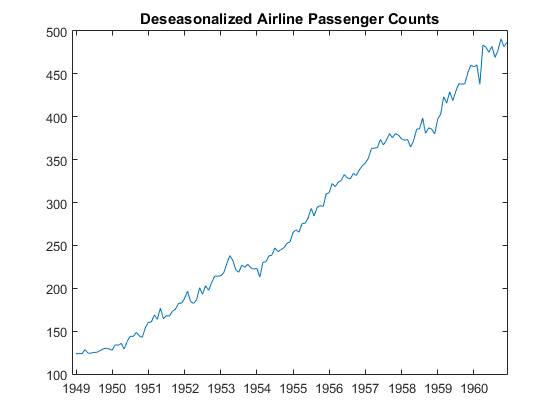
**Các thành phần của chuỗi thời gian thực**

Chuỗi thời gian thực thường được chia ra làm 4 thành phần:

* Thành phần xu hướng dài hạn
* Thành phần mùa
* Thành phần chu kì
* Thành phần bất thường

**1.2.1. Thành phần xu hướng dài hạn**

Thành phần xu hướng dài hạn thể hiện sự tăng trưởng hoặc giảm sút của một biến số theo thời gian với khoảng thời gian đủ dài. Thể hiện trên đồ thị thành phần này có thể biểu diễn bởi một đường thẳng hay một đường cong trơn.

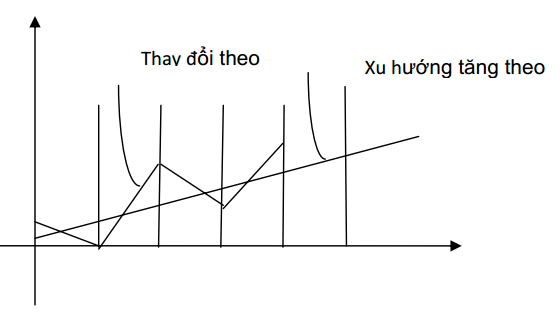


**1.2.2. Thành phần mùa**

Thành phần này dùng để chỉ xu hướng tăng hay giảm của đại lượng X tính theo mùa trong năm(có tính theo ngày/tháng trong năm)

Ví dụ: lượng tiêu thụ chất đốt sẽ tăng vào mùa đông và giảm vào mùa hè, ngược lại, lượng tiêu thụ xăng sẽ tăng vào mùa hè và giảm vào mùa đông.

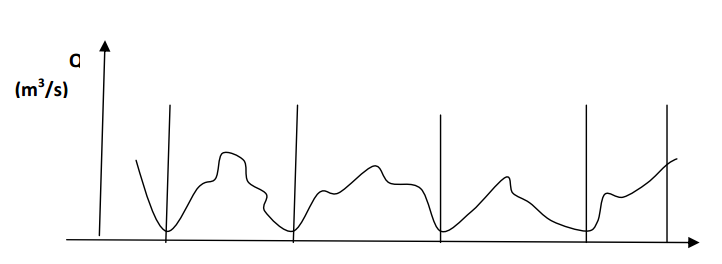
Lượng mưa sẽ tăng cao vào mùa mưa



**1.2.3.Thành phần chu kì**

Thành phần này chỉ sự thay đổi của đại lượng X theo chu kì. Thành phần này khác thành phần mùa ở chỗ chu kì của đại lượng X kéo dài hơn một năm. Để đánh giá thành phần này các giá trị của chuỗi thời gian được quan sát hàng năm

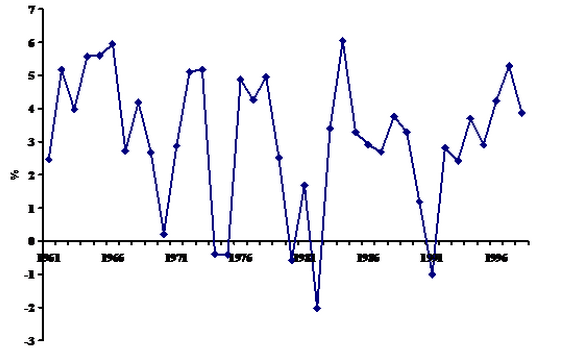
Ví dụ: lượng nước chảy đến hồ Trị An từ năm 1959-1983



**1.2.4.Thành phần bất thường**

Thành phần này dùng để chị sự thay đổi bất thường của các giá trị trong chuỗi thời gian. Sự thay đổi này không thể dự đoán bằng các số liệu kinh nghiệm trong quá khứ, về mặt bản chất thành phần này không có chu kì.

Ví dụ: chu kì ngẫu nhiên tăng trưởng kinh tế Hoa Kì giai đoạn 1961-1999



**Chương II: Cơ sở xác suất thống kê**

**1. Biến ngẫu nhiên**

1.1.Khái niệm biến ngẫu nhiên

Biến ngẫu nhiên X là đại lượng nhận các giá trị nào đó phụ thuộc vào các yếu tố ngẫu nhiên, nghĩa là với mọi giá trị thực thì là một biến cố ngẫu nhiên.

Như vậy đối với biến ngẫu nhiên người ta chỉ quan tâm xem nó nhận một giá trị nào đó hoặc nhận giá trị trong một khoảng thời gian nào đó với một xác suất là bao nhiêu

Ví dụ: các đại lượng sau là biến ngẫu nhiên:

* Số nốt xuất hiện khi gieo một con xúc xắc
* Tuổi thọ của một thiết bị đang hoạt động
* Số khách hàng vào một điểm phục vụ trong một đơn vị thời gian
* Số cuộc gọi đến một tổng đài
* Sai số khi đo lường một đại lượng vật lý

1.2. Phân loại

Người ta phân biến ngẫu nhiên thành hai loại:

* Biến ngẫu nhiên rời rạc: nếu nó chỉ nhận một số hữu hạn hoặc vô hạn đếm được các giá trị. Nghĩa là có thể liệt kê các giá trị thành một dãy
* Biến ngẫu nhiên liên tục nếu các giá trị của nó có thể lấp đầy một hoặc một số các khoảng hữu hạn hoặc vô hạn và xác suất bằng không với mội a

Ví dụ:

* Gọi X là số nốt xuất hiện khi gieo một con xúc sắc thì X là biến ngẫu nhiên rời rạc nhận các giá trị 1, 2, 3, 4, 5, 6.
* Gọi Y là tuổi thọ của một thiết bị đang hoạt động thì Y là biến ngẫu nhiên liên tục nhận giá trị trong một khoảng
* Gọi Z là số khách hàng vào một điểm phục vụ trong một đơn vị thời gian, Z là biến ngẫu nhiên rời rạc nhận các giá trị 0, 1, 2…
* Số cuộc gọi đến một tổng đài là biến ngẫu nhiên rời rạc nhận các giá trị 0, 1, 2,….
* Sai số khi đo lường một đại lượng vật lý Y nào đó là biến ngẫu nhiên liên tục nhận giá trị trong một khoảng

1. **Hàm phân bố xác suất**

Hàm phân bố xác suất (cumulative distribution function, viết tắt là CDF) của biến ngẫu nhiên X là hàm số F(x) được xác định với bởi công thức:

*Hàm phân bố xác suất có các tính chất sau:*

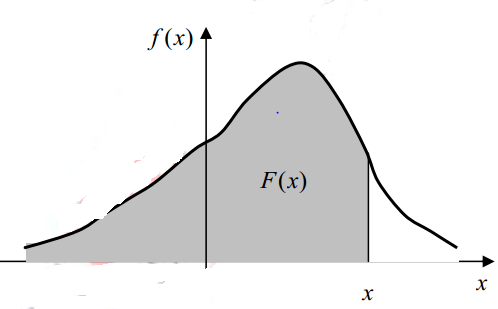
* với mọi .
* là hàm không giảm, liên tục bên trái. Nếu X là biến ngẫu nhiên liên tục thì là hàm liên tục.

1. **Biến ngẫu liên tục**

Giả sử X là một biến ngẫu nhiên liên tục có hàm phân bố F(x). Nếu tồn tại hàm f(x) sao cho với mọi

thì f(x) gọi là hàm mật độ của biến ngẫu nhiên X(probability density function, viết tắt là PDF)

Như vậy giá trị của hàm F(x) bằng diện tích hình phẳng giới hạn bởi đồ thị hàm mật độ f(x), trục hoành và đường thẳng song song với trục tung có hoành độ là x.



**Các tính chất của hàm mật độ:**

* tại các điểm x mà f(x) liên tục
* với mọi

**Ví dụ:**

Hàm phân bố xác suất của biến ngẫu nhiên liên tục X có dạng:

1. Hãy các định hệ số k
2. Tìm hàm mật độ xác suất f(x).

Giải:

1. Vì hàm phân bố xác suất F(x) liên tục, do đó tại x = 1
2. Theo tính chất trên của hàm mật độ xác suất ta có:

* với
* với
  1. **Phân phối chuẩn**

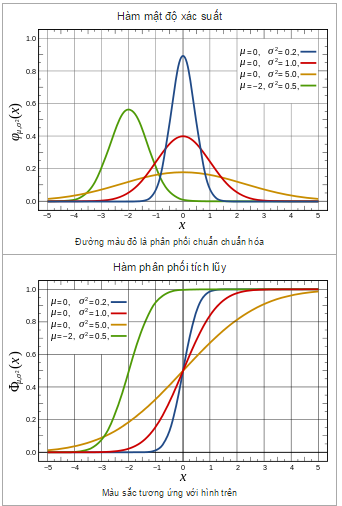
**Định nghĩa:** Biến ngẫu nhiên liên tục X có phân bố chuẩn , kí hiệu , nếu hàm mật độ có dạng:

trong đó là giá trị trung bình và là phương sai

Phân bố chuẩn được Gauss tìm ra năm 1809 nên nó còn được gọi là phân bố Gauss. Phân bố chuẩn thường được thấy trong các bài toán về sai số gặp phải khi đo đạc các đại lượng vật lý, thiên văn,..

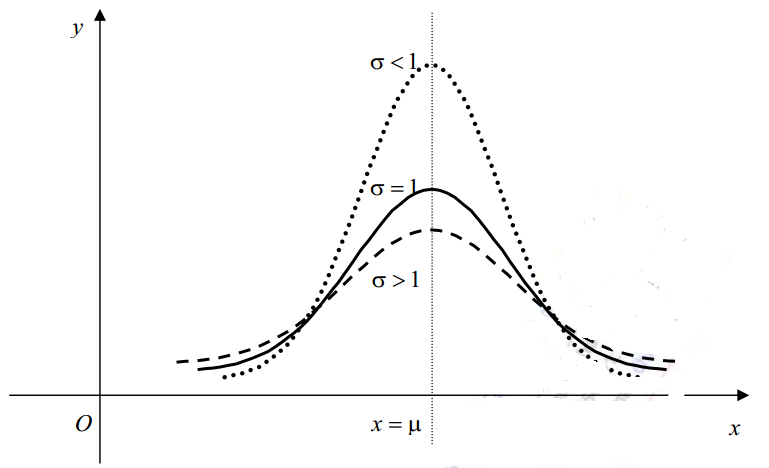
Trong thực tế, nhiều biến ngẫu nhiên tuân theo quy luật chuẩn hoặc tiệm cận chuẩn(Định lý giới hạn trung tâm). Chẳng hạn: trọng lượng, chiều cao của một nhóm người nào đó, điểm thi của thí sinh, năng suất cây trồng, mức lãi suất của một công ty, nhu cầu tiêu thụ của một mặt hàng nào đó.

**Phân phối chuẩn hóa**(standard normal distribution) là phân phối chuẩn với giá trị trung bình bằng 0 và phương sai bằng 1(đường cong màu đỏ trong hình bên dưới). Phân phối chuẩn còn được gọi là đường cong chuông(bell curve) vì đồ thị của hàm mật độ xác suất có dạng chuông.



**Tính chất đồ thị của hàm mật độ phân phối chuẩn:**

* Nhận trục làm trục đối xứng.
* Tiệm cận với trục hoành khi .
* Diện tích giới hạn với đồ thị và trục hoành bằng 1.
* Đạt cực đại tại và có giá trị cực đại bằng Có 2 điểm uốn tại
* Do đó khi tăng lên thì đồ thị dịch sang phải, còn khi giảm đồ thị dịch sang trái
* Khi tăng lên thì đồ thị sẽ thấp xuống, còn khi giảm đồ thị cao lên và nhọn hơn.



Nếu là hai biến ngẫu nhiên độc lập có phân bố chuẩn và ) thì tổ hợp tuyến tính bất kì của cũng có phân bố chuẩn, đặc biệt

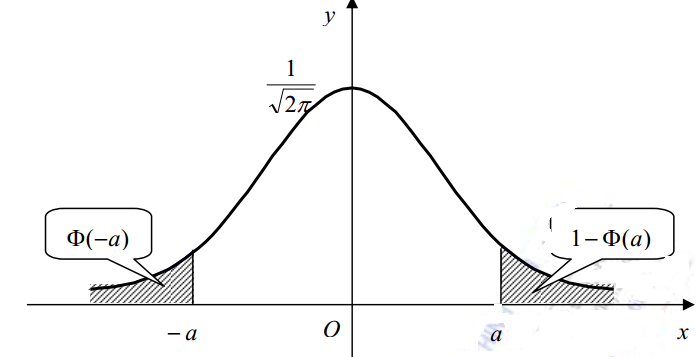
**3.1.2. Phân bố chuẩn tắc**

Phân bố chuẩn N(0, 1) với kì vọng bằng 0, phương sai bằng 1 gọi là phân bố chuẩn tắc. Hàm mật độ của

với mọi

Hàm phân bố của N(0; 1)

Đồ thị của hàm mật độ



**Các tính chất của hàm phân bố**

* Nếu thì

Gọi giá trị gọi là giá trị tới hạn của phân bố chuẩn tắc mức a nếu

Nếu thì

Người ta chứng minh được:

Nếu thì ~ N(0,1)

Từ đó ta có:

-

Xác suất của sự sai lệch giữa biến ngẫu nhiên có quy luật chuẩn và kì vọng của nó được tính theo công thức:

**Ví dụ:** Giả sử . Hãy tìm

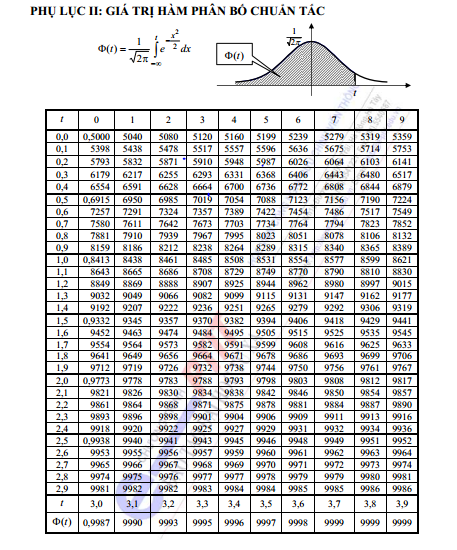
1. Xác định a để

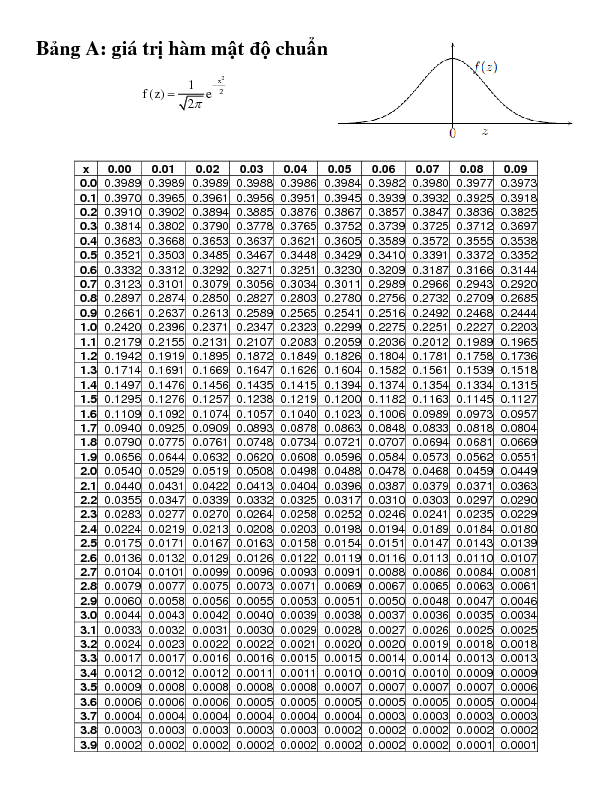
Giải:

. Tra bảng ta được:

3.1.3 Phân bố

**Bảng giá trị hàm mật độ và hàm phân phối chuẩn tắc**



****

1. **Các tham số đặc trưng của biến ngẫu nhiên**

**4.1.Kì vọng toán**

**Định nghĩa**

Kì vọng hoặc giá trị trung bình(average, mean value, expected value) của biến ngẫu nhiên X kí hiệu là EX và được xác định như sau:

* Nếu X rời rạc nhận các giá trị với xác suất tương ứng thì

(1)

* Nếu X liên tục có hàm mật độ f(x) thì: (2)

Kì vọng EX tồn tại nếu chuỗi (1)(trường hợp rời rạc) hội tụ tuyệt đối hoặc tích phân (2)(trường hợp liên tục) hội tụ tuyệt đối.

**Ý nghĩa của kì vọng**

Kì vọng mang ý nghĩa là giá trị trung bình mà biến ngẫu nhiên nhận được. Giả sử biến ngẫu nhiên X nhận các giá trị với các tần số tương ứng

là tổng giá trị X nhận được với cùng giá trị . Do đó là tổng tất cả các giá trị X nhận được. là giá trị trung bình của X, trong đó

được gọi là tần suất nhận giá trị của X. Trong trường hợp tổng quát thì tần suất được thay thế bằng xác suất

Khái niệm kì vọng được áp dụng rộng rãi trong nhiều lĩnh vực. Trong kinh doanh và quản lý, kì vọng được ứng dụng dưới dạng lợi nhuận kì vọng hay doanh số kì vọng

**Tính chất**

* E(C) = C với mọi hằng số C
* E(C.X) = C.E(X) với mọi hằng số C
* E(
* Nếu độc lập thì

**Ví dụ:**

Tung con xúc sắc n lần. Tìm kì vọng của tổng số nốt thu được

Gọi là số nốt thu được ở lần tung thứ i, gọi X là tổng số nốt thu được trong n lần tung. Như vậy

Các biến ngãu nhiên đều có phân bố xác suất như sau

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 |
| P |  |  |  |  |  |  |

Do đó

**4.2. Phương sai**

Định nghĩa:

Phương sai(variance) hay độ lệch bình phương trung bình của biến ngẫu nhiên X là đại lượng đo sự phân tán bình phương trung bình của X quanh giá trị trung bình EX.

Phương sai của X được kí hiệu là DX hay varX và được định nghĩa như sau:

(\*)

được gọi là độ lệch chuẩn(deviation) của X. Độ lệch chuẩn là một đại lượng thống kê mô tả dùng để đo mức độ phân tán của một tập dữ liệu đã được lập thành bảng tần số. Có thể tính ra độ lệch chuẩn bằng căn bậc hai của phương sai

Khai triển vế phải của công thức (\*) và các tính chất của kì vọng ta có thể tính phương sai theo công thức:

**Tính chất**

* D(a) = 0 với mọi hằng số a
* D(aX) = với mọi hằng số a
* với mọi hằng số a, b
* Nếu độc lập có các phương sai hữu hàn thì

**Phương sai của tổng thể chung và phương sai mẫu**

Trên nhiều tình huống thực tế, giá trị chính xác của phương sai của một tổng thể, kí hiệu bởi là không thể xác định được

Phương pháp chung để ước lượng phương sai của một tổng thể(hữu hạn hoặc vô hạn) là ta sẽ lấy một mẫu hữu hạn các cá thể từ quần thể. Giả sử mẫu thu được các giá trị đo được

Phương sai mẫu của một dãy các đại lượng được tính bởi:



trong đó  là bình quân số học của mẫu.

Tuy nhiên là một ước lượng chệch(biased) của phương sai quần thể. Ước lượng sau là một ước lượng không chệch(unbiased) của phương sai quần thể:



1. **Kiểm định giả thuyết thống kê**

**5.1.Khái niệm chung**

Giả thuyết thống kê là giả thuyết về dặng phân bố xác suất, các đặc trưng tham số của biến ngẫu nhiên hoặc giả thiết về sự độc lập của các biến ngẫu nhiên gốc

Giả thiết đưa ra kiểm nghiệm được kí hiệu là gọi là “giả thiết không”. Đó là giả thiết mà ta nghi ngờ muốn bác bỏ hoặc giả thiết ta muốn bảo vệ. Ngoài giả thiết ra, ta còn phải định ra một giả thiết cạnh tranh với , gọi là đối thiết, kí hiệu . Đối thiết sẽ được chấp nhận khi bị bác bỏ.

Cần chú ý rằng đối thiết không nhất thiết là phủ định của giả thiết . Chẳng hạn giả thiết : nhu cầu thị trường về loại hàng hóa này là đơn vị/tháng. Nếu nghi ngờ nhu cầu này không đúng thì đối thiết là , nhưng nếu do tiếp thị tốt, do chính sách hậu mãi tốt, người ta nghĩ rằng nhu cầu về mặt hàng này tăng lên thì đối thiết là

Quy tắc kiểm định dựa trên hai nguyên lí sau:

* Nguyên lí xác suất nhỏ: “Nếu một biến cố có xác suất rất nhỏ thì trong một hay vài phép thử thì biến cố đó coi như không xảy ra”
* Phương pháp phản chứng: “Để bác bỏ A ta giả sử A đúng thì dẫn đến một điều vô lí”

Dựa vào hai nguyên lý này ta đưa ra phương pháp chung để kiểm định một giả thiết thống kê như sau: Để kiểm định trước hết giả sử đúng từ đó ta tìm được biến cố A mà xác suất xuất hiện biến cố A là rất bé và ta có thể xem A không thể xảy ra trong một phép thử về biến cố này. Lúc đó nếu trên một mẫu cụ thể quan sát được mà biến cố A xuất hiện thì điều này trái với nguyên lý xác suất nhỏ. Vậy sai và bác bỏ nó. Còn nếu A không xảy ra thì ta có cơ sở để bác bỏ

Ta thực hiện phương pháp trên bằng các bước cụ thể như sau

**5.1.2. Các loại giả thuyết thống kê**

*Giả thuyết* (The null hypothesis)

Để dễ hiểu, kí hiệu được kí hiệu cho các tham số của tổng thể như trung bình (, phương sai hoặc tỉ lệ (p). Vậy giả thuyết là tham số của tổng thể thì bằng với giá trị với giá trị cụ thể nào đó trong trường hợp giả thuyết đó có giá trị đơn, nghĩa là (kiểm định hai đuôi) hoặc giả thuyết là một dãy của các giá trị, lúc đó hay (kiểm định một đuôi)

*Giả thuyết*

Giả thiết là kết quả ngược lại của giả thiết , nếu giả thuyết đúng thì giả thuyết sai và ngược lại. Vậy cặp giả thuyết và được thể hiện trong các trường hợp kiểm định như sau:

* Trường hợp kiểm định dạng hai đuôi(Two – tail test):
* Trường hợp kiểm định dạng một đuôi(One-tail test):

hoặc

**5.1.2. Tiêu chuẩn kiểm định giả thiết thống kê**

Từ biến ngẫu nhiên gốc X của tổng thể lập mẫu ngấu nhiên W = . Chọn thống kê **(\*)**

trong đó là tham số liên quan đến giả thiết cần kiểm định. Nếu đúng thì thống kê có quy luật phân bố xác suất xác định. Thống kê T được gọi là tiêu chuẩn kiểm định

**5.1.3. Miền bác bỏ giả thiết**

Sau khi đã chọn tiêu chuẩn kiểm định T, với  bé cho trước (thường được lấy bằng 0.05 hoặc 0.01) và với điều kiện đúng ta có thể tìm được miền sao cho T nhận giá trị trong miền  với xác suất bằng

Giá trị được gọi là mức ý nghĩa của kiểm định và miền gọi là miền bác bỏ giả thiết với mức ý nghĩa

**5.1.4. Giá trị quan sát của tiêu chuẩn kiểm định**

Thực hiện phép thử với mẫu ngẫu nhiên thu được mẫu cụ thể , thay giá trị này vào thống kê (\*) ta được giá trị quan sát của tiêu chuẩn kiểm định

**5.1.Quy tắc kiểm định giả thiết thống kê**

So sánh giá trị quan sát của tiêu chuẩn kiểm định với miền bác bỏ và kết luận theo quy tắc sau:

* Nếu , theo nguyên tắc kiểm định thì sai, do đó ta bác bỏ thừa nhận
* Nếu không thuộc thì điều này chưa khẳng định rằng đúng mà chỉ có nghĩa là qua mẫu cụ thể này chưa khẳng định được sai. Do đó chỉ có thể nói rằng qua mẫu cụ thể này chưa có cơ sở để bác bỏ (trên thực tế là thừa nhận

**5.2. Sai lầm loại một và sai lầm loại hai**

Sai lầm loại một (Type I error:

Là sai lầm của việc bác bỏ giả thuyết khi giả thuyết này đúng ở mức nghĩa nào đó của kiểm định, nghĩa là nếu quyết định xác suất bác bỏ giả thuyết khi giả thuyết này đúng là thì xác suất để chấp nhận nó là

Sai lầm loại hai(Type II error:

Ngược lại sai lầm lọa I là sai lầm loại II, là việc chấp nhận giả thuyết khi giả thuyết này sai. Nếu xác suất của việc quyết định chấp nhận một giả thuyết sai được kí hiệu là thì xác suất để bác bỏ giả thuyết này là

Những quyết định dựa trên giả thuyết được tóm tắt như sau:

|  |  |  |
| --- | --- | --- |
|  | đúng | sai |
| Bác bỏ | Sai lầm loại I  Xác suất | Quyết dịnh đúng  Xác suất = |
| Không bác bỏ | Quyết định đúng  Xác suất = | Sai lầm loại II  Xác suất = |

**5.3. Thủ tục kiểm định giả thuyết thống kê**

Thủ tục kiểm định giả thuyết thống kê trên bao gồm các bước:

* Phát biểu giả thiết và đối thiết
* Từ tổng thể, nghiên cứu lập mẫu kích thước n
* Chọn tiêu chuẩn kiểm định T và xác định quy luật phân bố xác suấy của T với điều kiện giả thiết đúng
* Với mức ý nghĩa , xác định miền bác bỏ tốt nhất tùy thuộc vào đối thiết
* Từ mẫu cụ thể tính giá trị quan sát tiêu chuẩn kiểm định
* So sánh giá trị quan sát của tiêu chuẩn kiểm định với miền bác bỏ và kết luận

**5.4. Kiểm định giả thiết kì vọng của biến ngẫu nhiên có phân bố theo quy luật chuẩn**

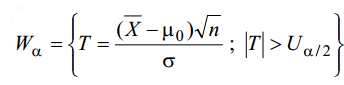
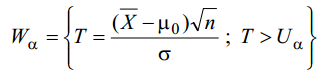
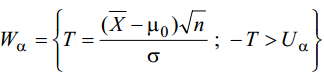
Giả sử biến ngẫu nhiên gốc X trong tổng thể có phân bố chuẩn , cần kiểm định kì vọng . Nếu có cơ sở để giả thiết rằng kì vọng bằng giá trị ta đưa ra giả thiết thống kê

Ta xét các trường hợp sau đây:

**Trường hợp đã biết phương sai**

Giả sử phương sai của biến ngẫu nhiên gốc X trong tổng thể có phân bố chuẩn đã biết. Từ tổng thể rút ra một mẫu ngẫu nhiên kích thước n: W = (. Xét thống kê

Nếu giả thiết đúng, ta xây dựng các miền bác bỏ dựa vào đối thiết

1. Bài toán 1: . Ta nói đây là bài toán kiểm định hai phía. Miền bác bỏ: 
2. Bài toán 2: . Đây là bài toán kiểm định một phía. Miền bác bỏ: 
3. Bài toán 3: . Đây là bài toán kiểm định một phía. Miền bác bỏ: 

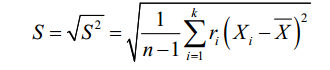
trong đó: , lần lượt là giá trị tới hạn mức và mức của phân bố chuẩn tắc

Lập mẫu cụ thể và tính giá trị quan sát của tiêu chuẩn kiểm định và so sánh với miền bác bỏ để kết luận

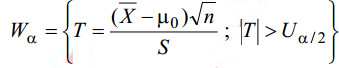
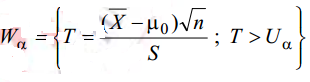
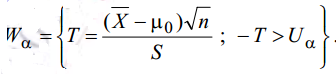
**Trường hợp chưa biết phương sai, kích thước mẫu**

Trường hợp phương sai chưa biết: với kích thước đủ lớn () và giả thiết đúng, ta có thống kê

**C:\Users\Administrator\Desktop\Capture.PNG**

trong đó S là độ lệch tiêu chuẩn mẫu được xác định bởi 

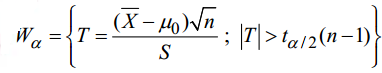
Thống kê T xấp xỉ phân bố chuẩn tắc . Ta xây dựng các miền bác bỏ dựa vào đối thiết :

1. Bài toán 1: . Ta nói đây là bài toán kiểm định hai phía. Miền bác bỏ: 
2. Bài toán 2: . Đây là bài toán kiểm định một phía. Miền bác bỏ: 
3. Bài toán 3: . Đây là bài toán kiểm định một phía. Miền bác bỏ: 

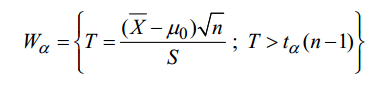
**Trường hợp chưa biết phương sai, kích thước mẫu**

Giả sử giả thiết đúng, xét thống kê

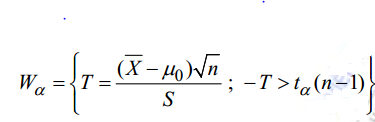
1. Bài toán 1: . Miền bác bỏ



trong đó (n-1) là giá trị tới hạn mức của mức phân bố Student n-1 bậc tự do

1. Bài toán 2: . Miền bác bỏ

trong đó là giá trị mức của phân bố Student n-1 bậc tự do

1. Bài toán 3: . Miền bác bỏ: 

**Chương II: Mô hình ARIMA**

**1. Khái niệm về chuỗi thời gian dừng, kiểm định tính dừng**

Trong toán học, tính dừng(Stationarity) được sử dụng như một công cụ trong phân tích chuỗi số liệu theo thời gian. Để hình thành một mô hình đầy đủ ý nghĩa thống kê thì chuỗi dữ liệu theo thời gian trước tiên cần kiểm tra tính dừng của chuỗi. Một quá trình có tính dừng của chuỗi. Một quá trình có tính dừng là một quá trình ngẫu nhiên, được thể hiện bởi trung bình mẫu và phương sai mẫu không đổi theo thời gian. Trên thực tế hầu hết các chuỗi dữ liệu(chuỗi gốc) đều không dừng. Điều này được hiểu là các chuỗi thời gian có trung bình mẫu và phương sai mẫu thay đổi theo thời gian. Nhưng khi ta lấy sai phân thì các chuỗi thời gian thường trở thành chuỗi dừng.

1. **Mô hình ARIMA**
   1. Hàm tự tương quan ACF(AutoCorrelation Function)

Tự tương quan là sự tương quan giữa các thành phần của chuỗi các quan sát được sắp xếp theo thứ tự thời gian.

Hàm tự tương quan đo lường phụ thuộc tuyến tính giữa các cặp quan sát y(t) và y(t+k) ứng với thời đoạn k = 1, 2, …(k còn gọi là độ trễ, giá trị max của k thường là 20 và được kí hiệu là maxLag = 20). Với mỗi độ trễ k, hàm tự tương quan tại độ trễ k được xác định qua độ lệch giữa các biến ngẫu nhiên y(t) và y(t+k) so với các giá trị trung bình, và được chuẩn hóa qua phương sai.

Giả sử có một chuỗi thời gian dừng thay đổi quanh giá trị trung bình với phương sai hằng số . Hàm tự tương quan tại các độ trễ khác nhau sẽ có giá trị khác nhau.

Ta có thể ước lượng hàm tự tương quan tại độ trễ thứ k qua phép biến đổi trung bình của tất cả các cặp quan sát. Giả sử có k độ trễ, giá trị trung bình mẫu là được chuẩn hóa bởi phương sai . Chẳng hạn cho một chuỗi có n điểm, giá r(k) của hàm tự tương quan tại độ trễ thứ k được tính như sau:

với và

Ghi chú:

* y(t): chuỗi thời gian dừng tại thời điểm t
* y(t + k) (k = 1, 2,… maxLag): chuỗi thời gian dừng tại thời điểm (t + k)
* : giá trị trung bình của chuỗi dừng
* : phương sai của chuỗi dừng
* : giá trị tương quan giữa y(t) và y(t + k) tại độ trễ k

Nếu r(k) = 0 thì không có hiện tượng tương quan. Tính r(k) cho các độ trễ k = 1, 2, 3, … sau đó vẽ biểu đồ ACF của mẫu dữ liệu. Biểu đồ có công dụng xác định xem hiện tượng dừng của chuỗi để nhận dạng mô hình ARIMA dự định.

**Tính chất của hàm ACF:**

* r(k) không có thứ nguyên
* Giá trị của r(k) nằm giữa -1 và 1. Hệ số tương quan bằng 1 trong trường hợp có tương quan tuyến tính đồng biến và -1 trong trường hợp tương quan tuyến tính nghịch biến. Các giá trị trong khoảng (-1, 1) cho biết mức độ phụ thuộc tuyến tính giữa các biến. Hệ số tương quan càng gần với -1 thì tương quan giữa các biến càng mạnh

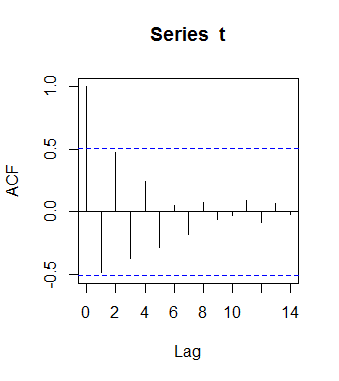
**Nhận xét:**

* Về mặt lý thuyết, chuỗi dừng khi tất cả các r(k) = 0 hay chỉ vài r(k) khác không. Do chúng ta xem xét hàm tự tương quan mẫu, do đó sai số mẫu sẽ xuất hiện vì vậy, hiện tượng tự tương quan khi r(k) = 0 theo ý nghĩa thống kê.
* Hàm tự tương quan ACF giảm đột ngột sau độ trễ k nếu không có những đỉnh ở độ trễ k lớn hơn k. Hầu hết hàm tự tương quan ACF sẽ giảm đột ngột sau độ trễ 1, 2

**Ví dụ:** giả sử t có một chuỗi các giá trị đưa vào (42, 59, 35, 66, 37, 58, 49,  
63, 52, 45, 36, 50, 43, 39, 52).

Khi đó:

* Giá trị trung bình mẫu:
* Phương sai mẫu:
* Tính hàm tự tương quan tại độ trễ k:
  + k = 0, hàm ACF có giá trị 1 (hàm tự tương quan với chính nó)
  + k = 1, hàm ACF có giá trị -0.482 (mô tả hàm tự tương quan tại độ trễ thứ một dao động trung bình giữa các quan sát liên tục: 42 và 59, 59 và 35, 35 và 66,…)
* Áp dụng tương tự cho các độ trễ khác



* 1. **Hàm tự tương quan riêng phần PACF (Partial AutoCorrelation Function)**

Song song với việc xác định hàm tự tương quan giữa các cặp ta xác định hàm tự tương quan từng phần cùng có hiệu lực trong việc can thiệp đến các quan sát . Hàm tự tương quan từng phần tại độ trễ k được ước lượng bằng hệ số liên hệ trong mối kết hợp tuyến tính bên dưới. Sự kết hợp được tính dựa trên tầm ành hưởng của và các giá trị trung gian

1. **Các phương pháp tính giá trị hàm PACF**

1.Giải phương trình hồi quy dựa trên phương pháp bình phương tối thiểu vì hệ số hồi quy phải được tính ở mỗi độ trễ k, với j chạy từ 1 đến k.

2. Giải pháp ít tốn kém hơn do Durbin phát triển dùng để xấp xỉ đệ quy hệ số hồi quy cho mô hình ARIMA chuỗi dừng, sử dụng giá trị hàm tự tương quan tại độ trễ k và hệ số hồi quy của độ trễ trước. Phương pháp sau sử dụng 3 độ trễ đầu tiên:

* Độ trễ 1: khởi tạo, giá trị của hàm tự tương quan từng phần tại độ trễ 1 có cùng giá trị với hàm tự tương quan tại độ trễ 1 vì không có trung gian giữa các quan sát kế tiếp:
* Độ trễ 2: Hai giá trị và được tính dựa vào hàm tự tương quan và , cùng với hàm tự tương quan từng phần trước đó
* Độ trễ 3: Tương tự, ba giá trị được tính vào các hàm tự tương quan trước cùng với các hệ số được tính ở độ trễ thứ 2: và

Tổng quan, hàm tự tương quan từng phần được tính theo Durbin:

**Ghi chú:**

* Hàm r(k): hàm tự tương quan tại độ trễ k
* C(kj): hàm tự tương quan riêng phần cho độ trễ k, loại bỏ những ảnh hưởng của các độ trễ can thiệp

**Nhận xét:**

* Khi độ trễ tăng, số các hệ số tăng theo. Phương pháp của Durbin cho phép tính đệ quy dựa vào việc sử dụng kết quả trước đó
* Tại độ trễ thứ một, giá trị hàm tự tương quan riêng phần PACF chính là giá trị hàm tự tương quan ACF. Giá trị âm cho thấy hệ số âm giữa các khoảng thời gian kề nhau

Tóm lại, hàm tự tương quan ACF và hàm tự tương quan từng phần PACF của chuỗi thời gian có các đặc tính khác nhau. Hàm tự tương quan ACF đo mức độ phụ thuộc tuyến tính giữa các cặp quan sát. Hàm tự tương quan từng phần PACF đo mức độ phụ thuộc tuyến tính từng phần. ARIMA sử dụng hai đặc tính này để xác định cấu trúc mô hình cho chuỗi thời gian được thực hiện trong luận văn này

Xu hướng đồ thị của hàm tự tương quan từng phần PACF có thể giảm đột ngột(thường sau độ trễ 1 hoặc 2) hay có thể giảm đều. Cũng như hàm tự tương quan ACF, xu hướng giảm đều của hàm tự tương quan từng phần PACF cũng có các dạng phân phối mũ, dạng sóng hình sin hoặc kết hợp cả 2 dạng này

* 1. **. Mô hình tự hồi quy bậc p (Auto Regression) – AR(p)**

Phân tích hồi quy là sự nghiên cứu mối liên hệ phụ thuộc của một biến(gọi là biến phụ thuộc) vào một hay nhiều biến khác(gọi là biến độc lập), với ý tưởng ước lượng và dự đoán giá trị trung bình tổng thể của biến phụ thuộc trên cơ sở giá trị biết trước(trong mẫu) của các biến độc lập.

Phương trình tự hồi quy:

Ý tưởng chính của mô hình AR(p) là hồi quy trên chính số liệu quá khứ ở những chu kì trước.

**Xây dựng công thức**

Mô hình tự hồi quy bậc một được đề cập chỉ xét tương quan giữa giá trị liên nhau (, ) trong chuỗi thời gian:

Mô hình tự hồi quy bậc 2 xem xét ảnh hưởng của tương quan hệ giữa 2 giá trị kì trước đó:

Tổng quát, mô hình hồi quy bậc p được biểu diễn bởi công thức

Trong đó:

* : quan sát dừng hiện tại
* ,…. :quan sát dừng quá khứ(thường sử dụng không quá 2 biến này)
* : các tham số phân tích hồi quy
* : sai số ngẫu nhiên của giai đoạn hiện tại. Giá trị trung bình được mong đợi bằng 0

**Nhận xét:**

* là một hàm tuyến tình của những quan sát dừng ,…Nói cách khác khi sử dụng phân tích hồi quy theo các giá trị chuỗi thời gian dừng có độ trễ, chúng ta sẽ được mô hình AR
* Số quan sát dừng quá khứ sử dụng trong mô hình hàm tự tương quan là bậc p của mô hình AR. Nếu ta sử dụng một quan sát dừng quá khứ, ta có mô hình tương quan bậc một AR(1). Nếu ta sử dụng hai quan sát quá khứ, ta có mô hình tương quan bậc hai AR(2).
* Điều kiện dừng là tổng các tham số hồi quy nhỏ hơn một:
* Mô hình AR(1):
* Mô hình AR(2):
  1. **Mô hình trung bình trượt bậc q(Moving Average) – MA(q)**

Quan sát dừng hiện tại là một hàm tuyến tính phụ thuộc các biến sai số dự báo quá khứ và hiện tại. Trong mô hình trung bình trượt, quá trình được mô tả hoàn toàn bằng tổng trọng số của các ngẫu nhiên hiện hành có độ trễ:

Trong đó:

* : quan sát dừng hiện tại
* sai số dự báo ngẫu nhiên, giá trị của nó không được biết và giá trị trung bình của nó là 0
* : sai số dự báo quá khứ(thông thường mô hình sẽ sử dụng không quá 2 biến này)
* : giá trị trung bình của và các hệ số bình quân di động
* : sai số trong quá khứ được dùng trong mô hình bình quân di động. Điều kiện cần là tổng các hệ số bình quân di động phải nhỏ hơn 1:
* Mô hình MA(1):
* Mô hình MA(2):
  1. **Sai phân I(d)**

**Chuỗi dừng:**

Theo Gajarati(2003) một chuỗi thời gian là dừng nếu như trung bình và phương sai của nó không đổi theo thời gian và giá trị đồng phương sai giữa hai thời đoạn chỉ phụ thuộc vào khoảng cách và độ trễ về thời gian giữa hai thời đoạn này chứ không phụ thuộc vào thời điểm thực tế mà đông phương sai được tính.

Chuỗi dừng có xu hướng trở về giá trị trung vình và những dao động quanh giá trị trung bình là như nhau. Nói cách khác, một chuỗi thời gian không dừng sẽ có giá trị trung bình thay đồi theo thời gian, hoặc phương sai thay đổi theo thời gian hoặc cả hai

Có nhiều phương pháp kiểm tra tính dừng của chuỗi thời gian: kiểm định Dickey-Fuller(DF), kiểm định Philip-Person(PP) và kiểm đinhk Dickey và Fuller mở rộng(ADF). Trong phạm vi của đồ án em chỉ đề cập đến kiểm định Dickey-Fuller(DF)

**Bậc tích hợp**

Hầu hết các chuỗi thời gian về kinh tế là chuỗi không dừng vì chúng thường có một xu hướng tuyến tính hoặc mũ theo thời gian. Tuy nhiên có thể biến đổi chúng về chuỗi dừng thông qua quá trình sai phân. Nếu sai phân bậc 1 của một chuỗi có tính dừng thì chuỗi ban đầu gọi là tích hợp bậc 1, kí hiệu là I(1). Tương tự, nếu sai phân bậc d của một chuỗi có tính dừng thì chuỗi ban đầu gọi là tích hợp bậc d, kí hiệu là I(d). Nếu chuỗi ban đầu(chưa lấy sai phân) có tính dừng thì gọi là I(0)

Sai phân chỉ sự khác nhau giữa giá trị hiện tại và giá trị trước đó. Phân tích sai phân làm cho ổn định giá trị trung bình của chuỗi dữ liệu, giúp cho việc chuyển đổi chuỗi thành một chuỗi dừng

**VD**

Sai phân lần một I(1): z(t) = y(t) – y(t-1)

Sai phân lần hai I(2): h(t) = z(t) – z(t-1)

* 1. **Mô hình ARIMA**
     1. Mô hình ARMA

Mô hình ARMA(p,q): là mô hình hỗn hợp của AR và MA. Hàm tuyến tính sẽ bao gồm những quan sát dừng quá khứ và sai số dự báo quá khứ và hiện tại

Trong đó:

* quan sát dừng hiện tại
* và : quan sát dừng và sai số dự báo quá khứ
* : các hệ số phân tích hồi quy

Ví dụ: ARMA(1, 1) là mô hình hỗn hợp của AR(1) và MA(1)

Đối với mô hình hỗn hợp thì dạng (p, q) = (1, 1) là khá phổ biến. p, q là tham số độ trễ của hàm ACF và PACF. Cả hai điều kiện trung bình di động và điều kiện dừng phải được thỏa mãn trong mô hình ARMA

* + 1. **Mô hình ARIMA(p, d, q)**

Hai tác giả George Box và Gwilym Jenkins(1976) đã nghiên cứu mô hình tự hồi quy tích hợp trung bình trượt(Autoregressive Intergrated Moving Average), viết tắt là ARIMA. Tên của họ(Box-Jenkins) được dùng để gọi cho các quá trình ARIMA tổng quát áp dụng vào phân tích dự báo chuỗi thời gian. Mô hình tự tương quan tích hợp với trung bình trượt có dạng ARIMA(p, d, q) là hỗn hợp của 2 mô hình AR và MA

Do mô hình Box\_Jekins chỉ mô tả chuỗi dừng hoặc những chuỗi đã sai phân hóa, nên mô hình ARIMA(d, p, q) thể hiện những chuỗi dữ liệu không dừng, đã được sai phân với d là bậc sai phân của chuỗi thời gian được khảo sát, p là bậc tự hồi quy và q là bậc trung bình trượt. Công thức chung của mô hình ARIMA:

Về mặt lý thuyết, không có trường hợp hàm tự tương quan ACF và hàm tự tương quan từng phần cùng giảm đột ngôt. Trong thực tế, hàm tự tương quan ACF và hàm tự tương quan từng phần PACF giảm đột biến khá nhanh. Trong trường hợp này, chúng ta nên phân biệt hàm nào giảm đột biến nhanh hơn, hàm còn lại xem là giảm đều. Do đôi lúc sẽ có trường hợp giảm đột biến đồng thời khi quan sát biểu đồ hàm tự tương quan ACF và hàm tự tương quan từng phần PACF, do đó ta cần tìm vài hàm dự định khác nhau cho chuỗi thời gian dừng. Sau đó, kiểm tra độ chính xác mô hình tốt nhất bằng tiêu chuẩn thống kê.

**Ví dụ:**

Mô hình ARIMA(1,1,1) với sai phân bậc 1 do d = 1:

)

hoặc

Mô hình ARIMA(1,2,1): với sai phân thứ hai: d = 2

* + 1. **Các bước xây dựng mô hình ARIMA**

Phương pháp Box-Jenkins gồm bốn bước là:

* Nhận dạng mô hình thử nghiệm
* Ước lượng các tham số p, d, q
* Kiểm định bằng chuẩn đoán
* Ước lượng, dự báo kết quả trong tương lai

**Bước 1: Nhận dạng mô hình**

Mô hình ARIMA chỉ được áp dụng với chuỗi dừng. Mô hình có thể trình bày theo dạng AR, MA hay ARMA.

Nhận dạng mô hình ARIMA(p, d, q) là việc tìm các giá trị thích hợp của p, d, và q. Với d là bậc sai của chuỗi thời gian được khảo sát, p là bậc tự hồi quy và q là bậc trung bình trượt.

Việc xác định p và q sẽ phụ thuộc vào các đồ thị SPAC = f(t) và SAC = f(t) với SAC là hàm tự tương quan mẫu(Sample Autocorreclation) và SPAC là hàm tự tương quan từng phần của mẫu(Sample Partial Autocorrelation):

* Chọn giá trị của p nếu đồ thị SPAC có giá trị cao tại độ trễ 1, 2,…, p và giảm nhiều sau p đồng thời hàm SAC giảm dần
* Chọn giá trị của q nếu đồ thị SAC có giá trị cao tại độ trễ 1, 2,.., q và giảm nhiều sau q đồng thời hàm SPAC giảm dần

Nếu chuỗi ARIMA không dừng: cần phải được chuyển đổi thành chuỗi dừng trước khi tính ước lượng tham số bình phương tối thiểu. Việc chuyển đổi này được thực hiện bằng cách tính sai phân giữa các giá trị quan sát.

**Bước 2: Ước lượng tham số**

Tính những ước lượng khởi đầu cho các tham số của mô hình dự định. Các tham số của mô hình ARIMA sẽ được ước lượng qua phương pháp maxium likelihood

**Bước 3: Kiểm định mô hình**

Sau khi ước lượng các tham số của một mô hình ARIMA được nhận dạng thử, chúng ta cần phải kiểm định để kiểm nghiệm rằng mô hình có thích hợp hay không. Cách thức để thực hiện điều này:

Kiểm tra phần dư có là nhiễu trắng hay không. Nếu là nhiễu trắng thì chấp nhận mô hình, trong trường hợp ngược lại chúng ta phải tiến hành từ đầu. Tính phù hợp của mô hình dựa trên tiêu chuẩn Schwarz(BIC) và sai số chuẩn SEE càng nhỏ càng tốt.

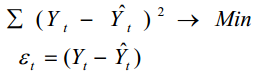
**Bước 4: Dự báo kết quả**

Dựa trên phương trình của mô hình ARIMA, tiến hành xác định dự báo điểm và khoảng tin cậy của dự báo. Khi mô hình thích hợp với dữ liệu đã tìm được, ta sẽ thực hiện dự báo tại thời điểm tiếp theo t.

* Dự báo điểm:
* Khoảng tin cậy: y:

Với độ tin caayh 95%, k = 2

**Ước lượng tham số với phương pháp Maximum likelihood(MLE)**

Các thông số của mô hình ARIMA sẽ được xác định theo phương pháp Maximum likelihood

Phương pháp có thể tổng quát như sau: Nếu x là biến ngấu nhiên với hàm phân bố: ,

với , là K tham số cần ước lượng với dãy n mẫu độc lập là x[1], x[2],…, x[N]. Thì hàm likelihood được cho bởi tích sau:

,= và hàm ln likelihood như sau:

MLE của đạt được khi hay  lớn nhất, chúng ta đã biết xác định giá trị lớn nhất với dễ hơn với , vậy MLE của là giải K hệ phương trình sau: = 0 ( j = 1, 2,…K)

Phương pháp MLE cho mô hình ARIMA

Xét mô hình ARIMA có dạng:

Giả sử có n quan sát đã biết: . Tương ứng với n quan sát này ta có n nhiễu trắng:

Gọi và lần lượt là trung bình và phương sai của dãy . Do là nhiễu trắng nên nó trung bình bằng 0, mỗi có hàm mật độ

g(

Gọi X là biến ngẫu nhiên có phân bỗ  .Do là các biến độc lập nên hàm mật độ của X có dạng

g(X) = g(

Tuy nhiên sẽ rất khó để xác định chính xác hàm mật độ trên. Do vậy, người ta giả sử đã biết p giá trị và q giá trị . Xét 2 vector đã biết và = .

Xét hệ phương trình sau:

……

+

Gọi a = (), b = , y =

Nhận xét nếu các vector a, b và tham số biết thì từ hệ phương trình trên ta hoàn toàn có thể xác định các giá trị . Vậy là một hàm xác định theo các tham số a, b, tức là

Vậy hàm mật độ của X được cho bởi và  là

f(y) = g(y| với

Ta cần xác định các vector a, b và tham số sao cho hàm f(y) trên đạt giá trị max. Xét hàm log likelihood của hàm f(y) trên

Lấy đạo hàm của hàm h(y) theo và cho bằng 0 ta được

Sử dụng phương pháp lặp để tìm min

**Kiểm tra mô hình phù hợp nhất**

Ta kí hiệu

TSS(Total Sum of Squares) là tổng bình phương của tất cả các sai lệch giữa các giá trị quan sát với giá trị trung bình của chúng



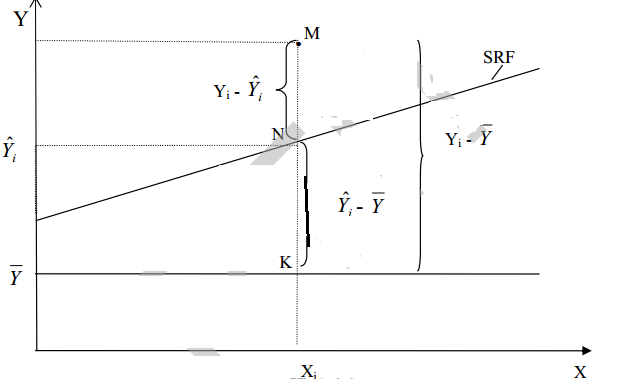
ESS(Explained Sum of Squres) là tổng bình phương của tất cả các sai lệch giữa giá trị của biến Y tính theo hàm hồi quy mẫu với giá trị trung bình. Phần này đo độ chính xác của các hàm hồi quy



RSS(Residual Sum of Squares) là tổng bình phương của tất cả các sai lệch giữa các giá trị quan sát của biến Y với các giá trị nhận được từ hàm hồi quy mẫu.

Người ta đã chứng minh được rằng:

Nếu hàm hồi quy mẫu phù hợp tốt với các số liệu quan sát thì ESS sẽ càng lớn hơn RSS. Nếu tất cả các giá trị quan sát của Y đều nằm trên SRF thì ESS sẽ bằng TSS và do đó RSS = 0. Ngược lại nếu hàm hồi quy mẫu kém phù hợp với các giá trị quan sát thì RSS sẽ càng lơn hơn ESS



Phân tích độ thích hợp hồi quy

Tóm lại trong các giá trị p và q tìm dược, chọn mô hình ARIMA nào có hệ số nhỏ nhất

**Kiểm tra chuẩn đoán mô hình**

Sau khi xác định p, d, q và các hệ số nghĩa là đã xác định được phương trình cho mô hình ARIMA, điểu cần phải làm là tiến hành xem số hạng của mô hình có phải là một nhiễu trắng(white noise, nhiễu ngẫu nhiên thuần túy) hay không. Đây là yêu cầu của một mô hình tốt

Về mặt lý thuyết, được tạo ra bởi quá trình nhiễu trắng nếu:

Việc kiểm tra tính nhiễu trắng sẽ dựa trên đồ thị SAC của chuỗi

**Chương III: Ngôn ngữ lập trình R**

**1. Tổng quan về R**

* 1. **R là gì**

R là một phần mềm sử dụng cho phân tích thống kê và đồ thị. Thật ta, về bản chất, R là ngôn ngữ máy tính đa năng, có thể sử dụng cho nhiều mục tiêu khác nhau, từ tính toán đơn giản, toán học giải trí(recreational methematics), tính toán ma trận(matrix) đến các phân tích thống kê phức tạp. Vì R là một ngôn ngữ, cho nên người ta có thể sử dụng R để phát triền thành các phần mềm chuyên môn cho một vấn đền tính toán đặc biệt

Hai người sáng tạo ra R là hai nhà thống kê học tên là Ross Ihaka và Robert Gentleman. Kể từ khi R ra đời, rất nhiều nhà nghiên cứu thống kê và toán học trên thế giới ủng hộ và tham gia vào việc phát triển R. Chủ trương của những người sáng tạo ra R là theo định hướng mở rộng (Open Access). Cũng một phần vì chủ trương này mà R hoàn toàn miễn phí. Bất cứ ai ở bất cứ nơi nào trên thế giới đều có thể truy nhập và tải toàn bộ mã nguồn của R về máy tính của mình để sử dụng. Cho đến nay, chỉ qua chưa đầy 5 năm phát triển, càng ngày càng có nhiều các nhà thống kê học, toán học, nghiên cứu trong mọi lĩnh vực đã chuyển sang sử dụng R để phân tích dữ liệu khoa học. Trên toàn cầu, đã có một mạng lưới gần một triệu người sử dụng R, và con số này đang tăng theo cấp số nhân. Có thể nói trong vòng 10 năm nữa, chúng ta sẽ không cần đến các phần mềm thống kê đắt tiến như SAS, SPSS hay Stata (các phần mềm này rất đắt tiền, có thể lên đến 100.000 USD một năm) để phân tích thống kê nữa, vì tất cả các phân tích đó có thể tiến hành bằng R.

* 1. **Cài đặt và chạy R**

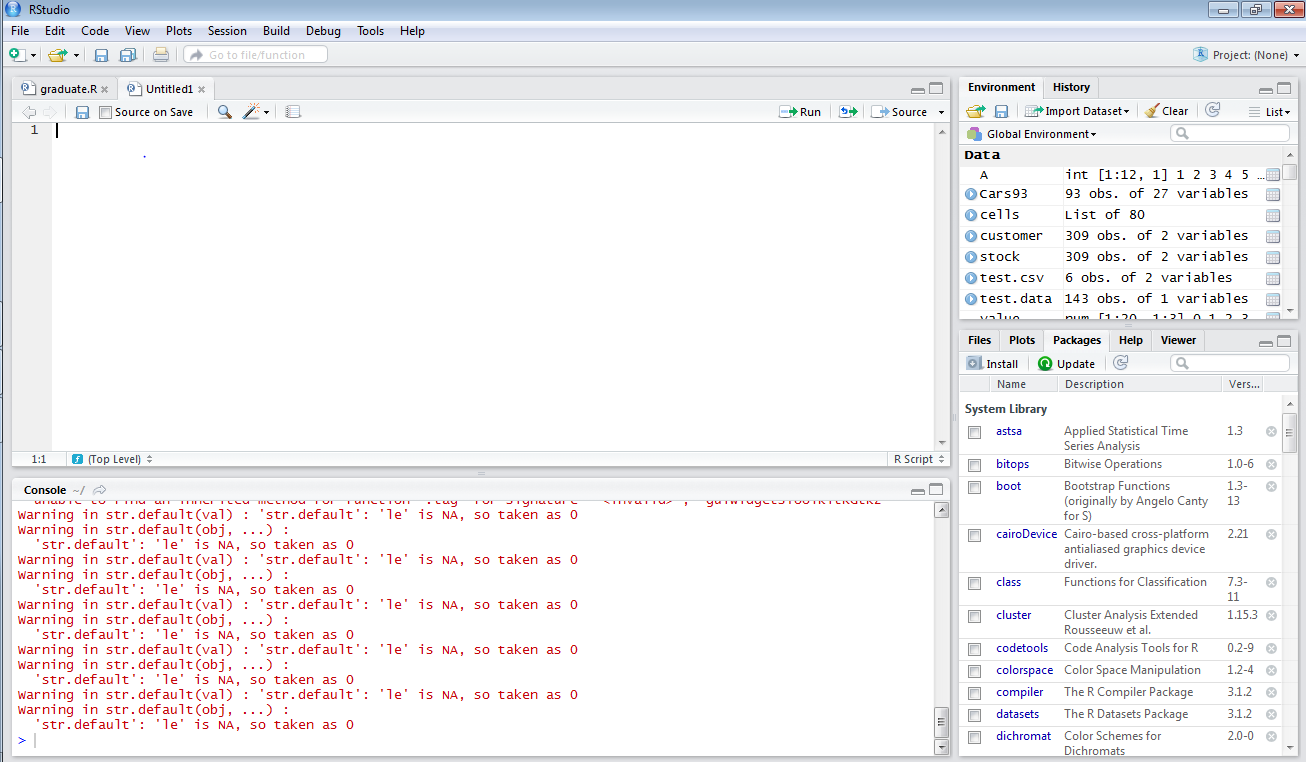
Truy cập vào trang web: <http://www.rstudio.com/products/rstudio/> để tải IDE R-studio

Việc cài đặt được thực hiện khá đơn giản

Sau khi hoàn tất việc cài đặt sẽ xuất hiện một icon:



sẽ xuất hiện trên desktop của máy tính. Đến đây thì chúng ta đã sẵn sàng sử dụng R có thể nhấp chuột vào icon này và chúng ta có một window như sau.



Giao diện của IDE R-studio gồm có:

* Source on save: nơi lưu trữ source code của bạn
* Console: nơi làm việc, R cho phép gõ từng dòng lệnh một
* Environment, history: nơi lưu trữ các biến trong quá trình code và lịc sử khi ta làm việc
* Packages: lưu trữ các packages quan trọng của R, muốn cài đặt chỉ cần ấn vào nút install
  1. **Tính toán dòng lệnh trong R**

R thường sử dụng dưới dạng “command line” có nghĩa là chúng ta phải gõ trực tiếp các lệnh vào prompt mầu xanh trên hình ở cửa sổ Console. Các lệnh phải tuân thử nghiêm ngặt các luật của ngôn ngữ R. Một câu lệnh sẽ được thực thi ngay sau khi nhấn phím Enter.

R phân biệt chữ hoa và chữ thường vd: library khác với Library. Một văn phạm khác nữa là khi có hai chữ rồi nhau, R thường dùng dấu chấm để thay khoảng trống, chẳng hạn như  
***data.frame, t.test, read.table*** *.*.. Điều này rất là quan trọng nếu không để ý sẽ làm mất thì giờ của người sử dụng.

Nếu lệnh gõ ra đúng “Văn phạm ” thì R sẽ cho chúng ta một cái prompt khác hay cho ra kết quả nào đó (tùy theo lệnh); nếu lệnh không đúng “Văn Phạm” thì R sẽ đưa ra một thông báo ngắn là không đúng hay không hiểu.

Ví dụ : khi chúng ta gõ.

> x <- rnorm(20)

>

thì R sẽ hiểu và cho chúng ta một cái prompt khác. Nhưng nếu chúng ta gõ lệnh sau :

*>*R is great

sẽ không hiểu và đưa ra một thông báo lỗi.

> Error: syntax error

Khi muốn rời khỏi R, chúng ta sẽ đơn giản nhấn nút (x) trên góc trái window hay gõ lệnh q()

* 1. **Văn phạm ngôn ngữ R**

Văn phạm chung của R là một lệnh (command) hay function. Mà đã là hàm thì phải có tham số; cho nên theo sau hàm là những tham số mà chúng ta phải cung cấp. chẳng hạn như:

> graph = plot.ts(data)

Để biết một hàm có những tham số nào, chúng ta dùng lệnh args(x), mà trong đó x là hàm mà chúng ta cần biết:

R là một ngôn ngữ “đối tượng ”. Điều này có nghĩa là các dữ liệu trong R được chứa trong object. Định hướng này cũng có vài ảnh hưởng đến cách viết câu R. Chẳng hạn như thay vì viết x = 5 như thông thường chúng ta vẫn viết, thì R yêu cầu viết x == 5.

Đối với R phép gán x = 5 tương đương với x <- 5. Cách viết sau (dùng kí hiệu <- ) được khuyến khích hơn là cách viết trước (=).

Một số kí hiệu hay dùng trong R:

|  |  |
| --- | --- |
| Kí hiệu | Ý nghĩa |
| x == y | x bằng y |
| x != y | x không bằng y |
| y < x | y nhỏ hơn x |
| x > y | x lớn hơn y |
| x <= y | x nhỏ hơn hoặc bằng y |
| x >= y | x lớn hơn hoặc bằng y |
| is.na(x) | có phải x là biến số missing |
| A & B | A và B |
| A | B | A hoặc B |
| ! | không là |

Với R thì tất cả các câu chữ hay lệnh sau kí hiệu # đều không có hiệu ứng, vì # là kí hiệu dành cho người sử dụng thêm vào các ghi chú.

* 1. **Cách đặt tên trong R**

Đặt tên một đối tượng hay một biến số trong R khá linh hoạt, vì R không có nhiều giới hạn như trong các phần mền khác. Tên một đối tượng phải được viết liền nhau. Chẳng hạn như R chấp nhận ***myobject*** nhưng không chấp nhận ***my object*** .

*> myobject <- rnorm(10)*  
*> my object <- rnorm(10)*  
*Error: syntax error in “my object”*

Nhưng đôi khi tên myobject khó đọc cho nên chúng ta nên tách rời bằng “.” Như my.object.  
*> my.object <- rnorm(10)*

Một điều quan trọng cần lưu ý là R phân biệt mẫu kí tự viết hoa và viết thường. Cho nên ***My.object*** khác so với ***my.object***. Ví dụ:

> My.object.u <- 15

> my.object.L <- 5

> My.object.u + my.object.L

> [1] 20

Một vài lưu ý khi đặt tên trong R:

* Không nên đặt tên một biến số bằng kí hiệu “\_” như ***my\_object*** hay ***my-object.***
* Không nên đặt tên một object giống như một biến số trong một dữ liệu. Ví dụ: nếu chúng ta có một data.frame với biến số ***age*** trong đó, thì chúng ta có một đối tượng trùng tên với ***age***, tức là không nên viết ***age <- age***. Tuy nhiên, nếu ***data.frame*** tên là data thì chúng ta có thể đề cập đến biến số ***age*** với mọt kí tự ***$*** như sau: ***data$age***. (tức là biến số age trong  
  ***data.frame*** *data*), và trong trường hợp đó, age <- data$age có thể chấp nhận được.
  1. **Trợ giúp trong R**

Ngoài lệnh ***args()*** R còn cung cấp lệnh ***help()*** để người sử dụng có thể hiểu “Văn phạm” của từng hàm. Chẳng hạn như muốn biết hàm ***lm*** có những tham số gì chúng ta chỉ cần gõ lệnh:  
*> help()* hay  
*> ?lm*

một cửa sổ sẽ hiện ra bên ngoài của màn hình chỉ rõ cách sử dụng ra sao và thậm chí có cả ví du. Sử dụng lệnh *help.start()* một cửa sổ sẽ xuất hiện chỉ dẫn toàn bộ hệ thống R.

Hàm ***apropos*** cũng rất có ích vì nó cung cấp cho chúng ta tất cả các hàm trong R bắt đầu bằng kí tự mà chúng ta muốn tìm. Chẳng hạn như chúng ta muốn biết hàm nào trong R có kí tự “lm” thì chỉ gõ lệnh:  
*> apropos(lm) .*

* 1. **Môi trường vận hành**

Dữ liệu phải được chứa trong một khu vực(directory) của máy tính. Trước khi sử dụng R, có lẽ cách hay nhất là tạo ra một directory để chứa dữ liệu, chẳng hạn như: c:\works\stats. Để R biết dữ liệu nằm ở đâu ta sử dụng lệnh setwd(set working directory) như sau:

> setwd(“c/works/stats”)

Lệnh trên báo cho R biết là dữ liệu sẽ chứa trong directory có tên là: c:\works\stats. Chú ý rằng, R dùng forward slash “/” chứ không phải backward slash “\” như trong hệ thống windows.

1. **R và lập trình GUI**

R cung cấp một số packages chuyên dùng cho lập trình GUI: gWidgets. Các gói gWidgets tạo ra một API để tương tác với bộ công cụ GUI. Ban đầu nó được xây dựng bởi lập trình viên SimSon Urbaner với tên gọi iWidgets, sau đó với những cải tiến được đề xuất bởi Philippe Grosjean , Michael Lawrence , Simon Urbanek và John Verzani .

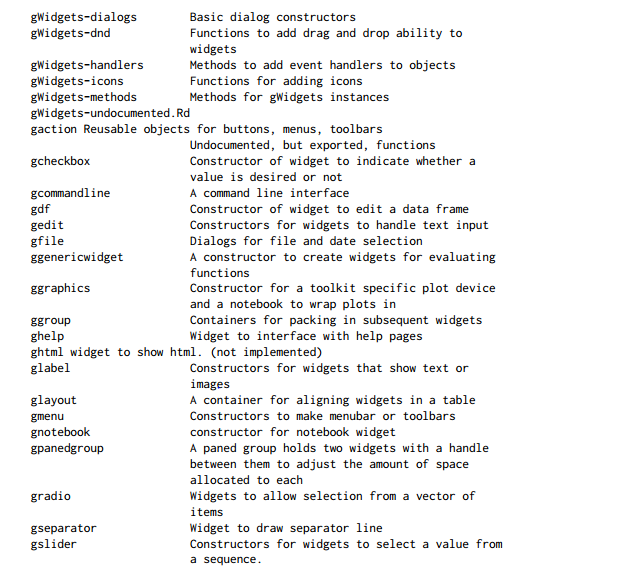
Muốn cài đặt packages gWidgets, bạn gõ vào Console dòng lệnh:

*> install.packages("gWidgetsRGtk2", dep = TRUE)*

Sau khi đã cài đặt packages gWidgets, ta gọi 2 lênh sau để sử dụng packages này:

*> require(gWidgets)*  
*> ##options("guiToolkit"="RGtk2"*  
*> options("guiToolkit"="tcltk")*

Sau đây là một số nội dung trong packages gWidgets



**Chương IV: Cài đặt và thử nghiệm mô hình**

* 1. **Mô tả bài toán**

Giới thiệu bài toán

Ra đời vào đầu năm 2000, thị trường chứng khoán Việt Nam đã trở thành một kênh đầu tư hết sức hấp dẫn đối với các nhà đầu tư, từ các tổ chức đầu tư chuyên nghiệp cho đến các nhà đầu tư cá nhân nghiệp dư nhỏ lẻ. Tuy nhiên, bên cạnh mức sinh lợi cao, đây cũng là hoạt động luôn tồn tại nhiều rủi ro tiềm ẩn bởi nhà đầu tư không phải lúc nào cũng dự đoán được chính xác xu hướng của giá cổ phiếu trong tương lai. Do đó, việc dự báo chính xác sự biến động giá của cổ phiếu để có một sách lược nhằm phục vụ cho công việc kinh doanh của các cá nhân, tổ chức hay hoạch định chiến lược của một quốc gia đã thu hút rất nhiều sự quan tâm của các nhà kinh tế lượng tài chính trong và ngoài nước. Tại thị trường Việt Nam, sự biến động của chỉ số VnIndex phản ánh rủi ro hệ thống, vì vậy, việc dự báo được sự tăng giảm của Vn-Index cũng đồng thời giúp các nhà đầu tư nhận biết chiều hướng biến động giá của các cổ phiếu trên thị trường này.

Trong khuôn khổ đề tài, chúng tôi đề xuất sử dụng mô hình ARIMA và phương pháp Box-jenkins để dự báo chỉ số VnIndex trong ngắn hạn căn cứ vào chuỗi dữ liệu quá khứ. George Box và Gwilym Jenkins (1976) đã nghiên cứu mô hình ARIMA (**Autoregressive Integrated Moving Average** - *Tự hồi qui tích hợp Trung bình trượt*), và tên của họ thường được dùng để gọi tên các quá trình ARIMA tổng quát, áp dụng vào việc phân tích và dự báo các chuỗi thời gian. Phương pháp Box-Jenkins với bốn bước: nhận dạng mô hình thử nghiệm, ước lượng, kiểm định bằng chuẩn đoán, và dự báo.

1. **Xây dựng mô hình ARIMA cho VnIndex**
   1. **Giới thiệu về số liệu**

**2.2. Cơ sở lý luận**

Dữ liệu chúng ta sử dụng là dữ liệu chuỗi thời gian. Đặc điểm chính để phân biệt giữ dữ liệu có phải là thời gian thực hay không đó chính là sự tồn tại của cột thời gian được đính kèm trong đối tượng quan sát. Nói cách khác, dữ liệu thời gian thực là một chuỗi các giá tị quan sát của biến Y:

với là giá trị của biến Y tại thời điểm t

Mục đích của việc phân tích chuỗi thời gian thực là thu được một mô hình dựa trên các giá trị quá khứ của biến quan sát cho phép ta dự đoán giá trị của biến Y trong tương lai, tức là có thể dự đoán được các giá trị

Mô hình tự hồi quy p – AR(p): trong mô hình tự hồi quy quá trình phụ thuộc vào tổng trọng số của các giá trị quá khứ và số hạng nhiễu ngẫu nhiên

Mô hình trung bình trượt q – MA(q): trong mô hình trung bình trượt, quá trình được mô tả hoàn toàn bằng tổng trọng số của các ngẫu nhiên hiện hành có độ trễ:

Mô hình hồi quy kết hợp với trung bình trượt ARIMA(p, d, q):

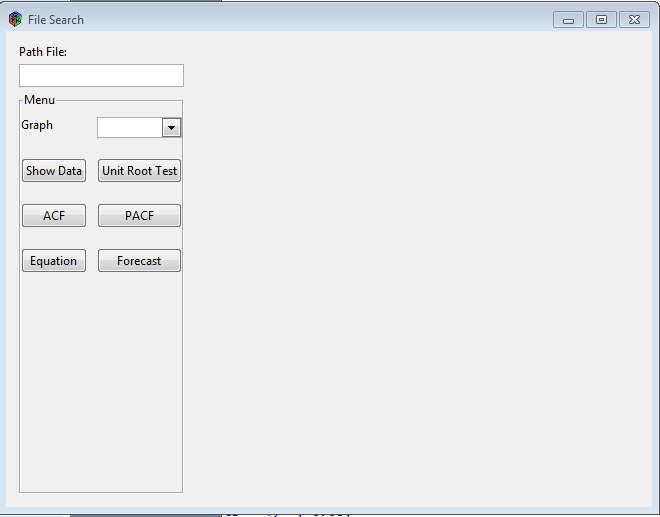
**2.3. Mô hình ARIMA cho bài toán dự báo giá cổ phiếu**

Để áp dụng mô hình ARIMA vào bài toán tài chính, ta xây dựng mô hình dự bảo. Mô hình gồm 3 quá trình:

* Xác định mô hình: với đầu vào là tập dữ liệu chuỗi thời gian trong tài chính giúp cho việc xác định ban đầu các thành phân mô hình p, d, q
* Ước lượng, kiểm tra: mô hình ARIMA là phương pháp lặp, sau khi xác định các thành phần, mô hình sẽ ước lượng các tham số, sau đó kiểm tra độ chính xác của mô hình: nếu hợp lý, tiếp bước sau, nếu không hợp lý, quay trở lại bước xác định
* Dự báo: sau khi đã xác định các tham số, mô hình sẽ đưa ra dự báo cho ngày tiếp theo

**2.4.Áp dụng**

**2.4.1. Giao diện chương trình**



**Các modun chức năng chính của chương trình**

* Path file: đường dẫn đến file dữ liệu, file dữ liệu có dạng .csv
* Graph: có ba bậc: bậc 0, bậc 1, bậc 2. Ba bậc này tương ứng với đồ thị của dữ liệu ban đầu, dữ liệu đã sai phân bậc 1 và dữ liệu đã sai phân bậc 2
* Show data: show dữ liệu file.csv mà ta muốn phân tích
* Unit Root Test: kiểm nghiệm đơn vị Dickey-Fuller
* ACF: hiển thị số liệu tính toán và đồ thị hàm ACF
* PACF: hiển thị số liệu tính toán và đồ thị hàm PACF
* Equation: nhập giá trị bằng tay các giá trị p, d và q và hiển thị các kết quả kiểm định: BIC, SEE,
* Forecast: dự báo kết quả và đưa ra đồ thị

**Cài đặt một số chức năng chính**

**//**Menu chính

|  |
| --- |
|  |

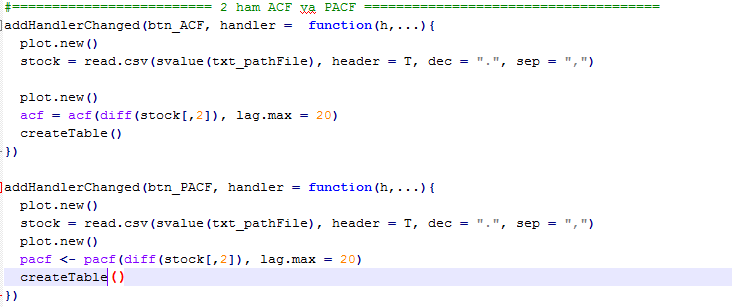
//show đồ thị

|  |
| --- |
|  |

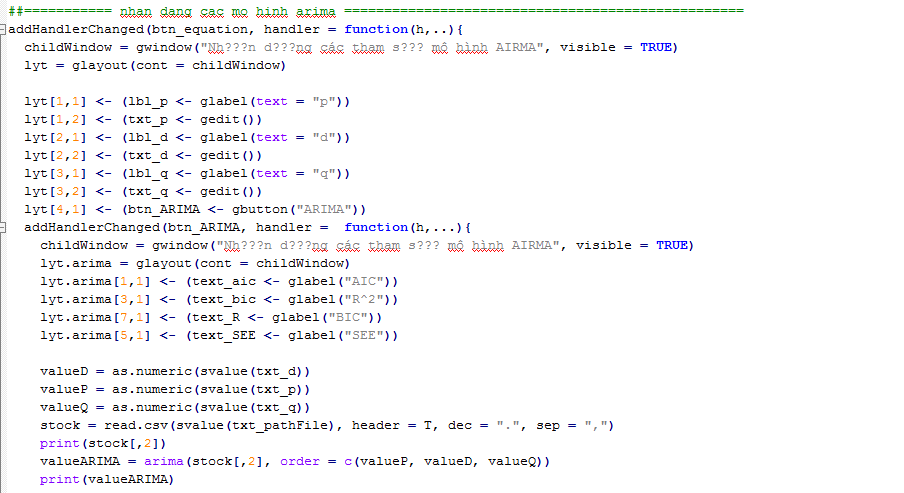
//show dữ liệu file.csv

|  |
| --- |
|  |

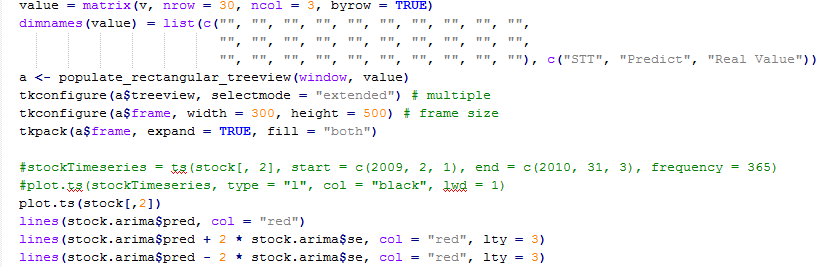
//hiển thị kết quả hàm ACF và PACF cùng đồ thị



//Nhận dạng các mô hình ARIMA



//Hiển thị dự báo



**2.4.2. Dữ liệu**

* Nguồn cập nhật số liệu là trang web cophieu68.com. Đây là trang web chuyên cung cấp số liệu về thị trường chứng khoán Việt Nam
* Số liệu VnIndex được lấy từ ngày 2/1/2009 tới ngày 31/3/2010 bao gồm 310 quan sát.
* Dữ liệu đầu vào là file.csv được lấy từ website xuống. Dữ liệu ở đây có dạng như sau:

|  |  |
| --- | --- |
| Year | Value |
| |  | | --- | | 20090102 | | 20090105 | | 20090106 | | 20090107 | | 20090108 | | 20090109 | | 20090112 | | 20090113  ……………..  ………………   |  | | --- | | 20100325 | | 20100326 | | 20100329 | | 20100330 | | 20100331 | | | |  | | --- | | 312.5 | | 308.8 | | 312.9 | | 320.4 | | 315.9 | | 313.8 | | 311.2 | | 308.6 |   ………  …………   |  | | --- | | 507.9 | | 505.5 | | 506.7 | | 504.7 | | 501.9 | |

**2.4.3. Kiểm tra tính dừng của chuỗi dữ liệu**

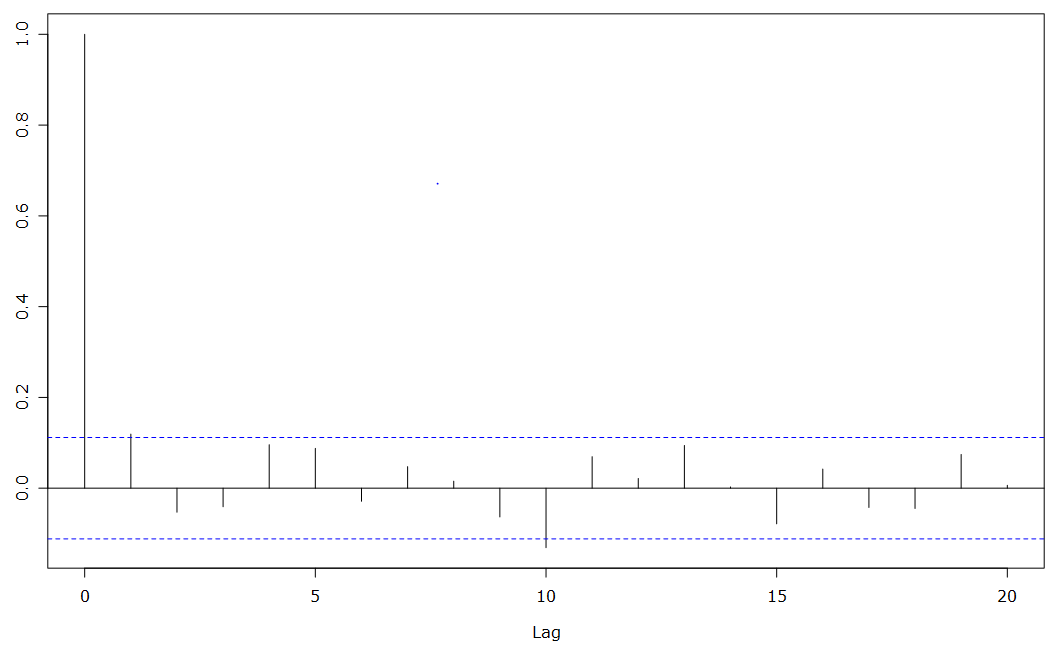
**2.4.4. Nhận dạng mô hình**

Trong phần này, chúng ta tính được giá trị của hàm ACF và của hàm PACF. Sau đó tiến hành vẽ đồ thị và rút ra một vài nhận xét

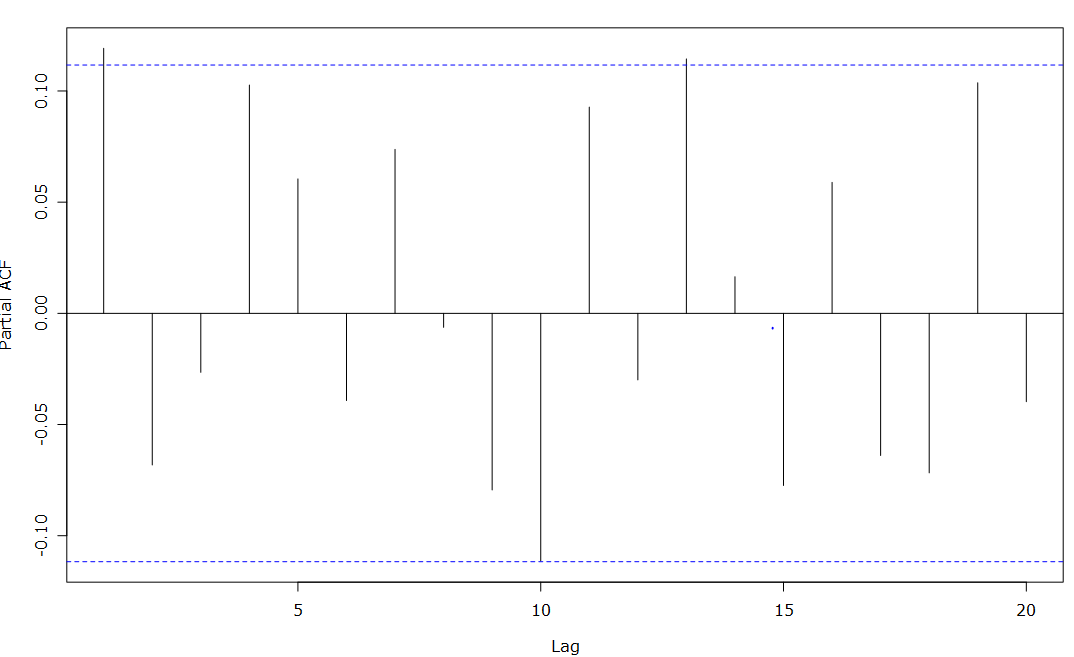
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| LAG | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| ACF | 1 | 0.12 | -0.05 | -0.04 | 0.1 | 0.09 | -0.03 | 0.05 | 0.02 | -0.06 | -0.13 |
| PACF | NA | 0.12 | -0.07 | -0.03 | 0.1 | 0.06 | -0.04 | 0.07 | -0.01 | -0.08 | -0.11 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| LAG | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| ACF | 0.07 | 0.02 | 0.09 | 0.00 | -0.08 | 0.04 | -0.04 | -0.04 | 0.07 | 0.01 |
| PACF | 0.09 | -0.03 | 0.11 | 0.02 | -0.08 | 0.06 | -0.06 | -0.07 | 0.1 | -0.04 |

Đồ thị của hàm ACF



Đồ thị của hàm PACF



**Nhận dạng**