TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - CƠ - TIN HỌC

BÁO CÁO CUỐI KÌ

PHÂN TÍCH CHUỖI THỜI GIAN

Đề tài:

DỰ ĐOÁN DOANH SỐ XE Ô TÔ TRONG 10 NĂM TỚI

Giảng viên: HOÀNG THỊ PHƯƠNG THẢO
NGUYỄN BẢO NGỌC

Sinh viên thực hiện: VŨ THUỲ TRANG 20002094
VƯƠNG THUỲ DƯƠNG 20002039

Hà Nội, 11-2023

LỜI NÓI ĐẦU

Trong bối cảnh ngày nay, sự phát triển của ngành công nghiệp và sự biến động liên tục đang đặt ra những thách thức và cơ hội mới. Đặc biệt, việc hiểu rõ và dự đoán xu hướng, biến động trong các dữ liệu theo thời gian trở thành một yếu tố quyết định quan trọng. Chính vì thế dữ liệu chuỗi thời gian đóng một vai trò cực kỳ quan trọng đối với sự phát triển của nhân loại.

Chuỗi thời gian, là một loại dữ liệu theo dõi sự biến động của các điểm dữ liệu đã chọn, không chỉ giới hạn trong lĩnh vực tài chính như giá chứng khoán mà còn có ứng dụng rộng rãi trong nhiều ngành công nghiệp khác nhau. Ví dụ như quan sát sóng điện não, đo lượng mưa, dự báo giá cổ phiếu, theo dõi doanh số bán lẻ hàng năm, người đăng ký hàng tháng, hay thậm chí là theo dõi nhịp tim mỗi phút.

Thông qua việc áp dụng phương pháp này, chúng ta có thể không chỉ nhận biết xu hướng và biến động mà còn dự đoán được những diễn biến trong tương lai. Chúng em hy vọng rằng việc khám phá về chuỗi thời gian sẽ giúp bạn có cái nhìn toàn diện hơn về cách phân tích và áp dụng dữ liệu thời gian, tạo ra những cơ hội mới và chiến lược thông minh trong quản lý và kế hoạch chiến lược của doanh nghiệp.

Với sự gia tăng đáng kể trong sự cạnh tranh và sự thay đổi nhanh chóng về ưu tiên của người tiêu dùng, khả năng dự đoán chính xác doanh số bán xe ô tô không chỉ là một thách thức mà còn là cơ hội lớn. Dự báo chính xác doanh số bán hàng không chỉ là một số liệu, mà còn là công cụ quan trọng hỗ trợ quyết định và định hình chiến lược kinh doanh. Từ đó, ta có thể đưa những chiến lược kinh doanh linh hoạt và phù hợp với nhu cầu của thị trường. Không chỉ thế, điều này còn giúp tối ưu hóa quản lý tài chính và nguồn lực, giảm thiểu rủi ro tài chính và chi phí lưu kho.

Trong bài báo cáo này, chúng em sẽ thực hiện dự đoán danh số bán xe ô tô tại Canada trong vòng 10 năm tới bằng các phương pháp phân tích chuỗi thời gian. Bài báo cáo bao gồm 4 phần:

Phần 1: Tổng quan về đề tài nghiên cứu

Phần 2: Đặt vấn đề, giới thiệu về dữ liệu và cách tiền xử lý dữ liệu

Phần 3: Cơ sở lý thuyết

Phần 4: Kết quả thực nghiệm và đánh giá

Contents

1 Tổng quan về đề tài nghiên cứu

Trong bối cảnh ngày nay, ngành công nghiệp ô tô không ngừng phát triển và biến động, tạo nên một thị trường đầy thách thức và cơ hội đối với các doanh nghiệp. Đặc biệt, khả năng dự đoán doanh số bán xe ô tô trở thành một yếu tố quyết định trong việc xây dựng chiến lược kinh doanh và duy trì sự cạnh tranh. Đề tài nghiên cứu "Dự Đoán Doanh Số Xe Ô Tô Trong 10 Năm Tới" nhằm tìm hiểu và áp dụng các phương pháp phân tích chuỗi thời gian để đưa ra những dự báo chính xác về xu hướng thị trường ô tô trong thập kỷ tới.

Mục tiêu của đề tài là xây dựng một mô hình dự đoán doanh số bán xe ô tô dựa trên phân tích chuỗi thời gian, từ đó đưa ra những nhận định chính xác về sự biến động của thị trường. Việc này không chỉ giúp doanh nghiệp ô tô nắm bắt xu hướng và điều chỉnh chiến lược kinh doanh một cách linh hoạt mà còn hỗ trợ quản lý nguồn lực và sản xuất một cách hiệu quả.

Nghiên cứu tập trung vào việc việc phân tích biểu đồ ACF (Autocorrelation Function) và PACF (Partial Autocorrelation Function) để hiểu rõ mức độ tương quan giữa các quan sát trong chuỗi thời gian, là cơ sở quan trọng cho quá trình xây dựng mô hình.

Tiếp theo, xác định tính mùa vụ để hiểu rõ các yếu tố ảnh hưởng theo chu kỳ trong dữ liệu. Mô hình mùa vụ giúp chúng ta cập nhật và tính toán độ biến động của dữ liệu theo thời gian, cung cấp thông tin cần thiết cho quá trình dự đoán. Sau đó thực hiện các biện pháp chuyển đổi và hiệu chỉnh dữ liệu để làm cho dữ liệu trở nên ổn định và phù hợp với mô hình.

Cuối cùng quá trình là xây dựng mô hình. Điều này bao gồm việc ước lượng tham số để mô tả mối quan hệ giữa các quan sát trong chuỗi thời gian. Quá trình này đòi hỏi sự cân nhắc và tối ưu hóa để tìm ra mô hình phù hợp nhất với dữ liệu. Sau đó, phân tích và kiểm định mô hình là quy trình quan trọng để đảm bảo rằng mô hình đáp ứng chính xác và đáng tin cậy đối với dữ liệu thực tế. Các phương pháp đánh giá mô hình, như phân tích phần dư và kiểm định giả thuyết, được sử dụng để đánh giá hiệu suất và tính toàn vẹn của mô hình dự đoán.

2 Đặt vấn đề, giới thiệu về dữ liệu và cách tiền xử lý dữ liệu

2.1 Giới thiệu về dữ liệu

Bộ dữ liệu bán hàng ô tô hàng tháng (Monthly Cars Sales) mô tả doanh số bán hàng ô tô tại Thành phố Québec, Canada. Như chúng ta biết, Canada có mùa đông cao điểm từ tháng 11 đến tháng 2 năm sau và mùa hè bắt đầu từ tháng 4 hoặc tháng 5. Hầu hết người mua thường mua ô tô vào mùa hè vì mùa đông có tuyết lạnh, chính quyền hay ban bố không cho phép người dân đi ra ngoài. Và đây cũng chính là kết quả dự đoán của chúng em sau khi sử dụng các model với kết quả được trực quan hoá bằng đồ thị trong các phần sau.

Bộ dữ liệu Monthly Cars Sales (Datasets/monthly-car-sales.csv) được thu thập từ năm 1960-1968 với nguồn dữ liệu được trích xuất ở Time Series Data Library (citing: Abraham & Ledolter (1983)). Dữ liệu bao gồm 108 bản ghi với 3 trường thuộc tính: Year (Năm), Month (Tháng) và Sales (Doanh số bán hàng). Tất cả dữ liệu đều ở dạng số nguyên (interger) và đây là dạng dữ liệu thô, chưa phải là dữ liệu theo thời gian.

	Year <int>	Month <int>	Sales <int>
1	1960	1	6550
2	1960	2	8728
3	1960	3	12026
4	1960	4	14395
5	1960	5	14587
6	1960	6	13791

6 rows

Figure 1: Dữ liệu với 6 bản ghi đầu tiên

2.2 Đặt vấn đề

Với mục đích là dự đoán doanh số xe ô tô, kết hợp với đặc điểm của dữ liệu và hiện trạng thực tế, chúng em sẽ tiếp cận bài toán theo chuỗi thời gian (mùa), cụ thể gồm các bước như sau:

- Model specification (Xác định mô hình)
- Residual approach (Tiếp cận dựa trên dư thừa)
- Model fitting (Gắn mô hình)
- Model diagnostics (Chẩn đoán mô hình)

- Dự đoán cho 10 năm tiếp theo dựa trên mô hình đã xây dựng

2.3 Tiền xử lý dữ liệu

Như đã phân tích ở trên, dữ liệu có hiện thời vẫn là dữ liệu thô (dạng bảng), chưa phải dữ liệu theo thời gian. Do đó, bước tiền xử lý quan trọng nhất chính là biến đổi dữ liệu về dạng dữ liệu theo chuỗi thời gian.

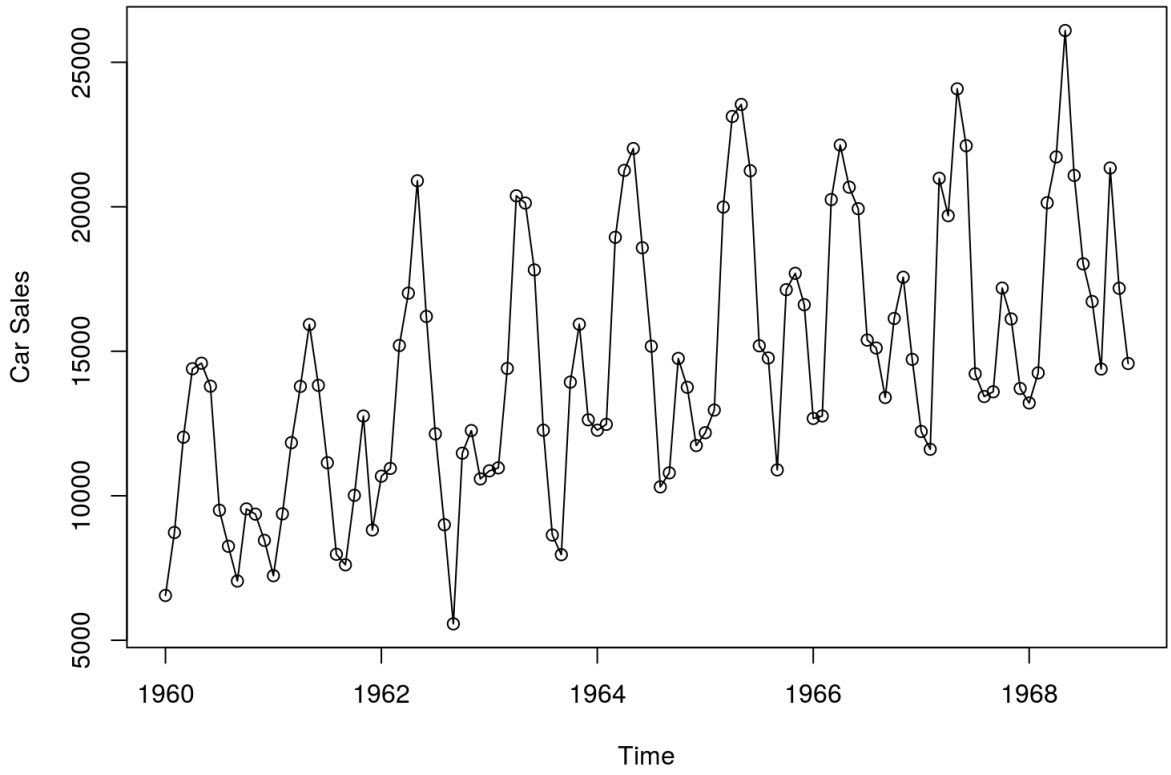


Figure 2: Đồ thị chuỗi thời gian về doanh số bán hàng ô tô hàng tháng 1960 - 1968.

Nhận xét về đồ thị 2, chúng em đưa ra được một số kết luận sơ bộ về xu hướng và mùa ảnh hưởng đến doanh số xe ô tô như sau:

- Xu hướng: Có một xu hướng tăng mạnh từ năm 1960 - 1968
- Mùa: Sau một khoảng thời gian đều đặn, dữ liệu đều có sự lên xuống theo chu kỳ. Đây chính là tác động theo mùa mà chúng em đã phân tích ở bên trên: mùa hè có xu hướng mua xe nhiều hơn so với mùa đông

Đồng thời, chúng ta thấy rằng ở đây là dữ liệu không dừng, không ổn định, nên ở những bước tiếp theo phải đặc biệt lưu ý đến tính chất của dữ liệu này.

3 Cơ sở lý thuyết

3.1 ACF - Autocorrelation Function

ACF là viết tắt của "Autocorrelation Function" (Hàm tự tương quan mẫu). ACF là một công cụ quan trọng để đo lường mức độ tương quan giữa các giá trị của chuỗi thời gian với chính nó tại các điểm thời gian khác nhau.

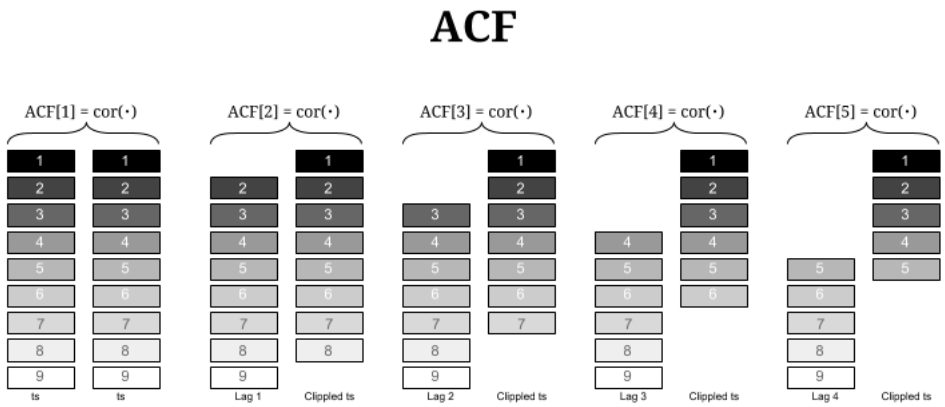


Figure 3: Mô hình minh hoạ cho ACF

ACF được xác định bởi công thức như sau:

$$R(k) = \frac{\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2} \tag{1}$$

Trong đó:

- $R(k)$ là giá trị của ACF tại khoảng cách thời gian k.
- T là chiều dài của chuỗi thời gian.
- y_t là giá trị tại thời điểm t.
- $\bar{y}$ là giá trị trung bình của chuỗi thời gian.

Hàm Tự tương quan (ACF) trong phân tích chuỗi thời gian có ý nghĩa quan trọng trong việc hiểu cấu trúc của dữ liệu thời gian. Điển hình có thể kể đến như là:

- Phát hiện mô hình chuỗi Thời gian: ACF thường được sử dụng để phát hiện mô hình chuỗi thời gian phù hợp. Các đỉnh và đáy của đồ thị ACF có thể chỉ ra các mức độ tương quan và không tương quan tại các khoảng cách thời gian khác nhau, cung cấp thông tin về lực lượng dự báo của mô hình.

- Xác định chu kỳ: Nếu chuỗi thời gian có yếu tố chu kỳ, ACF có thể giúp xác định độ lớn và độ dài của chu kỳ đó. Các đỉnh địa phương trên đồ thị ACF có thể chỉ ra sự lặp lại chu kỳ trong dữ liệu.
- Kiểm tra nguyên tắc: ACF có thể được sử dụng để kiểm tra xem có sự tự tương quan nào đó còn lại trong dữ liệu sau khi loại bỏ xu hướng. Điều này quan trọng trong việc xác định xem liệu mô hình có cần thêm yếu tố nào đó để mô tả dữ liệu một cách chính xác.
- Dự báo chuỗi thời gian: ACF cung cấp thông tin quan trọng để xác định các tham số của mô hình dự báo chuỗi thời gian, như mô hình ARIMA (Autoregressive Integrated Moving Average). Các thông số này giúp dự đoán giá trị tương lai của chuỗi thời gian.

3.2 PACF - Partial Autocorrelation Function

PACF là viết tắt của "Partial Autocorrelation Function" (Hàm Tự tương quan Riêng). Nó là một công cụ trong phân tích chuỗi thời gian, tương tự như ACF (Hàm Tự tương quan), nhưng PACF tập trung vào mức độ tương quan giữa các giá trị của chuỗi thời gian ở các khoảng cách thời gian cụ thể, giả định rằng tất cả các khoảng cách thời gian lớn hơn đều đã được loại bỏ.

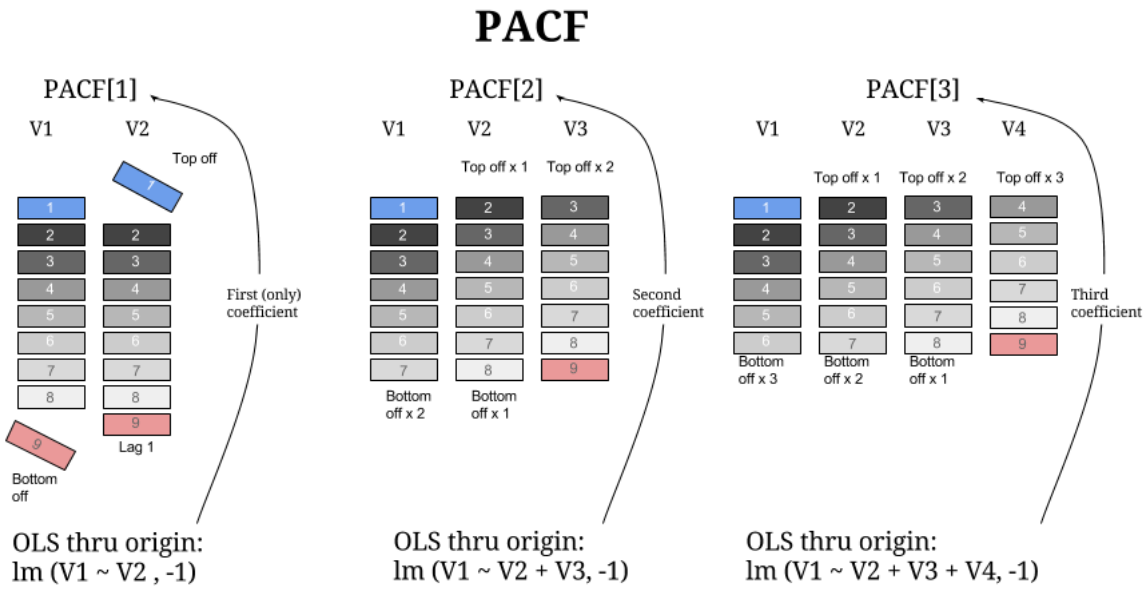


Figure 4: Mô hình minh hoạ cho PACF

PACF được xác định bởi công thức sau:

$$\phi_{kk} = \frac{\gamma_{kk}}{\gamma_{00}} \tag{2}$$

Trong đó:

- ϕ_{kk} là giá trị PACF tại khoảng cách thời gian k.
- γ_{kk} là giá trị tự tương quan (autocorrelation) tại lag k của chuỗi thời gian đã điều chỉnh cho ảnh hưởng của các lag nhỏ hơn.
- γ_{00} là giá trị tự tương quan tại lag 0 của chuỗi thời gian đã điều chỉnh cho ảnh hưởng của các lag nhỏ hơn.

Hàm Tự tương quan Riêng (PACF - Partial Autocorrelation Function) có ý nghĩa quan trọng trong phân tích chuỗi thời gian và xác định mô hình AR (Autoregressive). Cụ thể:

- Xác định bậc của mô hình AR: PACF thường được sử dụng để xác định cấp của mô hình AR. Cấp của mô hình là số lượng lag (độ trễ) mà giữ liệu tại thời điểm hiện tại tương quan với giữ liệu tại thời điểm trước đó. Khi PACF "tắt" sau một số lag cụ thể, điều này cho biết rằng chỉ có các lag đó ảnh hưởng đáng kể đến dữ liệu hiện tại và cấp của mô hình AR được xác định bởi số lượng lag mà PACF giữ lại.
- Lựa chọn tham số cho mô hình ARIMA: Trong quá trình xây dựng mô hình ARIMA (Autoregressive Integrated Moving Average), PACF cung cấp thông tin để lựa chọn tham số của thành phần tự hồi quy (AR). Khi xem xét đồ thị PACF, các đỉnh đầu tiên mà giảm về 0 thường là các lag quan trọng để xác định cấp của mô hình AR.
- Hiểu cấu trúc chuỗi thời gian: PACF giúp hiểu cấu trúc của chuỗi thời gian và xác định mối quan hệ tương quan giữa các giá trị tại các khoảng cách thời gian cụ thể. Các giá trị PACF giúp xác định xem liệu có sự tương quan riêng nào đặc biệt ngoài các lag gần đây hay không.

3.3 EACF - Extended Autocorrelation Function

EACF, hay Extended Autocorrelation Function, là một công cụ được sử dụng trong quá trình xác định các tham số cho mô hình ARIMA (Autoregressive Integrated Moving Average). EACF được thiết kế để giúp xác định cấp (order) của thành phần tự hồi quy (AR) và thành phần trung bình di chuyển (MA) trong mô hình ARIMA.

Quá trình sử dụng EACF thường bắt đầu bằng việc xây dựng một bảng, trong đó các hệ số AR và MA được liệt kê. EACF được tính toán cho mỗi giá trị của cả hai hệ số, và các giá trị được so sánh với một ngưỡng cụ thể, thường là giá trị tuyệt đối của 2 lần độ lệch chuẩn của EACF.

Các ô trong bảng EACF có giá trị nhỏ hơn ngưỡng được coi là quan trọng và có thể giúp xác định thứ tự tối ưu cho mô hình ARIMA. Nếu giá trị trong một ô là nhỏ hơn ngưỡng, nó có thể gợi ý rằng việc sử dụng một lag nhất định cho thành phần tương ứng là lựa chọn hợp lý.

Tuy EACF không phải là công cụ phổ biến như ACF và PACF, nhưng nó có thể cung cấp thông tin hữu ích trong quá trình xây dựng mô hình ARIMA, đặc biệt là khi cần xác định cả hai thành phần AR và MA.

3.4 AR - Autoregressive

AR là viết tắt của "Autoregressive" (tự hồi quy) mô tả mối quan hệ tương quan giữa giá trị tại một thời điểm cụ thể và giá trị tại các thời điểm trước đó (được gọi là các lags). Cụ thể, một mô hình AR(p) (tự hồi quy bậc p) sử dụng p giá trị lag gần nhất để dự đoán giá trị hiện tại.

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \varepsilon_t \tag{3}$$

Trong đó:

- X_t là giá trị của chuỗi thời gian tại thời điểm t.
- c là hệ số chệch (intercept).
- $\phi_1, \phi_2, ..., \phi_p$ là các hệ số tự hồi quy.
- $X_{t-1}, X_{t-2}, ..., X_{t-p}$ là giá trị của chuỗi tại các thời điểm lag.
- ε_t là sai số ngẫu nhiên tại thời điểm t.

Đánh giá mô hình AR, chúng em có những nhận xét và lưu ý sau:

- Về dữ liệu: Kiểm tra chuỗi thời gian: Đảm bảo rằng chuỗi thời gian của bạn thỏa mãn các điều kiện cần thiết cho mô hình AR, chẳng hạn như tính ổn định và tính đồng nhất.
- Về số lag của mô hình: Phân tích hàm tự tương quan (ACF) và hàm tự tương quan riêng (PACF) để xác định cấp của mô hình AR. Các đỉnh trên đồ thị PACF thường cho biết số lượng lag cần chọn cho mô hình AR.
- Về các xác định bậc của mô hình: Việc xác định mô hình AR thường được thực hiện tốt nhất với PACF. Do đối với mô hình AR, PACF lý thuyết "tắt" theo thứ tự của mô hình. Cụm từ "tắt" có nghĩa là về mặt lý thuyết, tự tương quan một phần bằng 0 ngoài điểm đó. Nói cách khác, số lượng tự tương quan một phần khác 0 mang lại thứ tự cho mô hình AR. Theo "thứ tự của mô hình", chúng tôi muốn nói đến độ trễ cực đại nhất của x được sử dụng làm yếu tố dự đoán.

3.5 MA - Moving Average

MA là viết tắt của "Moving Average" (Trung bình trượt). Trong mô hình MA, giá trị tại một thời điểm cụ thể được dự đoán dựa trên các sai số ngẫu nhiên từ các thời điểm trước đó. Mô hình MA(q) (Moving Average bậc q) sử dụng q sai số ngẫu nhiên gần nhất để dự đoán giá trị hiện tại.

Qúa trình trung bình trượt được hiểu là quá trình dịch chuyển hoặc thay đổi giá trị trung bình của chuỗi theo thời gian. Do chuỗi của chúng ta được giả định là dừng nên quá trình thay đổi trung bình dường như là một chuỗi nhiễu trắng. Qúa trình moving average sẽ tìm mối liên hệ về mặt tuyến tính giữa các phần tử ngẫu nhiên ε_t (stochastic term). Chuỗi này phải là một chuỗi nhiễu trắng thỏa mãn các tính chất:

$$\begin{cases} \mathrm{E}(\varepsilon_t) & = 0 \\ \sigma(\varepsilon_t) & = \alpha \\ \rho(\varepsilon_t, \varepsilon_{t-s}) & = 0, \forall s \leq t \end{cases} \tag{4}$$

Công thức tổng quát của một mô hình MA(q) là:

$$X_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + ... + \theta_q \varepsilon_{t-q} \tag{5}$$

Trong đó:

- X_t là giá trị của chuỗi thời gian tại thời điểm t.
- c là hệ số chệch (intercept).
- $\theta_1, \theta_2, ..., \phi_q$ là các hệ số trọng số của sai số ngẫu nhiên tại các thời điểm trước đó (lags).
- ε_t là sai số ngẫu nhiên tại thời điểm t.

Mô hình MA giúp mô tả và dự đoán các biến động ngẫu nhiên trong chuỗi thời gian. Đánh giá về mô hình MA, chúng em có những nhận xét và lưu ý sau:

- Về dữ liệu: Kiểm tra chuỗi thời gian: Đảm bảo rằng chuỗi thời gian của bạn thỏa mãn các điều kiện cần thiết cho mô hình AR, chẳng hạn như tính ổn định và tính đồng nhất. Điều này giống với mô hình AR.
- Về số lag của mô hình: Phân tích hàm tự tương quan (ACF) để xác định cấp của mô hình MA. Các đỉnh đầu tiên mà giảm về 0 thường là số lượng lag cần chọn cho mô hình MA.

- Về các xác định bậc của mô hình: Việc xác định mô hình MA thường được thực hiện tốt nhất với PACF. Do đối với mô hình MA, hàm PACF lý thuyết không "tắt" mà thay vào đó giảm dần về 0 theo một cách nào đó. Một mô hình MA thường có một mẫu rõ ràng hơn trong ACF. ACF chỉ sẽ có các tương quan riêng phần khác không tại các độ trễ có liên quan đến mô hình.

3.6 ARMA - Autoregressive Moving Average

ARMA là viết tắt của "Autoregressive Moving Average". Mô hình ARMA kết hợp cả hai thành phần chính là thành phần tự hồi quy (AR) và thành phần trung bình di chuyển (MA). Một mô hình ARMA(p, q) bao gồm p thành phần tự hồi quy và q thành phần trung bình di chuyển.

Công thức tổng quát của mô hình ARMA(p, q) là:

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} + \varepsilon_t + \phi_1 X_{t-1} + \phi_2 X_{t-2} + ... + \phi_p X_{t-p} \quad (6)$$

Trong đó:

- X_t là giá trị của chuỗi thời gian tại thời điểm t.
- c là hệ số chệch (intercept).
- $\phi_1, \phi_2, ..., \phi_p$ là các hệ số tự hồi quy.
- $X_{t-1}, X_{t-2}, ..., X_{t-p}$ là giá trị của chuỗi tại các thời điểm lag.
- ε_t là sai số ngẫu nhiên tại thời điểm t.
- $\theta_1, \theta_2, ..., \phi_q$ là các hệ số trọng số của sai số ngẫu nhiên tại các thời điểm trước đó (lags).

Mô hình ARMA được sử dụng để mô hình hóa và dự báo chuỗi thời gian có cả yếu tố xu hướng tự hồi quy và yếu tố biến động ngẫu nhiên từ sai số.

3.7 ARIMA - Autoregressive Integrated Moving Average

ARIMA là viết tắt của "Autoregressive Integrated Moving Average". Mô hình ARIMA là một phương pháp phổ biến được sử dụng để mô hình hóa và dự báo chuỗi thời gian. Mô hình này kết hợp cả ba thành phần chính: thành phần tự hồi quy (AR), thành phần trung bình di chuyển (MA), và thành phần chuyển động (I) để xử lý xu hướng không đồng nhất.

Thành phần (I - Intergrated): Là quá trình đồng tích hợp hoặc lấy sai phân. Yêu cầu chung của các thuật toán trong time series là chuỗi phải đảm bảo tính dừng. Hầu

hết các chuỗi đều tăng hoặc giảm theo thời gian. Do đó yếu tố tương quan giữa chúng chưa chắc là thực sự mà là do chúng cùng tương quan theo thời gian. Khi biến đổi sang chuỗi dừng, các nhân tố ảnh hưởng thời gian được loại bỏ và chuỗi sẽ dễ dự báo hơn. Để tạo thành chuỗi dừng, một phương pháp đơn giản nhất là chúng ta sẽ lấy sai phân. Một số chuỗi tài chính còn qui đổi sang logarit hoặc lợi suất. Bậc của sai phân để tạo thành chuỗi dừng còn gọi là bậc của quá trình đồng tích hợp (order of intergration). Qúa trình sai phân bậc của chuỗi được thực hiện như sau:

- Sai phân bậc 1: $\mathrm{I}(1) = \Delta(x_t) = x_t - x_{t-1}$
- Sai phân bậc d: $\mathrm{I}(d) = \Delta^d(x_t) = \underbrace{\Delta(\Delta(\ldots\Delta(x_t)))}_{\text{d times}}$

Thông thường chuỗi sẽ dừng sau quá trình đồng tích hợp I(0) hoặc I(1). Rất ít chuỗi chúng ta phải lấy tới sai phân bậc 2. Một số trường hợp chúng ta sẽ cần biến đổi logarit hoặc căn bậc 2 để tạo thành chuỗi dừng. Phương trình hồi qui ARIMA(p, d, q) có thể được biểu diễn dưới dạng:

$$\Delta x_t = \phi_1 \Delta x_{t-1} + \phi_2 \Delta x_{t-2} + \ldots + \phi_p \Delta x_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} \qquad (7)$$

Trong đó Δx_t là giá trị sai phân bậc d và ε_t là các chuỗi nhiễu trắng.

3.8 SARIMA - Seasonal Autoregressive Integrated Moving Average.

SARIMA là một viết tắt của "Seasonal Autoregressive Integrated Moving Average." Đây là một loại mô hình trong phân tích chuỗi thời gian, được phát triển để mô hình hóa và dự báo dữ liệu có yếu tố mùa vụ.

Mô hình SARIMA là sự kết hợp của mô hình ARIMA với yếu tố mùa vụ (seasonal) để xử lý biến động theo mùa trong dữ liệu.

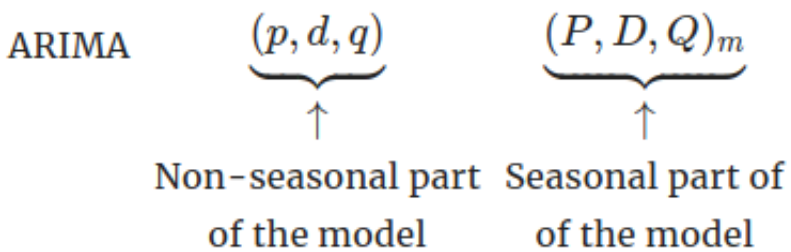


Figure 5: Minh hoạ cho mô hình SARIMA

Do đó, mô hình được biểu diễn như sau: SARIMA$(p,d,q)(P,D,Q)_s$ với các thông số:

- p, d, q là các tham số không mùa tương ứng với phần AR, I, và MA.
- P, D, Q là các tham số mùa vụ tương ứng với phần AR, I, và MA mùa vụ.
- s là chu kỳ mùa vụ, đại diện cho số lượng quan sát trong một chu kỳ mùa.

Do những yếu tố trên nên SARIMA là một mô hình mạnh mẽ cho dữ liệu có yếu tố mùa vụ và được sử dụng rộng rãi trong dự báo chuỗi thời gian.

4 Kết quả thực nghiệm và đánh giá

Với bộ dữ liệu Monthly Cars Sales đã giới thiệu ở bên trên kết hợp với cơ sở lý thuyết, chúng em sẽ thực hiện đưa ra dự đoán giá ô tô của 10 năm tới. Sau đây sẽ là cụ thể từng bước thực hiện, code, kết quả và đánh giá thực nghiệm trên từng bước.

4.1 Quy trình thực hiện

Bước 1: Khai báo thư viện

```r
{r}
library(TSA)
library(forecast)
library(lmtest)
library(fUnitRoots)
library(tseries)
library(knitr)
library(dLagM)
library(lattice)
library(bestglm)
library(leaps)
library(ltsa)
library(FitAR)
library(CombMSC)
library(lmtest)
library(fGarch)
library(zoo)
library(astsa)
```

Bước 2: Khai phá dữ liệu

1. Đọc dữ liệu

```r
cars <- read.csv("cars.csv",header=TRUE)
head(cars)
```

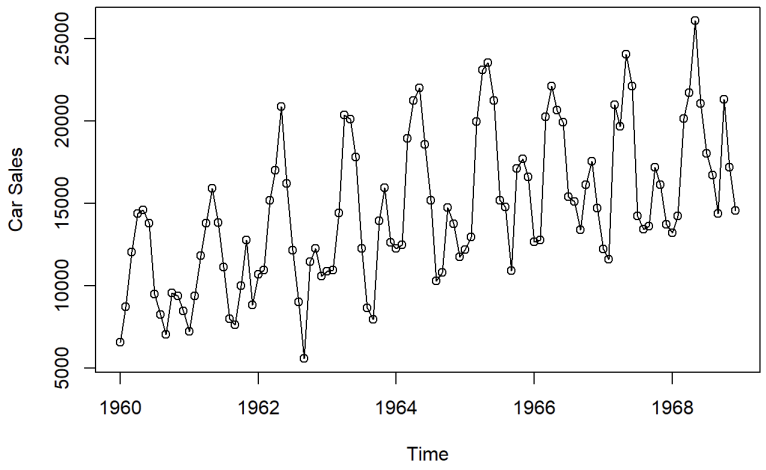
	Year <int>	Month <int>	Sales <int>
1	1960	1	6550
2	1960	2	8728
3	1960	3	12026
4	1960	4	14395
5	1960	5	14587
6	1960	6	13791

6 rows

2. Phân tích dữ liệu

   - Như đã nói ở trên, dữ liệu thu thập được là dữ liệu dạng thô, nên chúng ta phải chuyển dữ liệu về dạng chuỗi thời gian

```
cars.ts <- matrix(cars$Sales,nrow=108,ncol=1)
cars.ts<- as.vector(t(cars.ts))
cars.ts <- ts(cars.ts,start=c(1960,1), end=c(1968,12), frequency=12)
plot(cars.ts,type='o',ylab='Car Sales')
```

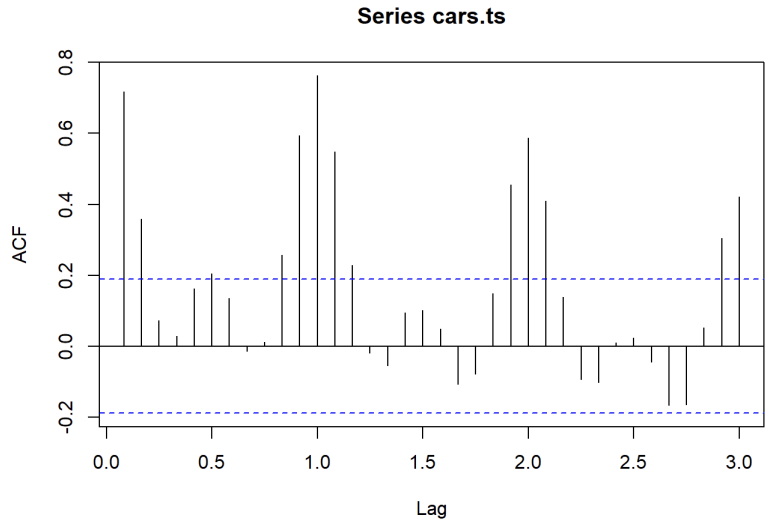


===> Nhận xét biểu đồ chuỗi thời gian:

- Xu hướng mua xe (Trend): Đang có xu hướng tăng mạnh.
- Tính mùa vụ (Seasonality): Số lượng mua xe cao/thấp là sự biến động có chu kỳ và lặp lại theo thời gian. Như đã phâ tích bên trên thì do yếu tố thời tiết và luật pháp, xu hướng mua xe cao vào mùa hè và thấp vào mùa đông.

• Vẽ biểu đồ ACF với dữ liệu chuỗi thời gian

```
acf(cars.ts,lag.max = 36)
```



===> Nhận xét biểu đồ ACF:

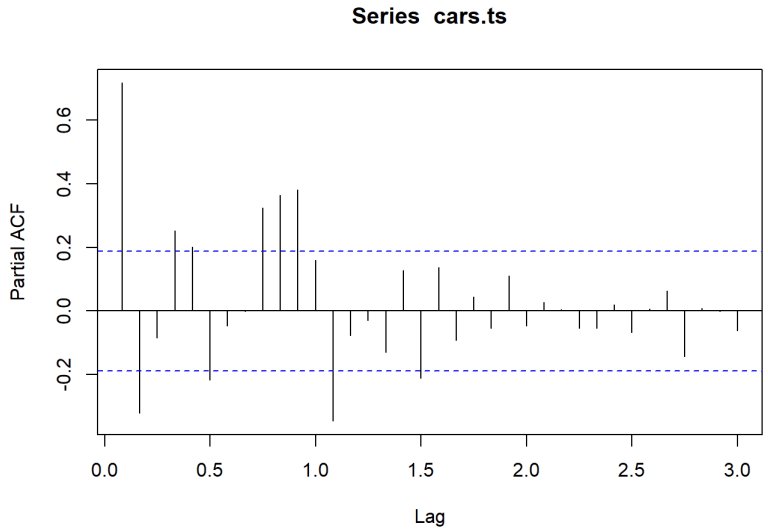
- Đồ thị AFC thể hiện mức độ tương quan giữa các phần tử trong chuỗi thời gian. Độ trễ (lags) là sự tương quan giữa quan sát chuỗi thời gian và các

quan sát trước đó.

– Bởi vì chúng ta quan sát có sự tương quan mạnh tại các lags 12, 24, 36 và các lag khác nhau theo cấp số nhân của 12, chúng ta có thể kết luận về sự tồn tại của mối quan hệ tương quan theo mùa vụ.

- Vẽ biểu đồ PACF với dữ liệu chuỗi thời gian

```
pacf(cars.ts, lag.max = 36)
```



===> Nhận xét biểu đồ PACF:

– Hàm PACF (Partial AutoCorrelation Function) đo lường mức độ tương quan giữa một biến và độ trễ của nó mà không giải thích được bởi tương quan ở các lags thấp hơn.

– PACF bên trên cho thấy độ trễ theo mùa là 12.

3. Yếu tố mùa vụ: Khi chuỗi dữ liệu có tính mùa vụ, ta cần phải kiểm tra sự chênh lệch giữa các điểm dữ liệu. Tính mùa vụ thường gây ra sự không ổn định trong chuỗi thời gian vì giá trị trung bình tại một số thời điểm cụ thể trong chu kỳ mùa vụ có thể khác biệt so với giá trị trung bình tại các thời điểm khác.

   Trong bài báo cáo của chúng em, chúng em sẽ lần lượt đi thử với các trường hợp sau: ARIMA(0,0,0)x(0,1,0), ARIMA(0,0,0)x(1,1,1), ARIMA(0,0,0)x(1,1,2)

3.1. Đặc điểm và cách giải quyết của mùa vụ với mô hình ARIMA(0,0,0)x(0,1,0)

(a) Đầu tiên, chúng ta thực hiện chênh lệch mùa để loại bỏ xu hướng theo mùa và điều chỉnh một mô hình đơn giản cho chuỗi thời gian, cho đến khi chuỗi thời gian và đồ thị ACF/PACF của phần dư không cho thấy dấu hiệu của tính mùa vụ.

(b) Sau đó, chúng ta xác định các tham số P & Q của mùa dựa trên đồ thị ACF/-PACF của phần dư cuối cùng.

(c) Ta thực hiện điều này bằng cách điều chỉnh mô hình ARIMA(0,0,0)x(0,1,0) và vẽ các biểu đồ.

(d) Mặc dù chúng ta đã giải quyết xu hướng tổng quát, nhưng vẫn cần kiểm tra tương quan tự (ACF) và tương quan riêng (PACF) để đánh giá các mối quan hệ tương quan trong phần dư.

```
m1.cars = arima(cars.ts,order=c(0,0,0),seasonal=list(order=c(0,1,0), period=12))
res.m1 = residuals(m1.cars);
par(mfrow=c(1,1))
plot(res.m1,xlab='Time',ylab='Residuals')
```

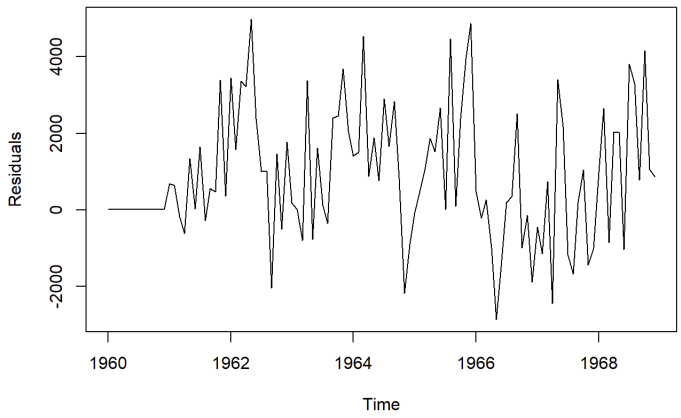


Figure 6: Biểu đồ chuỗi thời gian về sự chênh lệch mùa vụ đầu tiên

Dưới đây là biểu đồ ACF & PACF hiển thị độ trễ theo mùa 1, biểu thị SARMA(1,1)

```
par(mfrow=c(1,2))
acf(res.m1, lag.max = 36)
pacf(res.m1, lag.max = 36)
```

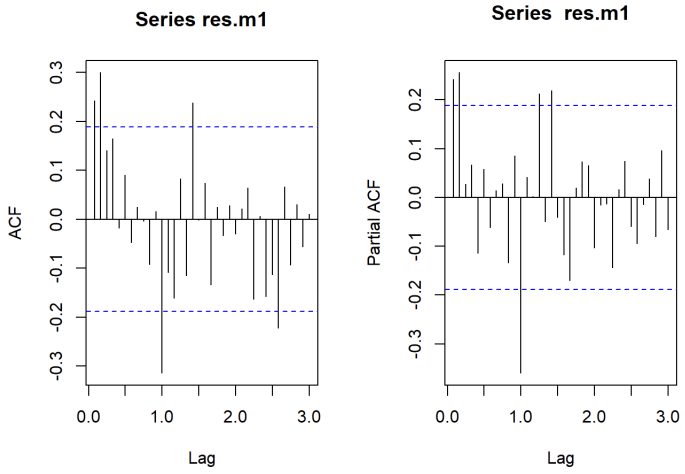


Figure 7: Biểu đồ ACF/PACF của phần dư.

3.2. Đặc điểm và cách giải quyết của mùa vụ với mô hình ARIMA(0,0,0)x(1,1,1)

Xu hướng chung đi lên không còn nữa, do đó chúng em sẽ tạo biểu đồ ACF và PACF của phần dư.

```
m2.cars = arima(cars.ts,order=c(0,0,0),seasonal=list(order=c(1,1,1), period=12))
res.m2 = residuals(m2.cars);
par(mfrow=c(1,1))
plot(res.m2,xlab='Time',ylab='Residuals',main="Time series plot of the residuals")
```

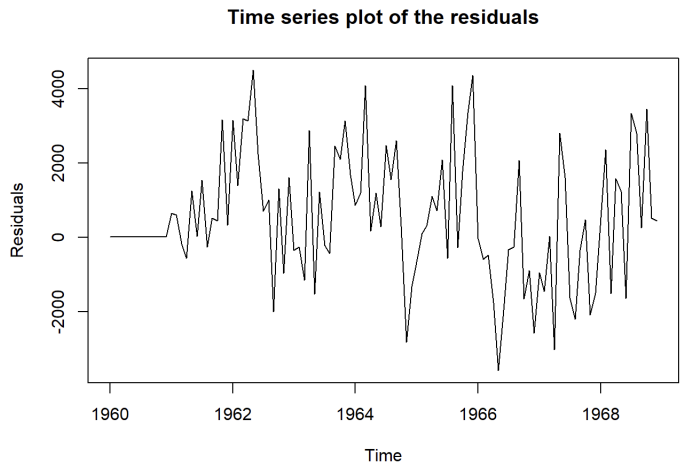


Figure 8: Biểu đồ chuỗi thời gian với hệ số AR & MA theo mùa (P=1,Q=1)

```
par(mfrow=c(1,2))
acf(res.m2, lag.max = 36)
pacf(res.m2, lag.max = 36)
```

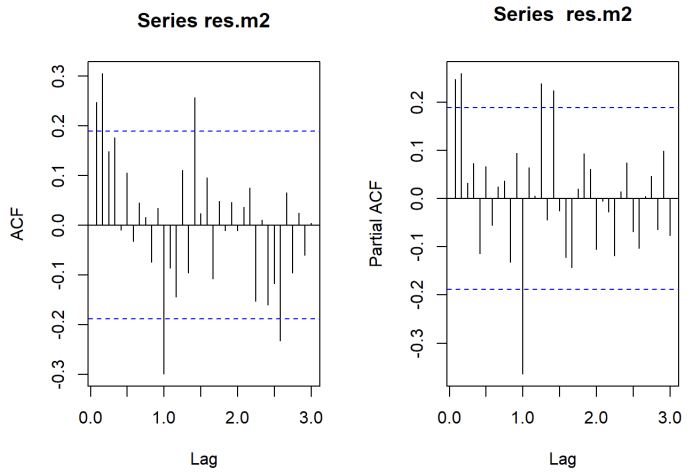


Figure 9: Biểu đồ ACF/PACF của phần dư có hệ số AR và MA (P=1,D=1,Q=1)

Nhìn vào các đồ thị ở trên, chúng em rút ra nhận xét: sự tự tương quan vẫn tồn tại ở các độ trễ mùa. Do đó, chúng ta cần lặp lại quá trình với Q cao hơn cho đến khi loại bỏ hết yếu tố mùa vụ.

3.3. Đặc điểm và cách giải quyết của mùa vụ với mô hình ARIMA(0,0,0)x(1,1,2)

(a) Hiện tại không có độ trễ theo mùa nào xuất hiện. Do đó, việc đặc tả tính thời vụ hoàn thành tại P=1, D=1, Q=2. Chúng ta nhận thấy có nhiều độ trễ không theo chu kỳ mùa vụ trong phần phi mùa vụ. (tức là độ trễ trước độ trễ mùa vụ đầu tiên).

(b) Vì chúng ta không thấy xu hướng nào, chúng ta đầu tiên áp dụng biến đổi và kiểm tra các độ trễ đáng kể.

```
m3.cars = arima(cars.ts,order=c(0,0,0),seasonal=list(order=c(1,1,2), period=12))

res.m3 = residuals(m3.cars)
par(mfrow=c(1,2))
acf(res.m3, lag.max = 36)
pacf(res.m3, lag.max = 36)
```

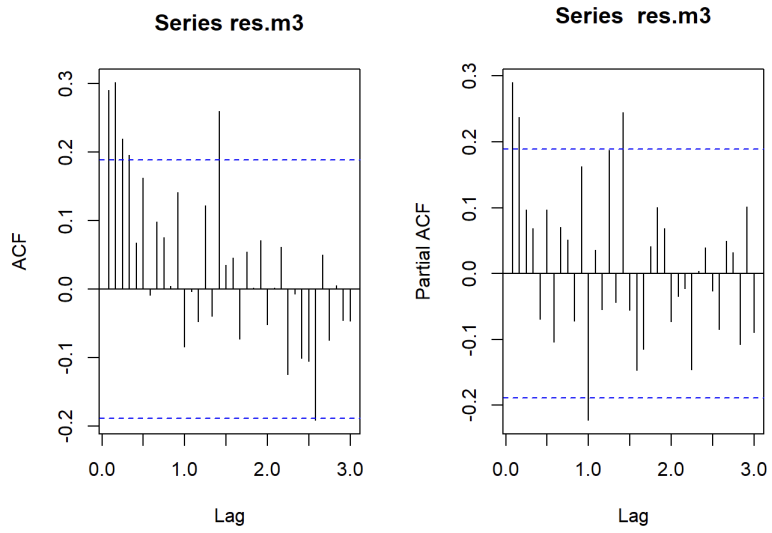


Figure 10: Biểu đồ ACF/PACF của phần dư với P=1,D=1,Q=2

4. Biến đổi (Transformation): Mục tiêu của bước biến đổi này chính là biến từ chuỗi không dừng thành chuỗi dừng. Phương pháp chúng em lựa chọn là biến đổi hàm log trên dữ liệu chuỗi thời gian

```
log.cars.ts = log(cars.ts)
par(mfrow=c(1,1))
plot(log.cars.ts,ylab='log of sales count',xlab='Year',type='o')
```

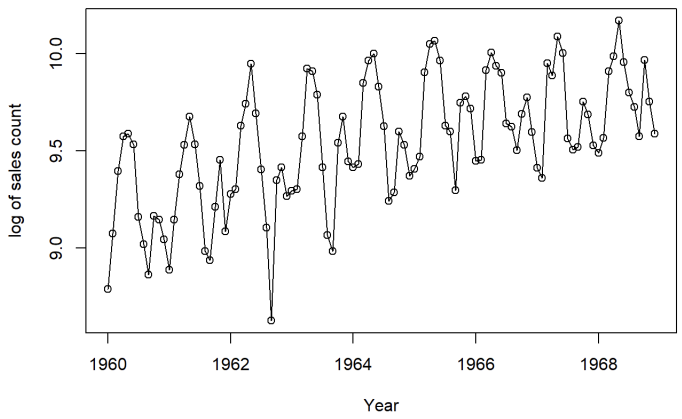


Figure 11: Biểu đồ chuỗi thời gian với dữ liệu được chuyển đổi

```
m4.cars = arima(log.cars.ts,order=c(0,0,0),seasonal=list(order=c(1,1,2), period=12))
res.m4 = residuals(m4.cars)
plot(res.m4,xlab='Time',ylab='Residuals',main="Time series plot of the residuals")
```

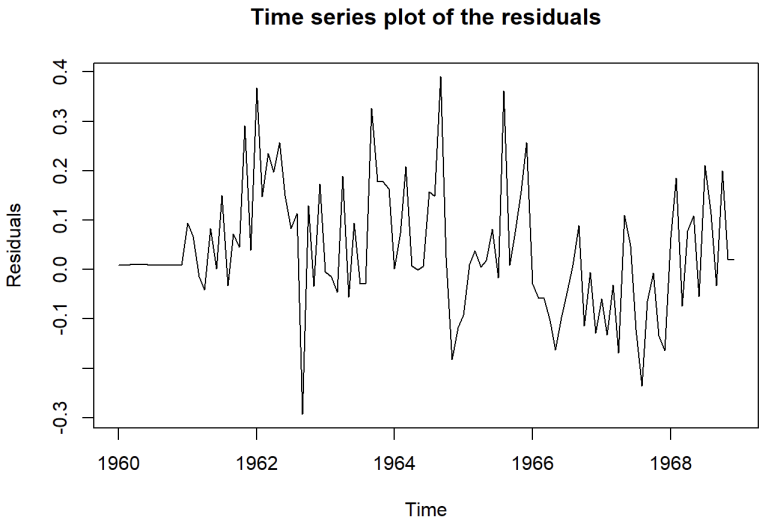


Figure 12: Biểu đồ chuỗi thời gian của phần dư sau khi chuyển đổi

```
acf(res.m4, lag.max = 36)
```

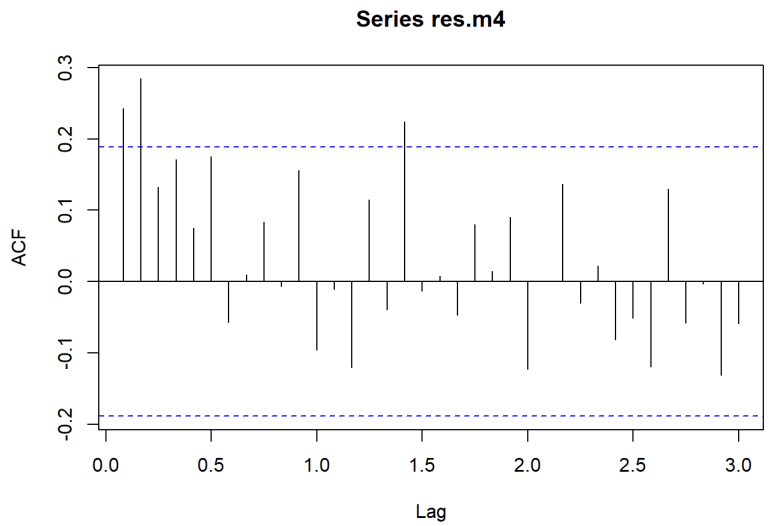


Figure 13: Biểu đồ ACF của phần dư sau biến đổi.

```
pacf(res.m4, lag.max = 36)
```

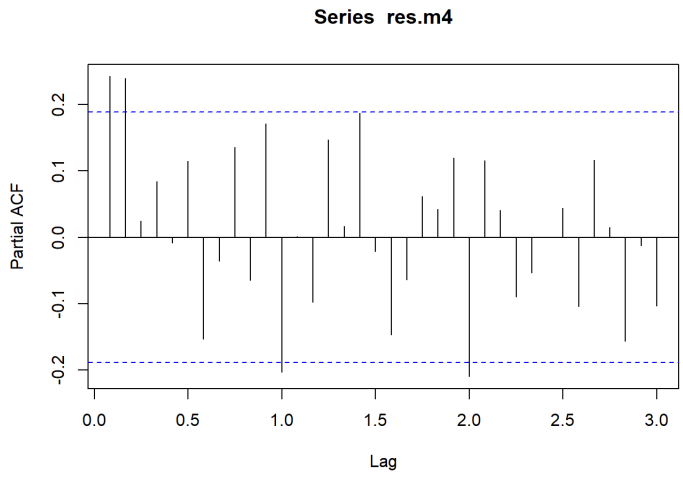


Figure 14: Biểu đồ PACF của phần dư sau biến đổi.

Nhìn vào biểu đồ ACF, chúng em nhận thấy rằng hai độ trễ có ý nghĩa trước độ trễ mùa vụ đầu tiên.

5. Non-seasonal Differencing (Sự khác biệt không theo mùa). Mục đích chính ở đây là đưa ra tập hợp các mô hình có thể sử dụng. Biến đổi từ chuỗi không dừng sang chuỗi dừng với sai phân bậc d = 1 để loại bỏ xu hướng còn lại và mối tương quan còn lại trong biểu đồ ACF/PACF. Và dưới đây sẽ là các biểu đồ thống kê mà chúng em rút ra.

===> Nhận xét:

```
# SARIMA(0,1,0)x(1,1,2)

m5.cars = arima(log.cars.ts,order=c(0,1,0),seasonal=list(order=c(1,1,2), period=12))
res.m5 = residuals(m5.cars)
#plot(res.m5,xlab='Time',ylab='Residuals',main="Time series plot of the residuals")
par(mfrow=c(1,2))
acf(res.m5, lag.max = 36)
pacf(res.m5, lag.max = 36)
```

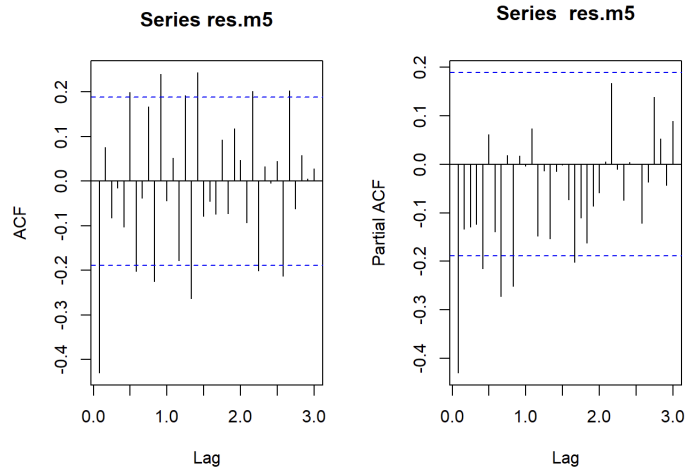


Figure 15: Biểu đồ ACF/PACF của phần dư với sai phân thông thường.

- Chúng em thấy một tương quan cao tại độ trễ đầu tiên và cũng quan sát thấy một số độ trễ có ý nghĩa. Tất cả điều này là do điểm can thiệp.
- Chúng em đề xuất sử dụng MA(3 hoặc 4) và AR(3) từ các biểu đồ ACF và PACF.
- Ngoài ra, không có bằng chứng cho một xu hướng thông thường. Chúng em vẫn có thể áp dụng kiểm tra ADF (Augmented Dickey-Fuller) trên phần dư để đảm bảo.

```
adf.test(res.m5)
```

```
## Warning in adf.test(res.m5): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  res.m5
## Dickey-Fuller = -6.9914, Lag order = 4, p-value = 0.01
## alternative hypothesis: stationary
```

6. EACF: EACF về cơ bản là một công cụ khác giống như ACF/PACF, được sử dụng để xác định thứ bậc của mô hình ARIMA. Trong bảng EACF với hệ số MA được liệt kê ở phía trên và hệ số AR được liệt kê ở phía dưới, hầu hết EACF ở trên cùng bên trái nhỏ hơn 2 lần giá trị tuyệt đối của độ lệch chuẩn của EACF thường là sự lựa chọn tốt cho các thứ bậc của mô hình.

   Chúng em sử dụng EACF trên phần dư của bước trước đó (tức là res.m5), để xem

thông tin về các phần AR và MA còn lại trong phần dư. Các ứng viên cho phần ARMA là ARMA(1,2), ARMA(2,2) & ARMA (2,1).

```
eacf(res.m5)
```

```
## AR/MA
##   0 1 2 3 4 5 6 7 8 9 10 11 12 13
## 0 x o o o o x x o o x x  o  o  o
## 1 x x o o o o x o o o x  o  o  o
## 2 x o o o o o o o o o x  o  o  o
## 3 x o o o o o o o o o o  o  o  o
## 4 x o o o o o o o o o x  o  o  o
## 5 x x x x o o o o o o o  o  o  o
## 6 x o x o x o o o o o o  o  o  o
## 7 x x x o x o x o o o o  o  o  o
```

Các mô hình dự kiến được xác định như sau:

- SARIMA (0,1,4)x(1,1,2) dựa trên ACF nếu xem xét rằng PACF đang giảm đi, vì vậy chỉ lấy hệ số MA.
- SARIMA (0,1,3)x(1,1,2) dựa trên ACF nếu xem xét rằng PACF đang giảm đi, vì vậy chúng ta chỉ lấy hệ số MA.
- SARIMA (3,1,4)x(1,1,2) dựa trên ACF/PACF.
- SARIMA (2,1,1)x(1,1,2) dựa trên EACF.
- SARIMA (2,1,2)x(1,1,2) dựa trên EACF.
- SARIMA (3,1,2)x(1,1,2) để tránh quá mức phù hợp với SARIMA (2,1,2)x(1,1,2).

Bước 3: Khớp mô hình - Model Fitting: Ước lượng tham số: Từ tập hợp các mô hình đã cho, bắt đầu điều chỉnh từng mô hình một và xem xét xem có mô hình nào phù hợp.

- Điều chỉnh các mô hình này trên dữ liệu gốc để thu được các phần dư
- phân tích phần dư để xem xét xem chúng có phải là nhiễu trắng không.

1. SARIMA(0,1,3)x(1,1,2): ACF/PACF cho thấy sự hiện diện của nhiễu trắng và đây thực sự là một trong những mô hình tốt.

```
model2.cars = arima(log.cars.ts,order=c(0,1,3),seasonal=list(order=c(1,1,2), period=12),method = "ML")
coeftest(model2.cars)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ma1  -0.750678   0.103762 -7.2346 4.668e-13 ***
## ma2   0.074003   0.137082  0.5398  0.589307
## ma3  -0.217749   0.108075 -2.0148  0.043927 *
## sar1  0.480677   0.334715  1.4361  0.150979
## sma1 -1.084238   0.393826 -2.7531  0.005904 **
## sma2  0.084263   0.325977  0.2585  0.796027
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

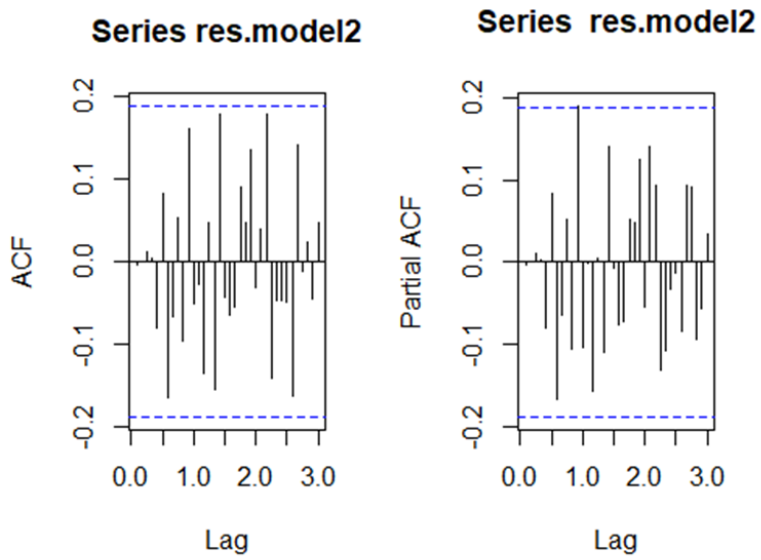


Figure 16: Biểu đồ ACF/PACF của phần dư cho ARIMA(0,1,3)x(1,1,2)

2. SARIMA(0,1,4)x(1,1,2): Kiểm tra hệ số cho thấy rằng MA(2), MA(4) không có ý nghĩa trong khi MA(3) có ý nghĩa nhưng rất nhỏ. Do đó, chúng ta không thể sử dụng mô hình này cho phân tích tiếp theo. Ngoài ra, trong PACF, chúng ta thấy một số độ trễ đáng kể nhưng cũng rất nhỏ.

```
model1.cars = arima(log.cars.ts,order=c(0,1,4),seasonal=list(order=c(1,1,2), period=12),method = "ML")
coeftest(model1.cars)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value  Pr(>|z|)
## ma1  -0.716539   0.096911 -7.3938 1.427e-13 ***
## ma2   0.070897   0.122981  0.5765   0.56429
## ma3  -0.205801   0.119286 -1.7253   0.08448 .
## ma4  -0.041378   0.083517 -0.4954   0.62028
## sar1 -0.197613         NA      NA        NA
## sma1 -0.290238         NA      NA        NA
## sma2 -0.336851         NA      NA        NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. SARIMA(3,1,4)x(1,1,2): Kiểm tra hệ số không cho thấy bất kỳ giá trị quan trọng nào nhưng nhận thấy sự hiện diện của nhiễu trắng từ phần dư. Do đó kiểm tra xem mô hình này có phù hợp hay không sau.

```
model3.cars = arima(log.cars.ts,order=c(3,1,4),seasonal=list(order=c(1,1,2), period=12),method = "ML")
coeftest(model3.cars)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value Pr(>|z|)
## ar1   0.726370   0.912899  0.7957  0.42622
## ar2   0.538303   0.349953  1.5382  0.12400
## ar3  -0.472199   0.411453 -1.1476  0.25112
## ma1  -1.458156   0.908939 -1.6042  0.10866
## ma2   0.047285   0.723648  0.0653  0.94790
## ma3   0.608440   0.565129  1.0766  0.28164
## ma4  -0.170793   0.238443 -0.7163  0.47382
## sar1  0.438596   0.384708  1.1401  0.25426
## sma1 -1.047716   0.546545 -1.9170  0.05524 .
## sma2  0.072712   0.371171  0.1959  0.84469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

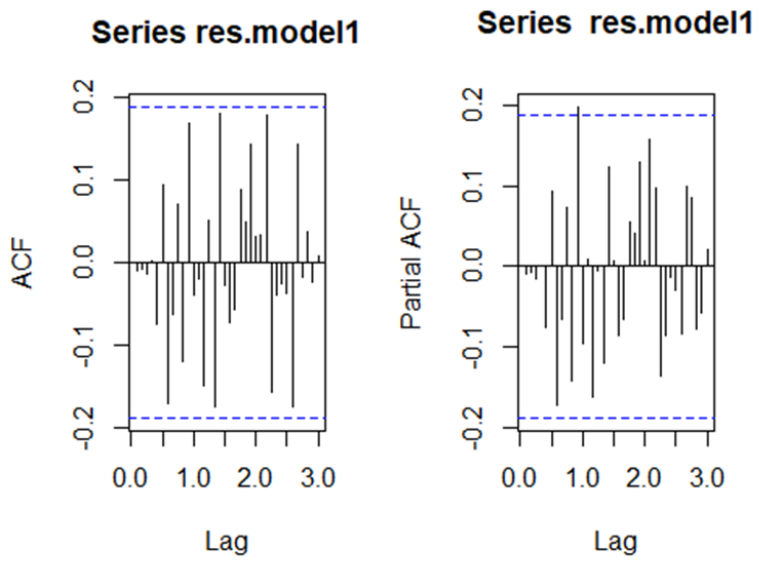


Figure 17: Biểu đồ ACF/PACF của phần dư cho ARIMA(0,1,3)x(1,1,2)

4. SARIMA(2,1,1)x(1,1,2): Hệ số MA(1) hiển thị các giá trị quan trọng nhưng thành phần AR thì không. Mô hình không có ý nghĩa

```
model4.cars = arima(log.cars.ts,order=c(2,1,1),seasonal=list(order=c(1,1,2), period=12),method = "ML")
coeftest(model4.cars)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error  z value  Pr(>|z|)
## ar1   0.182219   0.111112   1.6400  0.101014
## ar2   0.207476   0.116201   1.7855  0.074180 .
## ma1  -0.943115   0.051848 -18.1901 < 2.2e-16 ***
## sar1  0.511863   0.312659   1.6371  0.101604
## sma1 -1.102519   0.380088  -2.9007  0.003723 **
## sma2  0.102524   0.307102   0.3338  0.738497
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

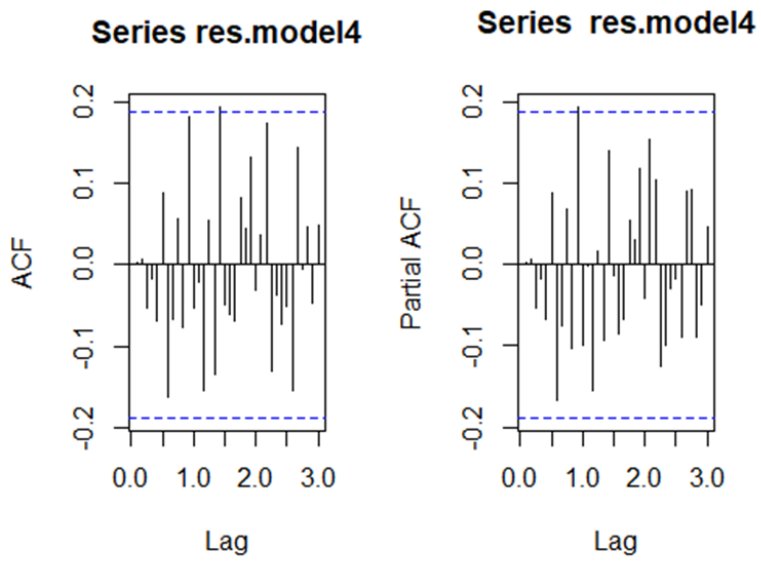


Figure 18: Biểu đồ ACF/PACF của phần dư cho ARIMA(2,1,1)x(1,1,2)

5. SARIMA(2,1,2)x(1,1,2): Chỉ AR(2) là có ý nghĩa và ACF/PACF cho thấy sự hiện diện của nhiễu trắng.

```
model5.cars = arima(log.cars.ts,order=c(2,1,2),seasonal=list(order=c(1,1,2), period=12),method = "ML")
coeftest(model5.cars)
```

```
##
## z test of coefficients:
##
##       Estimate Std. Error z value Pr(>|z|)
## ar1  -0.309047   0.378002 -0.8176 0.413597
## ar2   0.299031   0.119764  2.4968 0.012531 *
## ma1  -0.423538   0.392802 -1.0782 0.280924
## ma2  -0.476045   0.347773 -1.3688 0.171049
## sar1  0.457113   0.333679  1.3699 0.170712
## sma1 -1.045249   0.393968 -2.6531 0.007975 **
## sma2  0.045396   0.329495  0.1378 0.890417
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

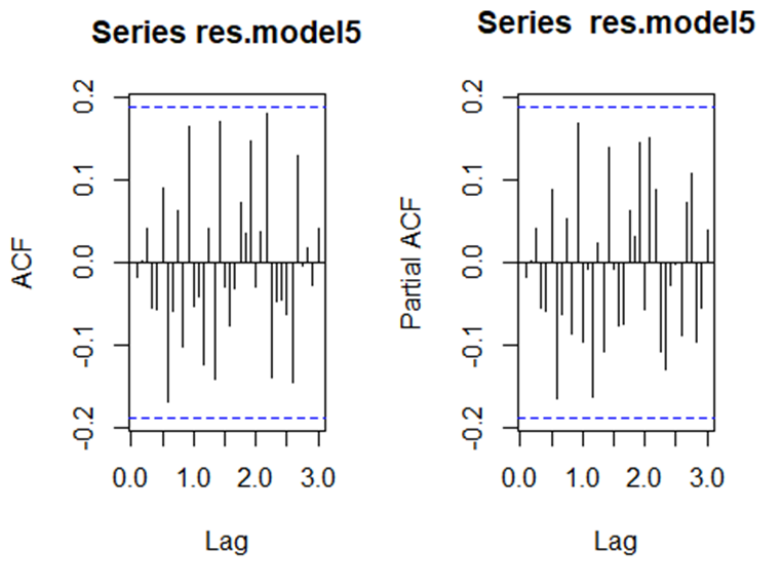


Figure 19: Biểu đồ ACF/PACF của phần dư cho ARIMA(2,1,2)x(1,1,2)

Bước 4: Chẩn đoán mô hình - Model Diagnostics

Trong bài toán dự báo chuỗi thời gian, việc lựa chọn mô hình phù hợp là rất quan trọng để đảm bảo độ chính xác của dự báo. Để đánh giá và so sánh các mô hình, chúng ta thường sử dụng hai tiêu chí phổ biến là Akaike Information Criterion (AIC) và Bayesian Information Criterion (BIC).

Chúng em đã tính toán giá trị AIC và BIC cho 5 mô hình SARIMA khác nhau, và kết quả như sau:

```
sort.score(sc.AIC, score = "aic")
```

	df <dbl>	AIC <dbl>
model2.cars	7	-129.3116
model4.cars	7	-129.1109
model5.cars	8	-127.7769
model1.cars	8	-126.5361
model3.cars	11	-122.6328

5 rows

```r
sort.score(sc.BIC, score = "aic")
```

	df <dbl>	AIC <dbl>
model2.cars	7	-110.53671
model4.cars	7	-110.33597
model5.cars	8	-106.31981
model1.cars	8	-105.07908
model3.cars	11	-93.12932

5 rows

Cả hai tiêu chí AIC và BIC đều chọn mô hình **SARIMA(0,1,3)x(1,1,2)** là mô hình tốt nhất.

Một yếu tố quan trọng khác trong việc chọn lựa mô hình là kiểm tra phần dư của mô hình. Một mô hình tốt là mô hình có phần dư là nhiễu trắng, nghĩa là phần dư không có cấu trúc tự tương quan. Điều này đảm bảo rằng mô hình đã nắm bắt đầy đủ thông tin có trong chuỗi thời gian và các dự báo sẽ không bị thiên vị.

```r
residual.analysis(model=model2.cars)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res.model
## W = 0.9716, p-value = 0.02057
```

```
## Warning in (ra^2)/(n - (1:lag.max)): longer object length is not a multiple of
## shorter object length
```

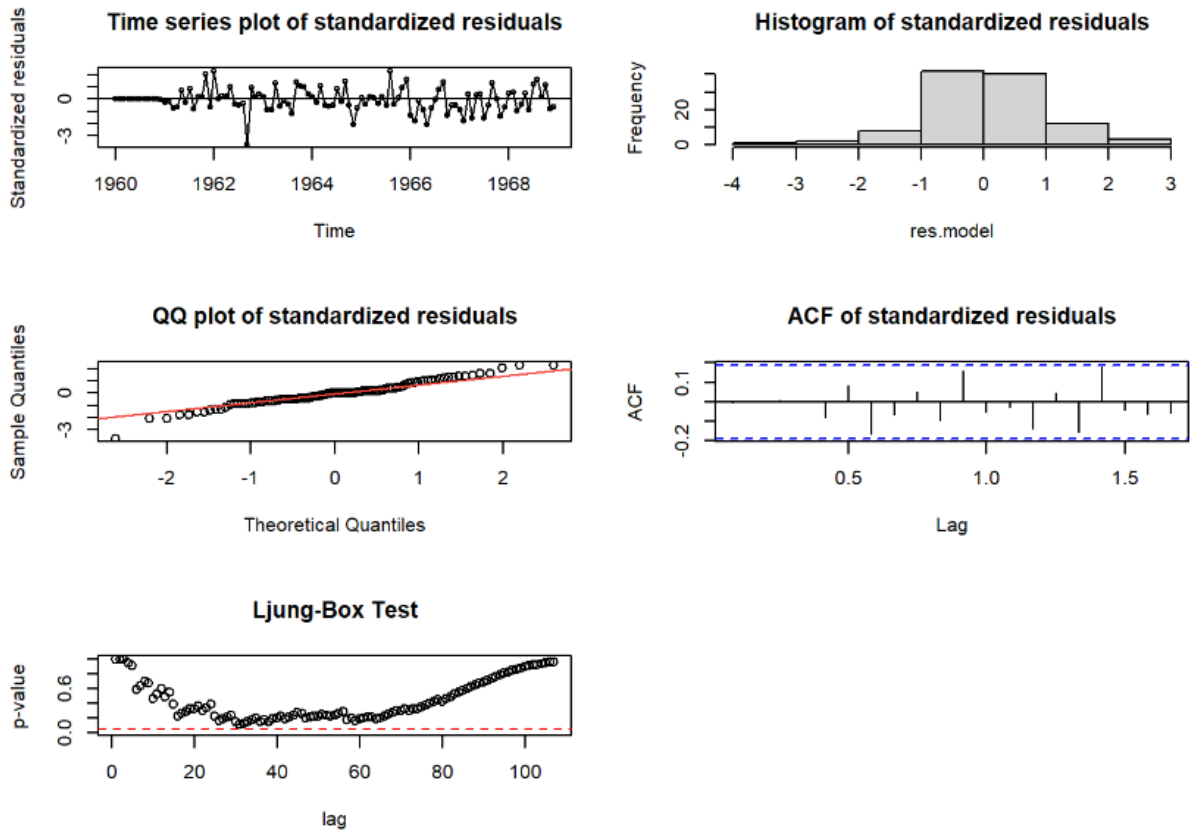


Figure 20: Phân tích dư cho mô hình ARIMA(0,1,3)x(1,1,2)

Phân tích phần dư của mô hình SARIMA(0,1,3)x(1,1,2) cho thấy mô hình hoạt động khá tốt. Biểu đồ histogram cho thấy phần dư phân phối gần như chuẩn, phản ánh rằng mô hình đã phù hợp với dữ liệu. Biểu đồ ACF chỉ ra rằng phần dư là nhiễu trắng, không có sự tự tương quan đáng kể, chứng tỏ mô hình đã xử lý hiệu quả các yếu tố tự tương quan. Kết quả từ kiểm định Ljung-Box cũng không phát hiện vấn đề nghiêm trọng, xác nhận rằng mô hình không còn sót lại cấu trúc chưa được mô hình hóa. Mặc dù biểu đồ QQplot phát hiện một điểm ngoại lệ ở đầu, cho thấy một số điểm dữ liệu không hoàn toàn tuân theo phân phối chuẩn, phần lớn phần dư vẫn nằm gần đường chéo chuẩn. Cuối cùng, đồ thị chuỗi thời gian của phần dư không cho thấy xu hướng rõ ràng nào ngoài một điểm đặc biệt, điều này đồng nhất với các phân tích trước và cho thấy phần dư chủ yếu không có cấu trúc hệ thống rõ ràng.

4.2 Kết quả

Sau khi xác định mô hình ARIMA(0,1,3)x(1,1,2) là mô hình tối ưu, chúng em sẽ sử dụng mô hình này để dự đoán các giá trị tương lai của chuỗi thời gian. Dưới đây là biểu đồ dự đoán cho khoảng thời gian 10 năm tiếp theo dựa trên mô hình ARIMA(0,1,3)x(1,1,2).

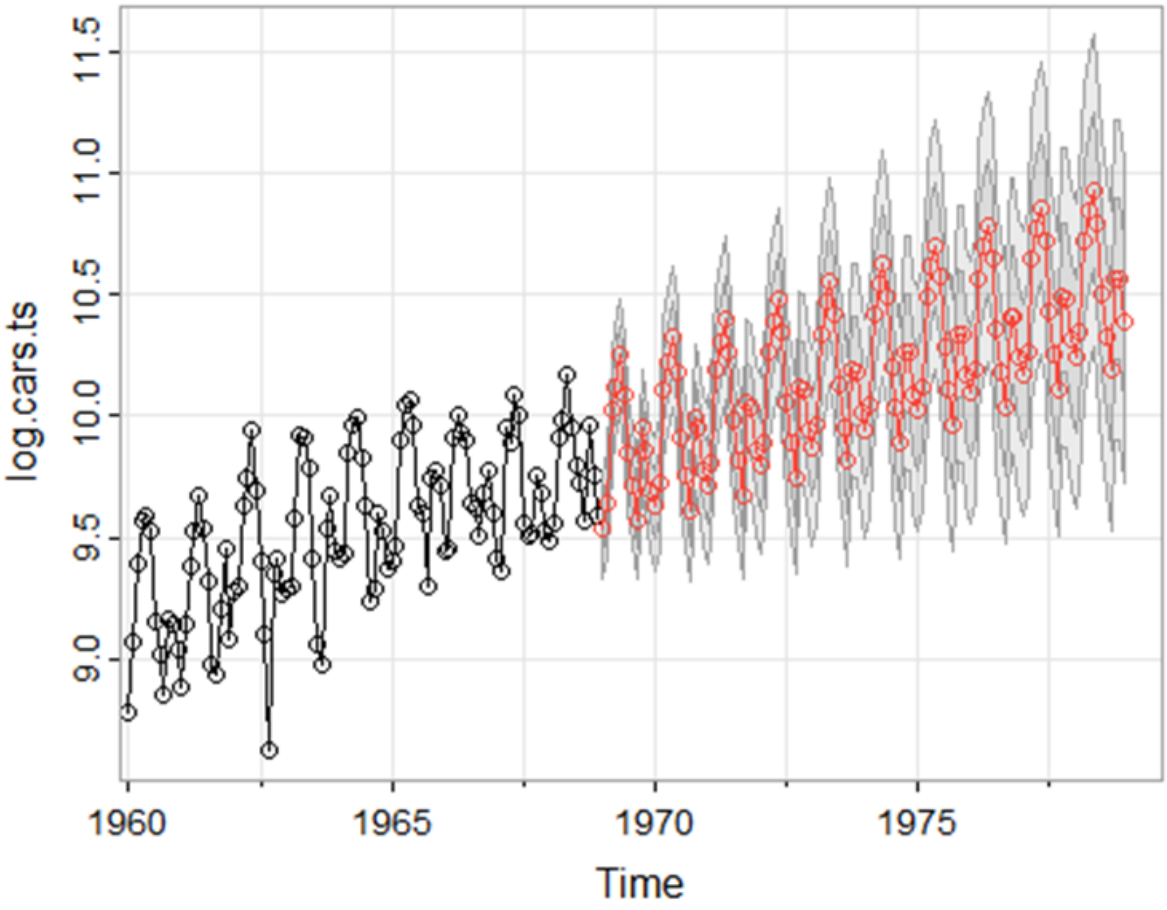


Figure 21: Biểu đồ dự đoán doanh số bán ô tô trong 10 năm tới

4.3 Đánh giá

Dựa trên mô hình ARIMA(0,1,3)x(1,1,2), kết quả dự đoán cho 10 năm tới cho thấy mô hình có khả năng dự đoán chính xác các giá trị tương lai của chuỗi thời gian. Biểu đồ dự đoán cung cấp cái nhìn rõ ràng về các xu hướng và biến động dự kiến trong tương lai, từ đó hỗ trợ việc ra quyết định dựa trên phân tích dữ liệu.