Time Series Project: Housing Sales
Winter 2014

Yu-Ching Eugene Chen

March 27, 2014

1 Introduction

House is often considered an expensive item to purchase. Housing sales and price are subject to serial correlation and seasonal fluctuation. In this project, I study the monthly housing sales under time series framework. I break down the variable further and identify the time effect on each of the components. I then fit time series models and evaluate the goodness-of-fit. Finally, I reserve 12 monthly data for each model and compare the forecast with real data.

2 Data

I gather monthly housing data from the NEAS forum. Housing sales are presented in thousands. Those numbers are broken down into 3 categories: "Not Started," "Under Construction," and "Completed." Because they are expressed in terms of housing unit, so there is no need for CPI adjustment. The monthly data is available from January 1963 to March 2008, which is in the midst of housing market meltdown. A glance from Figure 1 suggests that the downward trend started since 2006.

I also downloaded the 10-year Treasury rates over the same time span. The general consensus is that housing sales are negatively affected by mortgage rates, which is tightened to 10-year Treasury rates. Mortgage rates varies by lender, borrower, and states, just to list a few factors. 10 year treasury rate is considered the benchmark of mortgage rates. I therefore include 10-year Treasury rates as the benchmark long-term interest rate. I acquire the 10 year treasury rate from U.S. Department of the Treasury, dating back to January 1, 1963 to March 1, 2008. [1]

A preliminary analysis of the variables in Table 1 shows a very high correlation among Total Sales and Not Started (91%) and Total Sales and Under Construction (93%), while it is only moderately high between Total Sales and Completed (65%). On the contrary, the total sales does not seem to be affected by 10-year treasury as much (-40%). The correlation can also be observed from Figure 1.

For the simplicity of the project, I merge the Completed and Under Construction as Started, as opposed to Not Started. From a statistical standpoint, I am interested in Not Started because it seems to capture the overall fluctuation well. 10-year Treasury is then omitted, but it is worth researching for a more comprehensive model. In R, the

[1] http://www.treasury.gov/resource-center/data-chart-center/interest-rates/Pages/TextView.aspx?data=yield

Variables	Correlation
TO & NO	0.915
TO & CO	0.651
TO & UN	0.934
TO & i10	-0.399
NO & ST	0.707

Table 1: Correlation

final 3 variables - Total Sales, Not Started, and Started - are coded as TO, NS, and ST, respectively. Note that NS and ST are also highly correlated (71%.)

In Figure 2, all three time series shows left skewed distributions in the Q-Q plots comparing to the theoretical normal line, indicating that the residual after fitting a normal distribution does not behave randomly. The Box-Pierce tests of the three variables are showing p-value close to zero, which means the time effect is significant for all series.

```
> Box.test(tsto)

Box-Pierce test

data:  tsto
X-squared = 473.7546, df = 1, p-value < 2.2e-16

> Box.test(tsno)

Box-Pierce test

data:  tsno
X-squared = 496.334, df = 1, p-value < 2.2e-16

> Box.test(tsst)

Box-Pierce test

data:  tsst
X-squared = 429.511, df = 1, p-value < 2.2e-16
```

3 Analysis

3.1 Correlograms

Correlograms and partial autocorrelations are presented in Figure 3, 4, and 5. All three of them have significant lag effects over the span of two years. Note that each bar represents the autocorrelation of one month. The seasonal fluctuations are observable in these graphs: the sales peak in January and bottom in July. With all the correlations

Variable	Model	AIC	Log Likelihood
TO	AR(2)	3589.50	-1790.75
NO	AR(2)	2654.63	-1323.32
ST	AR(1)	3257.33	-1625.67

Table 2: Model Selection

above the critical values (the blue dashed lines), the time effect is strong. The pattern repeats every 12 periods, which meets my expectation that housing sales are seasonal.

After removing linear independence, I obtain partial autocorrelograms of the three variables. The seasonal effect is still observable for the first few lags.

I include the cross-correlations to study the pairwise relationships among three time series and the relationship between Total Sales and interest rate. The lagged correlations over plus and minus two years are revealed in Figure 6. The positive lag effect among three sales slowly dies down but still stay above the critical value. Seasonal fluctuation is easily observable here. Interest rate on the lower-left corner of Figure 6 has a negative effect since the ACF lies below the 0 benchmark. The lower absolute ACF values and smoothed curve suggest less autocorrelation and seasonality comparing to those among the other three housing sales.

3.2 Model Selection

For each variable, I fit eight different ARIMA models: AR(1), AR(2), MA(1), ARMA(1,1), ARI(1,1), ARI(2,1), IMA(1,1), and ARIMA(1,1,1). R generates several goodness-of-fit criteria, including log-likelihood and Akaike information criterion:

$$AIC = -2ln(loglikelihood) + 2K$$

where K is the number of free parameter. Higher likelihood and lower number of free parameter (lower penalty) are preferable criteria. This implies that the smaller the AIC, the better model fits. With model simplicity in mind, I choose AIC because it penalizes extra parameters. Larger likelihood is emphasized without losing the balance towards potential over-fitting.

From the AIC scores, the best model out of 8 options for each variables are summarized in Table 2. Total Sales and Not Started Sales are best modeled by 2-terms autocorrelation while Started are best captured by 1-term autocorrelation. It is slightly counterintuitive from an individual standpoint: it is rare that a purchase this month will trigger one next month. It is doubtful that individuals with moderate income can afford two houses in consecutive months. This might be better explained by investment, though. A rise housing demand raise the price, and a higher property value might attract more investors due to higher rate of return.

A more detailed view on the coefficient shows that TO and NO have $\phi_1 > 1$ and a negative ϕ_2 while ST has $\phi_1 < 1$. Although the estimated 95% confidence intervals raises my concern: the significancy of ϕ_2 in NO, 0 falls into the 95% confidence interval. In terms of hypothesis test, I cannot reject the null hypothesis and claim that this is an

	Coefficient	S.E.	2.5%	97.5%
ϕ_1	1.0465	0.0426	0.9630806	1.12993698
ϕ_2	-0.1204	0.0426	-0.2039897	-0.03684711

Table 3: Fitting AR(2) of TO

	Coefficient	S.E.	2.5%	97.5%
ϕ_1	1.0143	0.0428	0.9304365	1.09821435
ϕ_2	-0.0615	0.0428	-0.1454500	0.02248411

Table 4: Fitting AR(2) of NO

AR(2) instead of AR(1). Nevertheless, AR(2) reflects a better log-likelihood as well as smaller AIC scores.

In Figure 7, 8, and 9, I take first difference for all three time series in (a) and comparisons of original versus detrended graph in (b). Differencing does not eliminate the seasonal fluctuations, so other model design might be needed to capture the random components.

In (b), I detrend the time series by regressing the series over time and plot the residuals. The upward dashed lines are flattened and are shown as dashed lines. However, the detrending is slightly observable for Total Sales but not for Not Started and Started.

3.3 'Best Fit' and Seasonality

I then utilize the `auto.arima()` function in search of the 'best fit'. R generates the desired output by allowing more parameters and detect seasonality. Table 6 summarizes the result and compare AICs for both seasonal and non-seasonal AICs. Clearly, the complicated models demonstrate superiority.

For Total Sales, R uggests ARIMA(2,1,4) model with seasonal AR(1) and MA(2) of 12 months period. AIC value 3233 is the smallest AIC among all the models I fit. The parameters of ARIMA(2,1,4) are shown in Table 9. Their 95% confidence interval are estimated in Table 10. Despite the statistical significancy, the non-seasonal model alone is difficult to explain: for the first difference, there is an permanent effect of 2 periods and temporary effect for 4 periods. Same challenge lies in the interpretation of NO and ST.

	Coefficient	S.E.	2.5%	97.5%
ϕ_1	0.8894	0.0194	0.8513505	0.9273963

Table 5: Fitting AR(1) of ST

Variable	Seasonal Model	Seasonal AIC	Non-Seasonal AIC
TO	ARIMA(2,1,4)(1,0,2)[12]	3233.44	3589.50
NO	ARIMA(3,1,4)(2,0,0)[12]	2444.10	2654.63
ST	ARIMA(5,1,1)(1,0,2)[12]	2943.19	3257.33

Table 6: Model Selection: auto.arima()

	ϕ_1	ϕ_2	θ_1	θ_2	θ_3	θ_4	Φ_1	Θ_1	Θ_2
Coefficient	- 0.1419	-0.5480	-0.0862	0.4914	-0.1714	-0.1468	0.9871	-0.6861	-0.1172
S.E.	0.1268	0.0874	0.1278	0.0827	0.0429	0.0464	0.0101	0.0472	0.0443

Table 7: Best Fit of TO - ARIMA(2,1,4)(1,0,2)

	ϕ_1	ϕ_2	ϕ_3	θ_1	θ_2	θ_3	θ_4	Φ_1	Φ_2
Coefficient	0.4812	0.5193	-0.6029	-0.6739	-0.5695	0.7747	-0.1965	0.3800	0.2810
S.E.	0.1438	0.0953	0.1037	0.1465	0.1101	0.1280	0.0701	0.0448	0.0436

Table 8: Best Fit of NO - ARIMA(3,1,4)(2,0,0)

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	θ_1	Φ_1	Θ_1	Θ_2
Coefficient	-0.9594	-0.3801	-0.2230	-0.1576	-0.1239	0.6833	0.9870	-0.7283	-0.1271
S.E.	0.1961	0.0857	0.0676	0.0581	0.0499	0.2043	0.0075	0.0433	0.0422

Table 9: Best Fit of ST - ARIMA(5,1,1)(1,0,2)

	2.5 %	97.5 %
ϕ_1	-0.3903844	0.10667266
ϕ_2	-0.7192424	-0.37676723
θ_1	-0.3366312	0.16428790
θ_2	0.3293093	0.65343703
θ_3	-0.2555484	-0.08728077
θ_4	-0.2378219	-0.05585379
Φ_1	0.9673379	1.00681726
Θ_1	-0.7785762	-0.59369039
Θ_2	-0.2040221	-0.03046210

Table 10: 95% CI of ARIMA(2,1,4)(1,0,2) : TO

	2.5 %	97.5 %
ϕ_1	0.1993971	0.76296440
ϕ_2	0.3325591	0.70607633
ϕ_3	-0.8061097	-0.39971800
θ_1	-0.9611182	-0.38667222
θ_2	-0.7853569	-0.35373310
θ_3	0.5238807	1.02560318
θ_4	-0.3338350	-0.05915535
Φ_1	0.2921994	0.46785117
Φ_2	0.1954599	0.36653448

Table 11: 95% CI of ARIMA(3,1,4)(2,0,0) : NO

	2.5 %	97.5 %
ϕ_1	-1.3437015	-0.57514995
ϕ_2	-0.5480570	-0.21204488
ϕ_3	-0.3554864	-0.09058013
ϕ_4	-0.2714431	-0.04373513
ϕ_5	-0.2216991	-0.02610741
θ_1	0.2829646	1.08362039
Φ_1	0.9722973	1.00167813
Θ_1	-0.8131099	-0.64349299
Θ_2	-0.2097842	-0.04437002

Table 12: 95% CI of ARIMA(5,1,1)(1,0,2) : ST

	Oct '07	Nov'07	Dec'07	Jan'08	Feb'08	Mar'08
AR(2)	71.2	70.1	69.1	68.2	67.3	66.6
ARIMA	63.1	54.8	53.8	57.6	64.7	78.3
Actual	57	45	44	44	47	51

Table 13: Prediction vs Actual: TO

	Apr'07	May'07	Jun'07	Jul'07	Aug'07	Sep'07
AR(2)	79.8	78.2	76.5	75.0	73.6	72.3
ARIMA	74.3	78.3	75.7	70.1	71.6	63.1
Actual	83	79	73	68	60	53

Table 14: Prediction vs Actual: TO

3.4 Goodness of Fit

Model diagnostic for AR(2) and ARIMA(2,1,4)(1,0,2) models of Total Sales are shown in Figure 10. ARIMA(2,0,4)(1,0,2) outperform because the ACF residuals are all below the critical values, while those of AR(2) are still significant for several lags.

This is also evidenced by the Ljung-Box statistics: for ARIMA(2,0,4)(1,0,2), the p-values are high among all lags, which is the lack of proof for significant lag effects. The same statistics shows low values for all lags except for lag 0 in AR(2), implying that the lag effects are statistically different from 0. Similar patterns can be seen in 11 and 12 except that AR(2) removes the significancy of lag 2 for Not Started. This resulted in a high p-value for lag 2 Ljung-Box statistics in the third graph of Figure 11(a).

4 Forecasting

I remove a year worth of data from April 2007 to March 2008, and I project forward 12 months and compares the result against actuals in Table 13 to 16. The forecasting performance is also evidenced by graphs. In Figure 13, the prediction captured the seasonality, as appeared in the upward shape. However, the overall actual sales are lower almost consistently. In reality, it was when subprime mortgage crisis stroke the housing market.

In Figure 13, the prediction of Total Sales with AR(2) and ARIMA(2,1,4)(1,0,2) are presented side-by-side. Judging from the gaps between actual and predicted lines, I

	Oct '07	Nov'07	Dec'07	Jan'08	Feb'08	Mar'08
AR(2)	21.1	20.9	20.6	20.4	20.2	20.0
ARIMA	14.6	11.6	11.4	13.0	12.0	15.5
Actual	12	9	10	10	11	14

Table 15: Prediction vs Actual: NO

	Apr'07	May'07	Jun'07	Jul'07	Aug'07	Sep'07
AR(2)	22.8	22.5	22.2	21.9	21.6	21.4
ARIMA	20.1	19.1	17.6	17.7	19.5	14.7
Actual	22	20	18	15	14	11

Table 16: Prediction vs Actual: NO

	Oct '07	Nov'07	Dec'07	Jan'08	Feb'08	Mar'08
AR(1)	48.5	47.7	47.1	46.5	46.0	45.5
ARIMA	49.5	45.3	44.1	43.7	48.5	56.9
Actual	45	37	33	35	36	36

Table 17: Prediction vs Actual: ST

believe the more complexed ARIMA(2,1,4)(1,0,2) model capture the trend better than AR(2). Although, the upward trend in prediction widens the gap in the final months. Similar conclusion can be drawn for Not Started except that the gap does not seem to be widened towards the end of ARIMA(3,1,4)(2,0,0). That leaves me more confident with forecasting on Not Started over Total Sales. Future sales might be better modeled by Not Started.

Started sales forecast is more ambiguous in terms of the trade-off between model complexity and pattern fitting. The ARIMA(5,1,1)(1,0,2) demonstrates a wide difference towards the end, while the flat AR(1) seems to have a moderate gap without a complicated model.

5 Conclusion

Overall, the ARIMAs capture the seasonality of all three series to a certain extend. ARIMA(3,1,4)(2,0,0) on Not Started Sales appears to be more accurate than the other two variables. Moving average, lagged regression, and differences are applied in the end result.

Seasonality effect is captured by second difference. The graph of the final output has a very similar shape comparing to the actual numbers.

	Apr'07	May'07	Jun'07	Jul'07	Aug'07	Sep'07
AR(1)	55.3	53.8	52.5	51.3	50.2	49.3
ARIMA	54.4	58.5	57.5	53.2	54.1	49.8
Actual	61	59	55	53	47	43

Table 18: Prediction vs Actual: ST

5.1 Potential Improvements

- Log-transformation

  Box-Cox plots in Figure 16 shows that λ lies between (0,1). They suggest that log-transformation might be feasible. However, it is not attempted due to the difficulty of interpretation. It can be difficult to justify the underlying assumption of multiplicativity.

- Recasting

  In practice, the difference between theory and actual can be taken into consideration. The behavior of the residuals can be carefully studied and further contributed to refine the model. In practice, it is also acceptable to assign certain credibility factors in reflection of the plausibility of data over models.

- 10-year treasury rate

  Although the 10-year treasury rate is omitted in this project, a negative 40% correlation cannot be ignored in reality. The causality needs further verification, but there is little doubt that a crucial macroeconomic variable like interest rate impacts the housing market.

- Overfitting

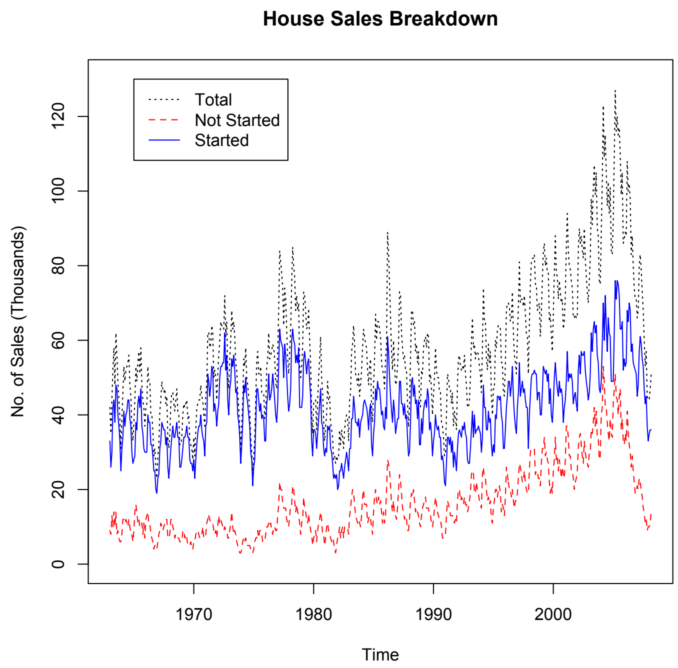
  The ARIMA models generated by R are appealing because of the goodness of fit. However, it is still challenging to explain the large amount of parameters in a meaningful way. Over-fitting might exists, so further verification is required before an extensive use of model. It is particularly true for a huge investment like housing sales.

References

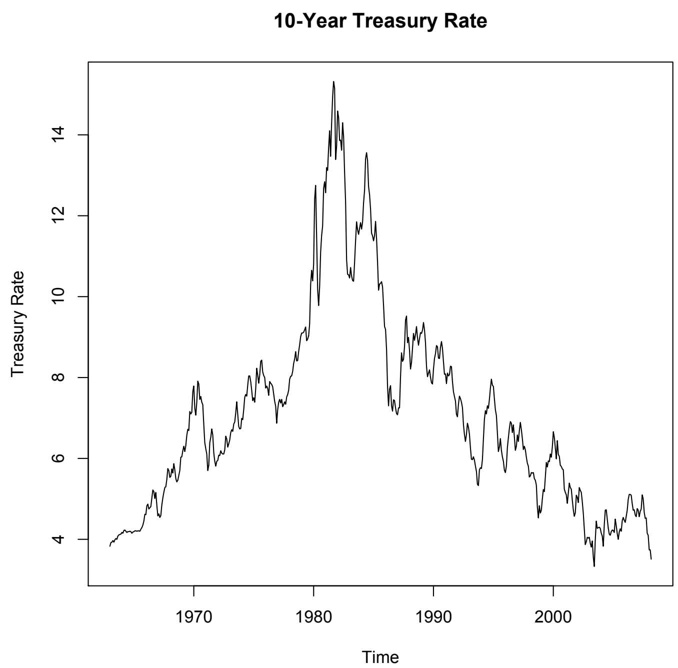
[1] Jonathan D. Cryer and Kung-Sik Chan *Time Series Analysis - With Applications in R, 2nd Ed* Taylor & Springer 2008.

[2] Paul S.P. Cowpertwait and Andrew V. Metcalfe *Introductory Time Series with R* Taylor & Springer 2009.

[3] Paul Teetor *R Cookbook* O'Reilly 2011.

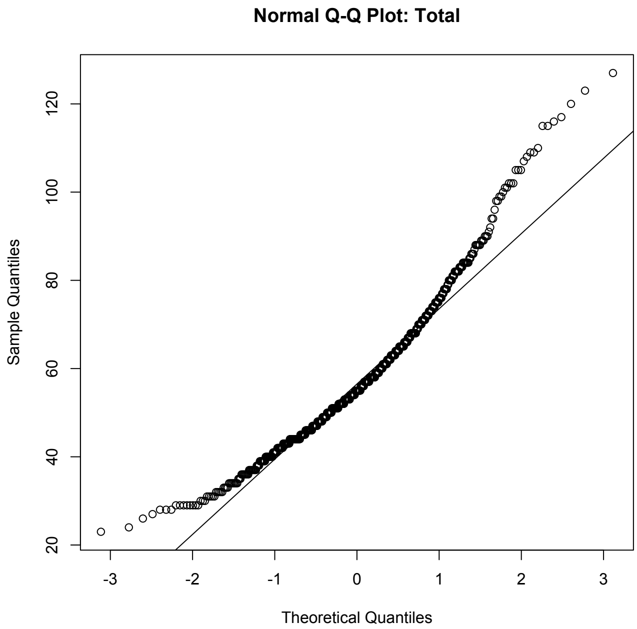


(a)

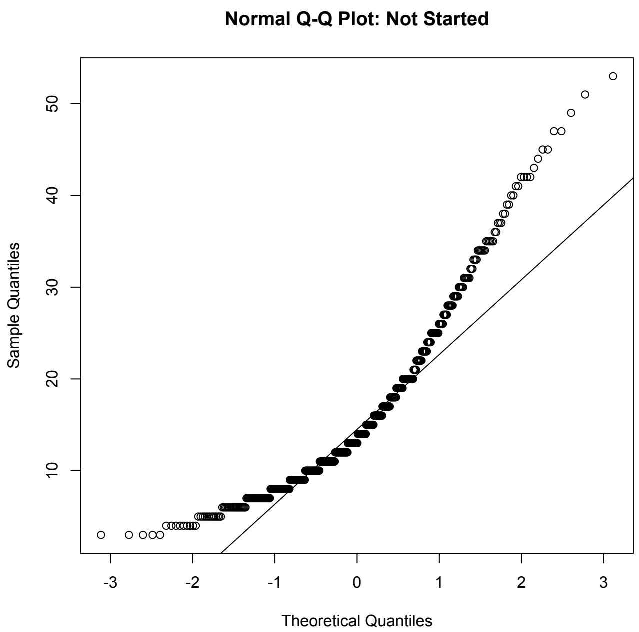


(b)

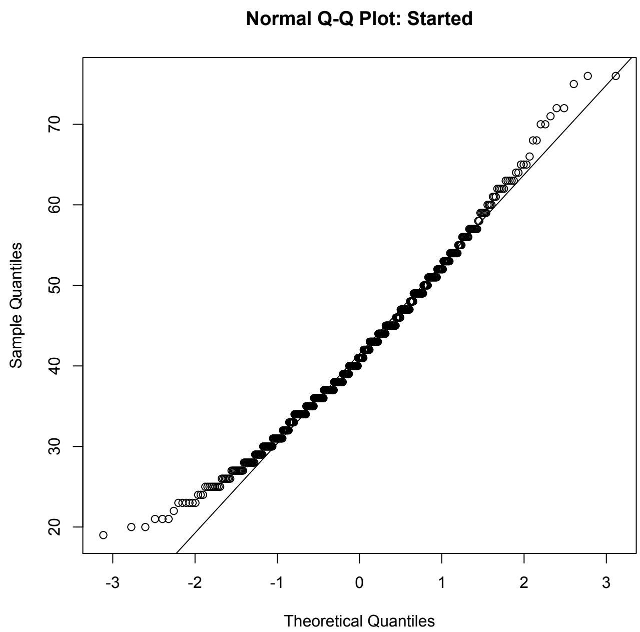
Figure 1: Housing Sales Breakdown versus 10-Year Treasury Rates: 1963-2008



(a)



(b)



(c)

Figure 2: Q-Q Plots: (a) Total, (b) Not Started, and (c) Started

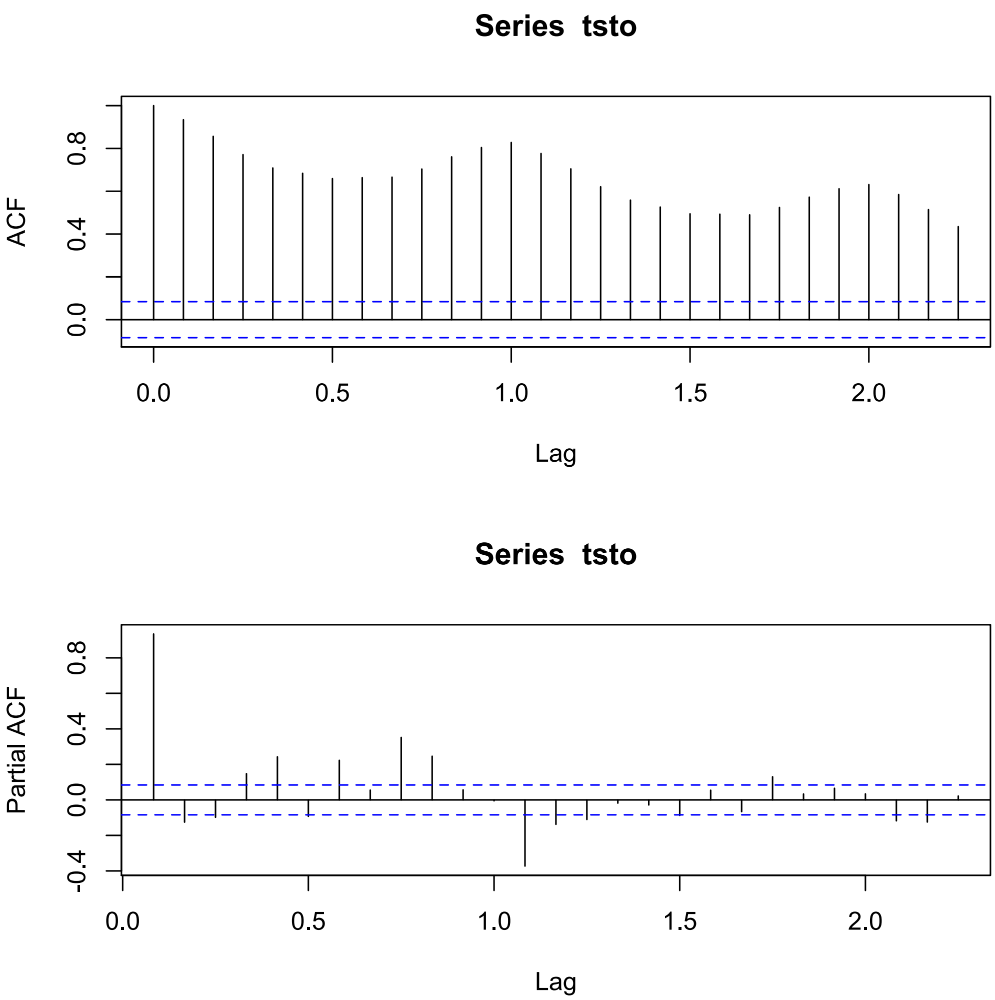


Figure 3: Autocorrelation and Partial Autocorrelation: Total

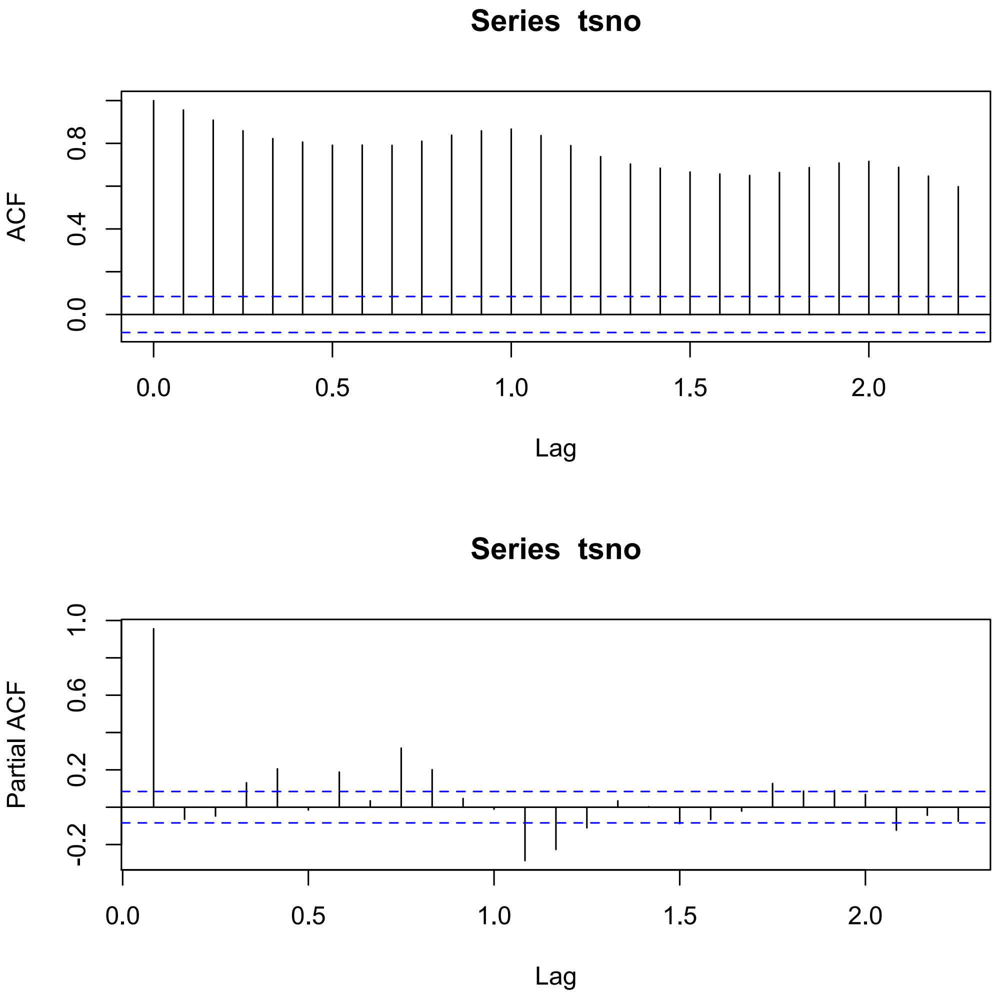


Figure 4: Autocorrelation and Partial Autocorrelation: Not Started

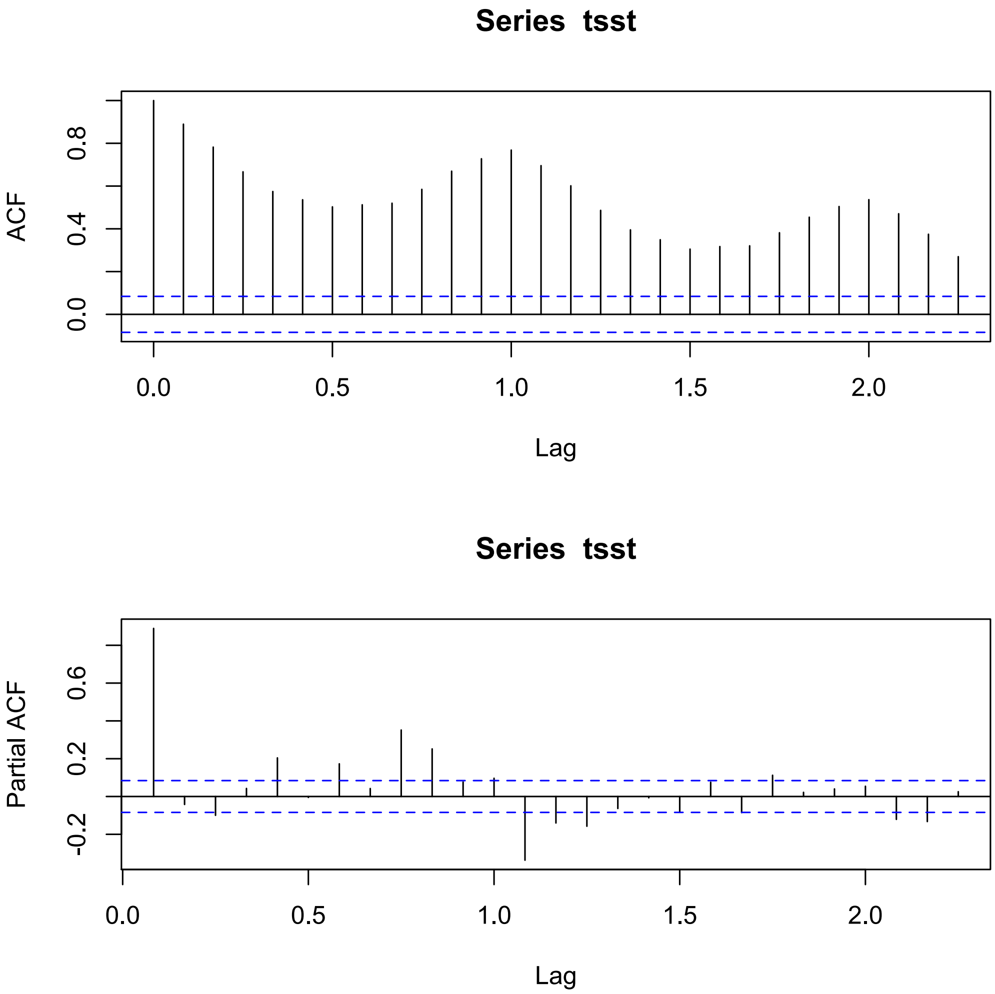


Figure 5: Autocorrelation and Partial Autocorrelation: Not Started

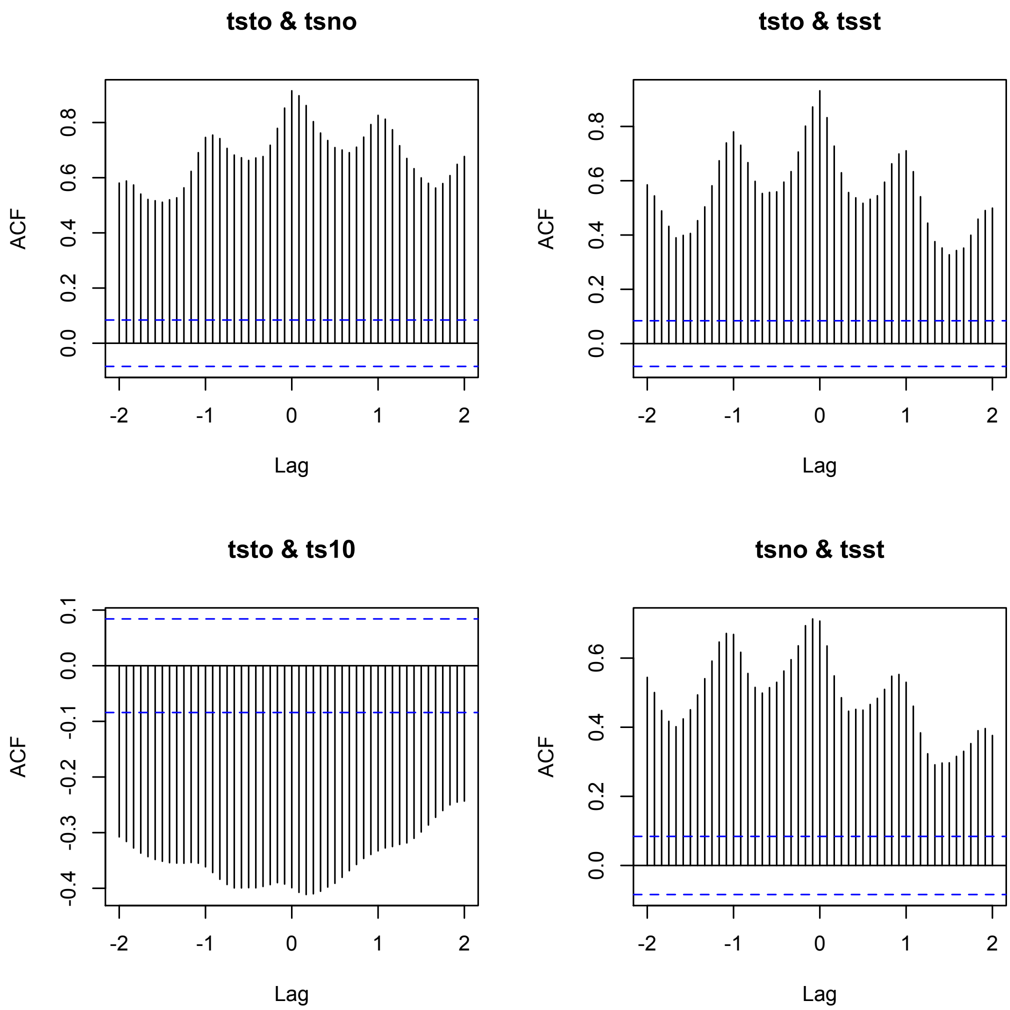
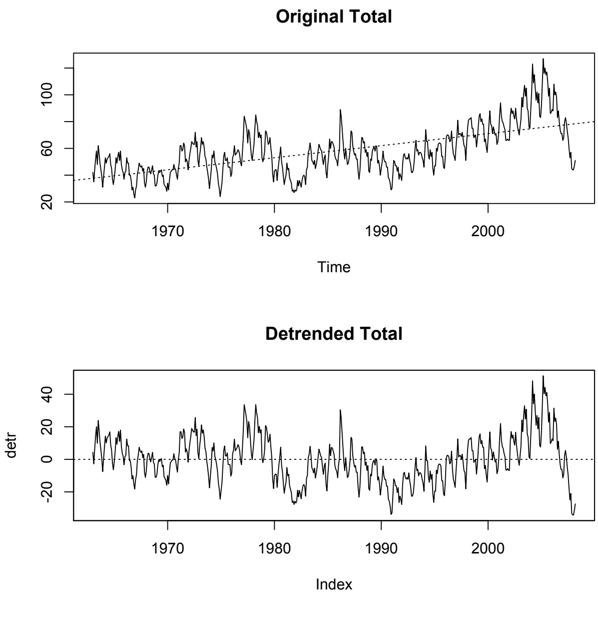


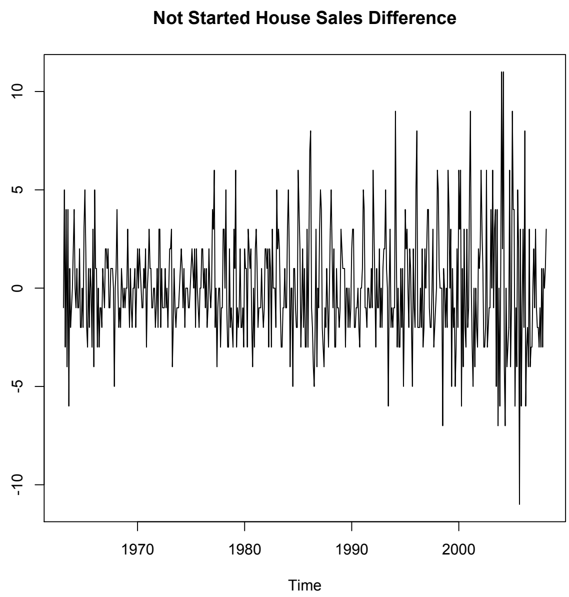
Figure 6: Lagged Correlations

(a)

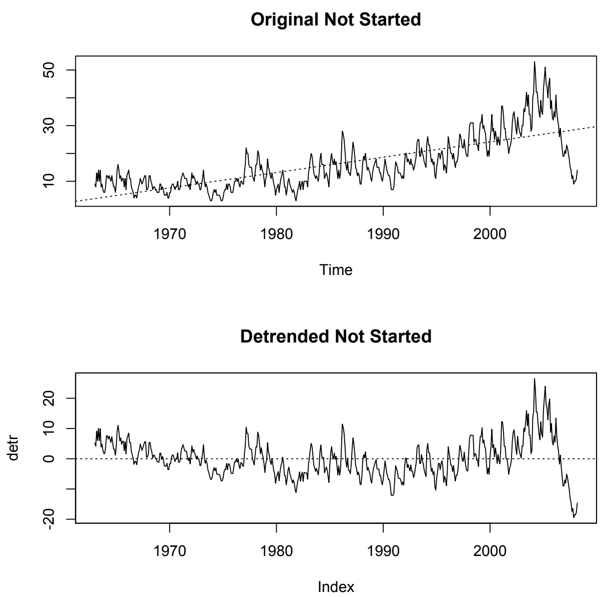


(b)

Figure 7: (a) First Difference and (b) Detrending: Total Sales



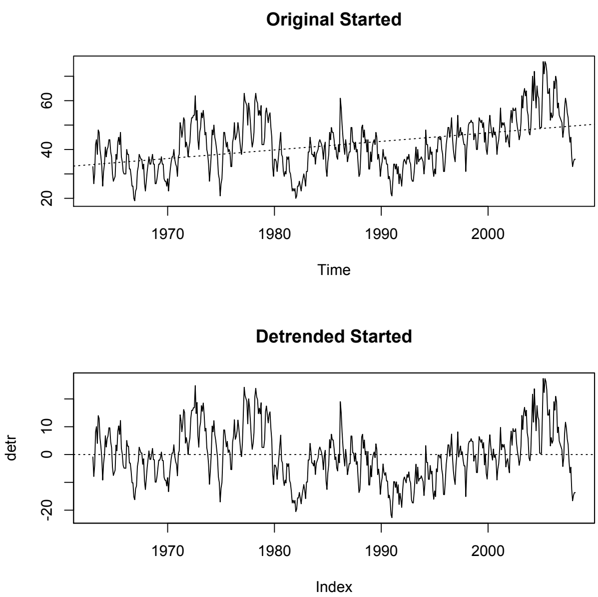
(a)



(b)

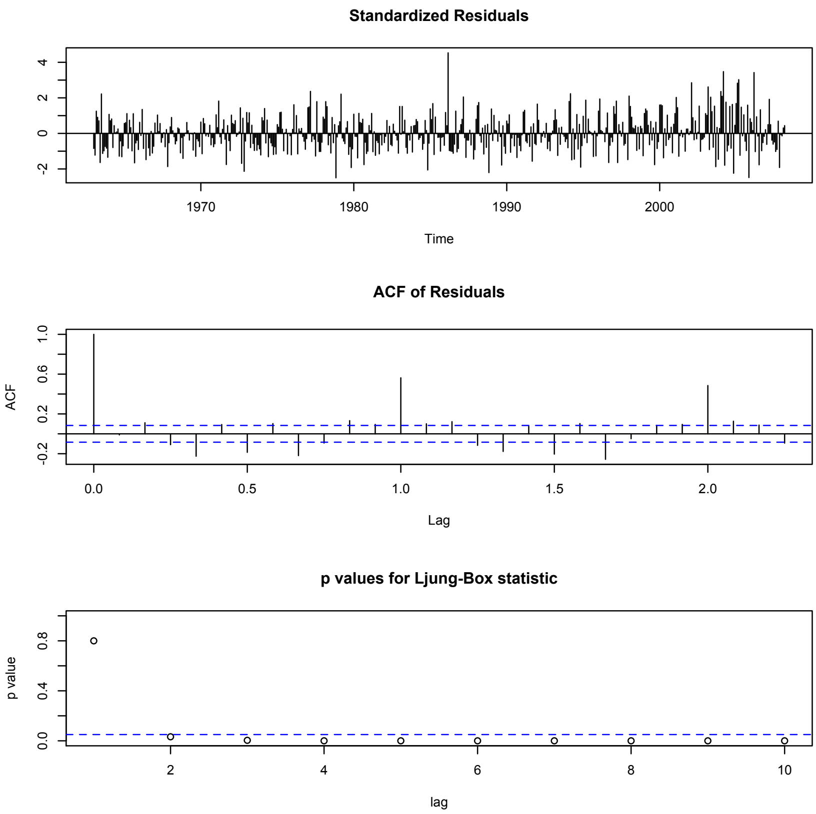
Figure 8: (a) First Difference and (b) Detrending: Not Started

(a)

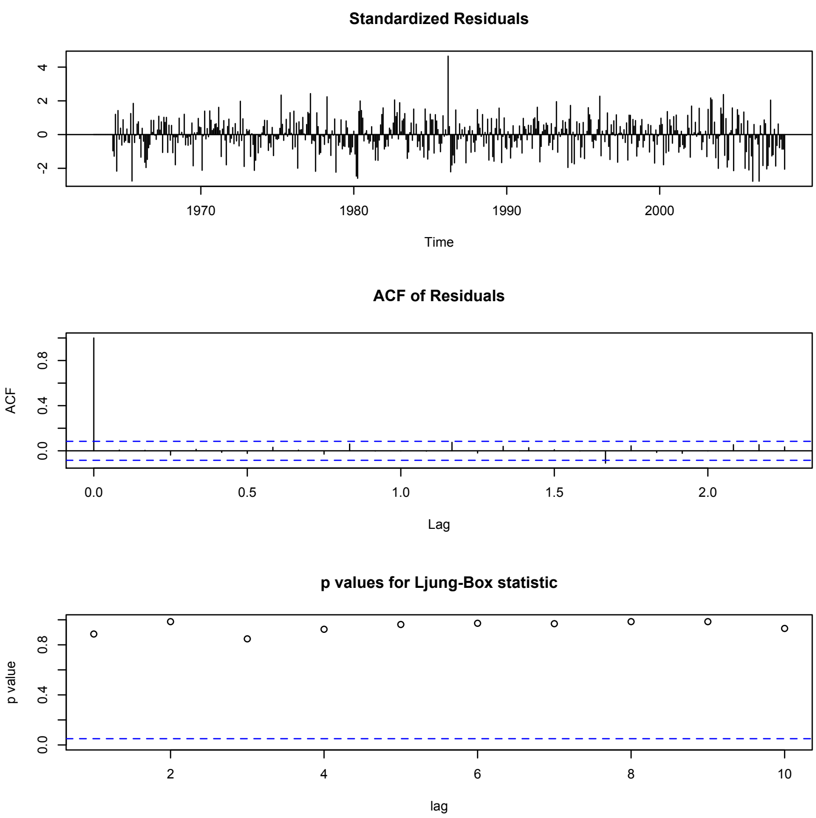


(b)

Figure 9: (a) First Difference and (b) Detrending: Started

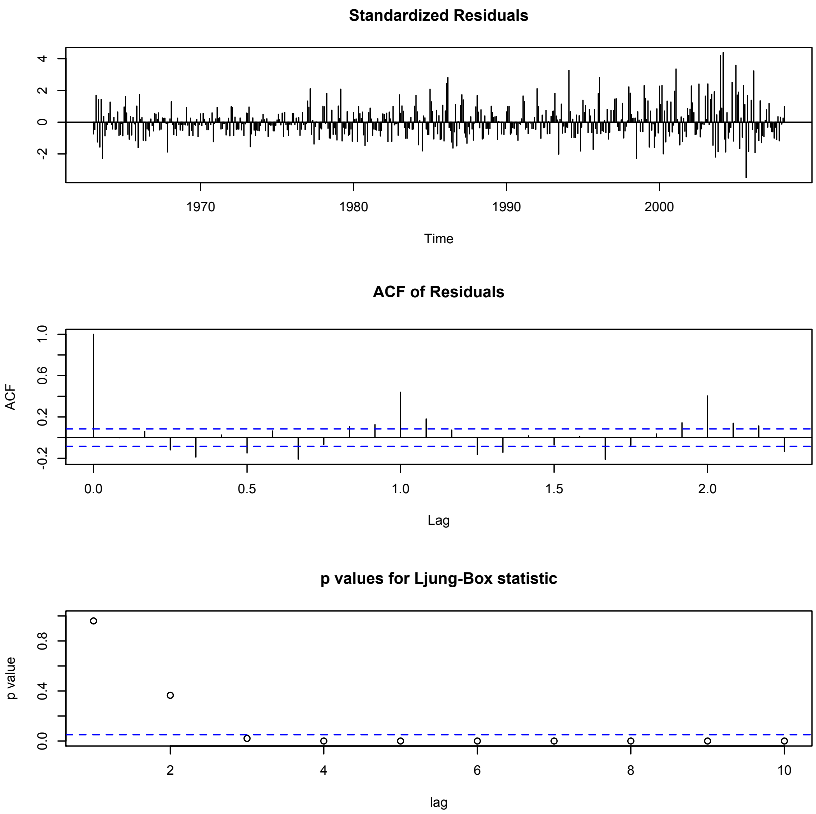


(a)

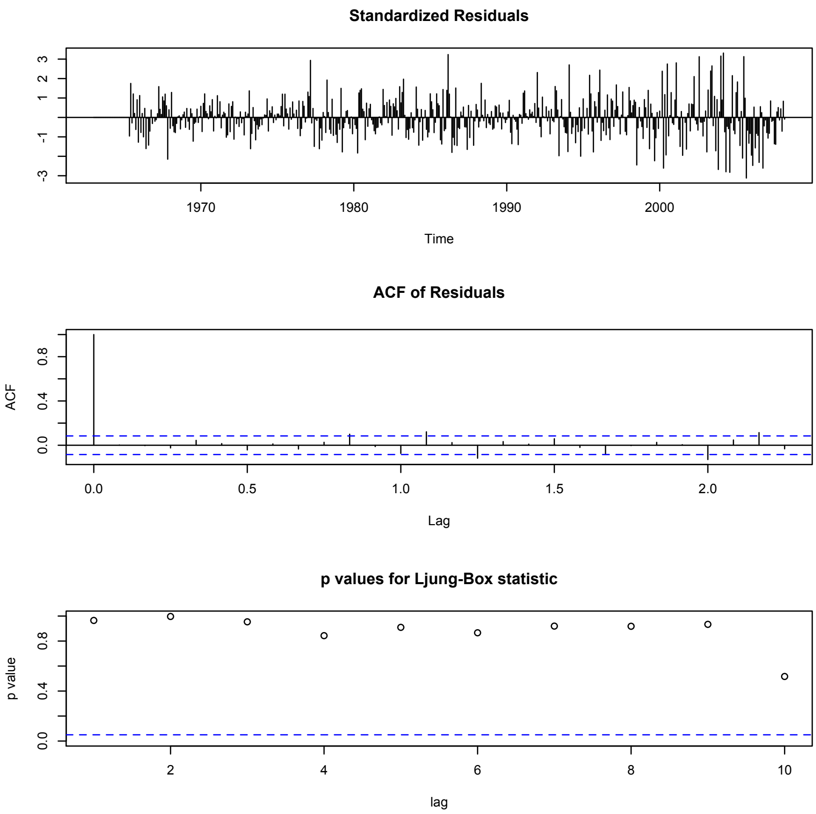


(b)

Figure 10: Diagnostics for Total Sales (a) AR(2) and (b) ARIMA(2,1,4)

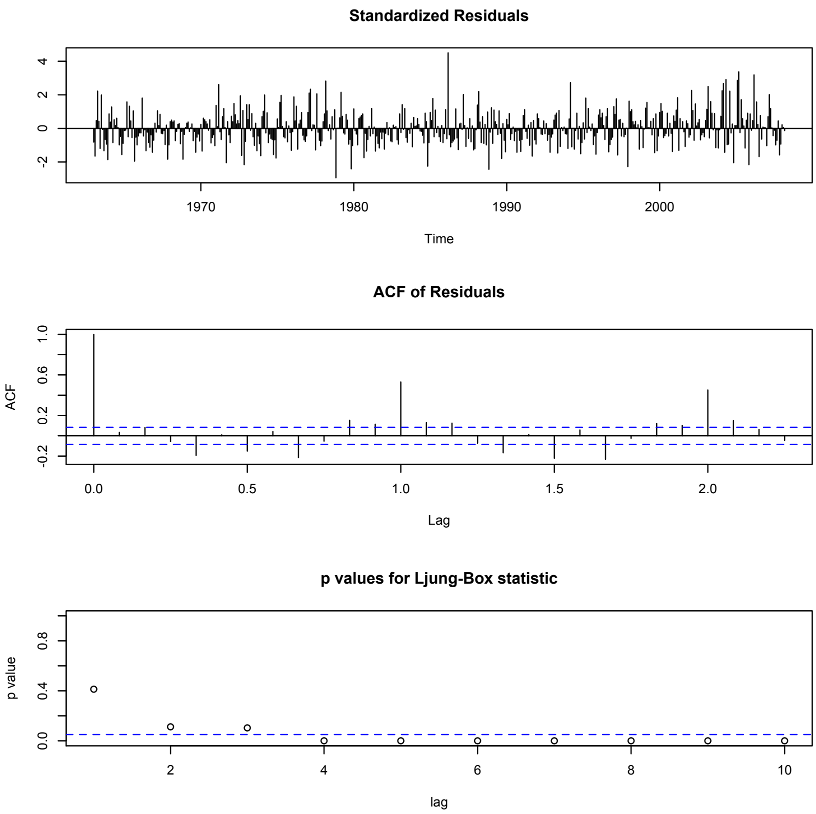


(a)

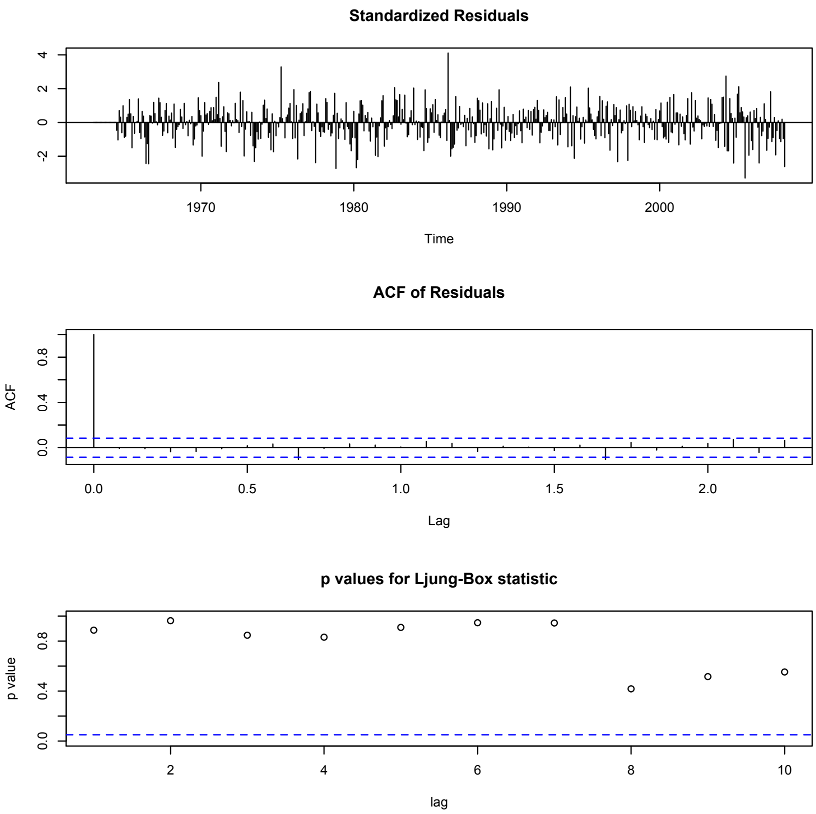


(b)

Figure 11: Diagnostics for Not Started Sales (a) AR(2) and (b) ARIMA(3,1,4)

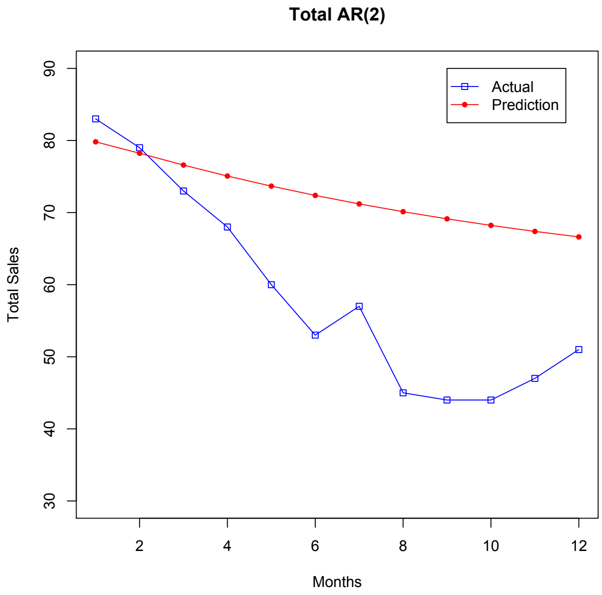


(a)

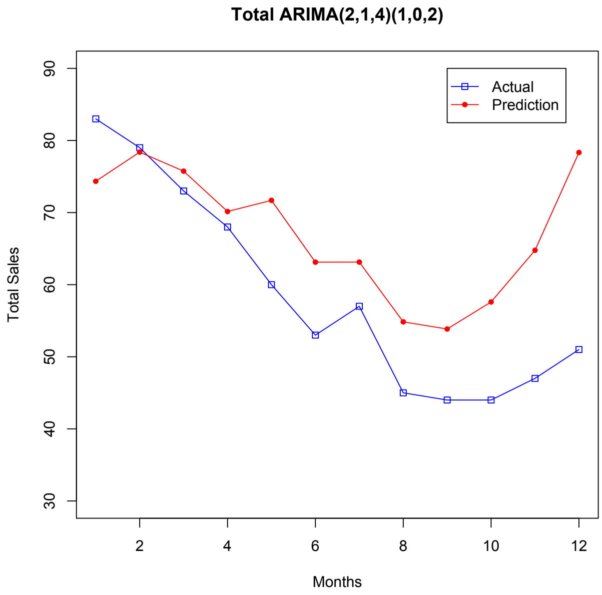


(b)

Figure 12: Diagnostics for Started Sales (a) AR(1) and (b) ARIMA(5,1,1)

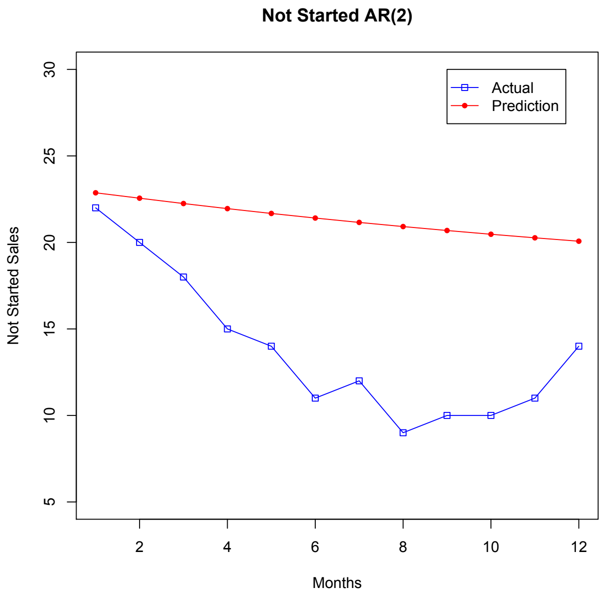


(a)

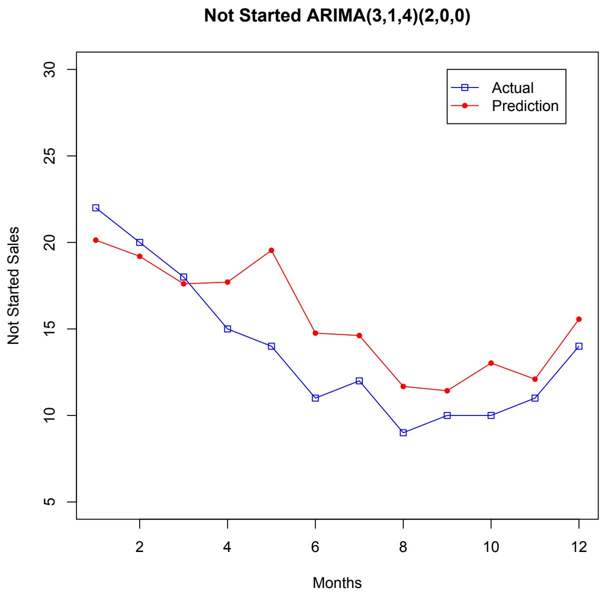


(b)

Figure 13: Prediction and Actual Series: Total

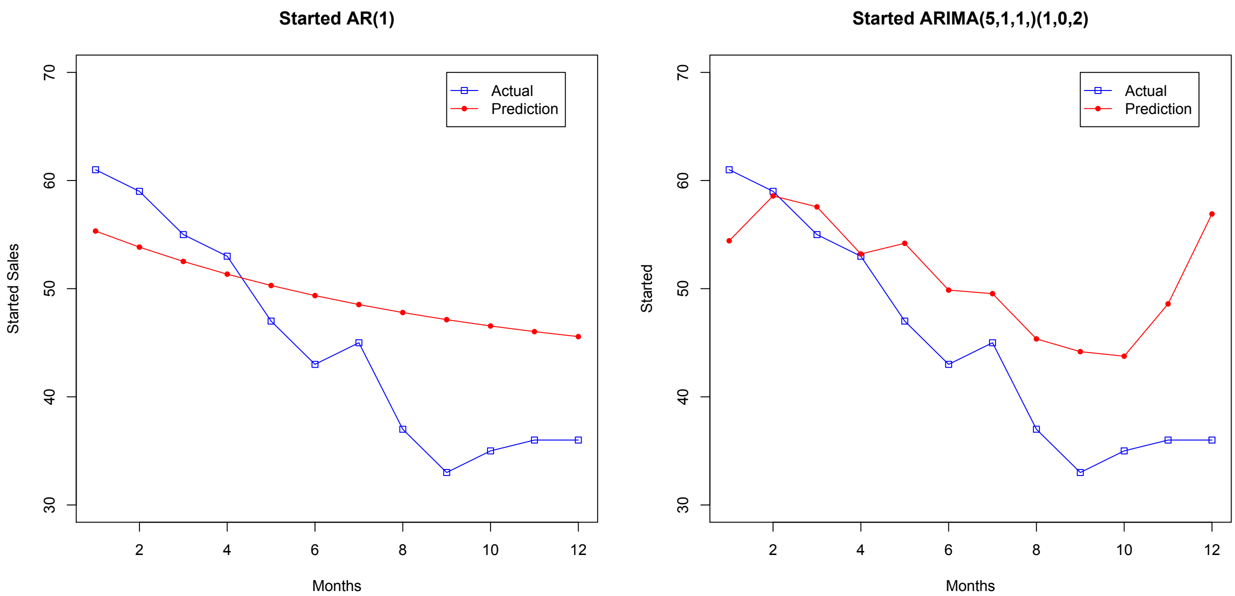


(a)



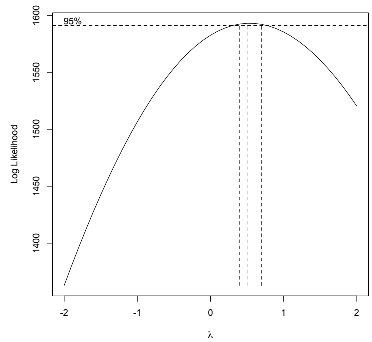
(b)

Figure 14: Prediction and Actual Series: Not Started

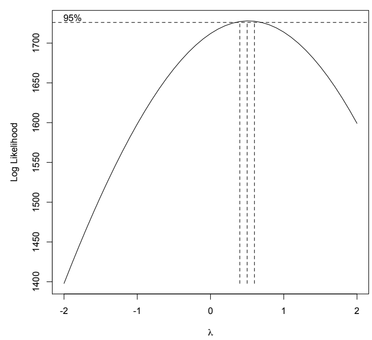


(a) (b)

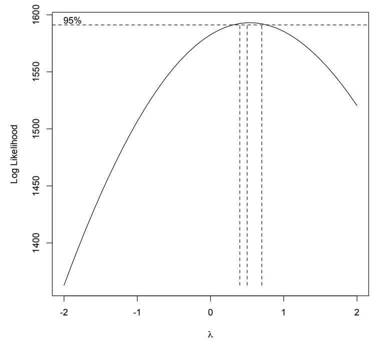
Figure 15: Prediction and Actual Series: Started



(a)



(b)



(c)

Figure 16: Box-Cox Tests for (a) Total, (b) Not Started, and (c) Started