Time Series Analysis

Midterm Project

Chenyin Gao*

(Dept. Mathematics School Major Statistics)

1 Sunspot

1.1 Exploitation Data Analysis

Figure 1 shows the yearly Sunspot data from 1700 to 1984, and we note the repetitive nature of the value and rather regular periodicities. Modeling such series begins by observing the primary patterns in the time history.

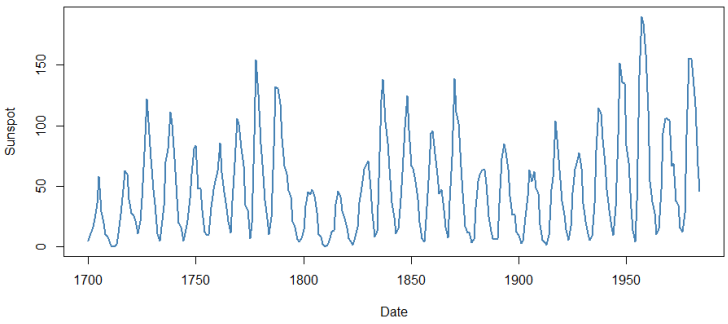

Figure 1. Yealy Sunspot from 1700 to 1984

First, we conduct exploitation data analysis to see if necessary transformation is required. We define the *sample ACFs function* as,

$$\hat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-h} (x_{t+h} - \bar{x})(x_t - \bar{x}) \tag{1}$$

with $\hat{\gamma}(-h) = \hat{\gamma}(h)$ for $h = 0, 1, \cdots, n-1$

*gaochy5@mail2.sysu.edu.cn

Compute the sample ACFs as previous definition and shows the ACFs of the sunspot in Figure 2. Since the original series appears to contain a sequence of repeating short signals, the ACF confirms this behavior showing the repeating peaks.

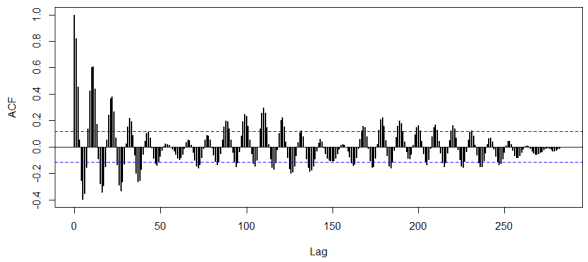

Figure 2. ACFs of the sunspot

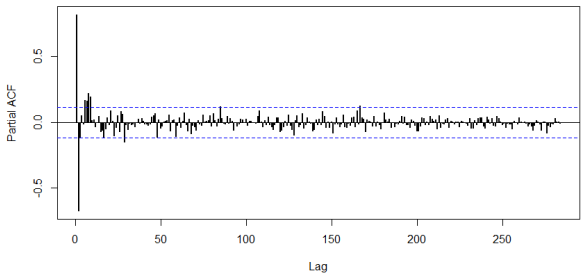

Figure 3. PACFs of the sunspot

The Figure 2 and Figure 3 show patterns which are consistent with the original series. The ACF has cycles corresponding roughly to a 11-year period and the PACF has large values for $h = 1,2$ and then is significantly zero for higher order lags. These results preliminarily suggest that a second-order AR(2) autoregressive model might provide a good fit. As

$$x_t = (\phi_1 + \phi_2)x_{t-1} - \phi_2(x_{t-1} - x_{t-2}) + w_t \tag{2}$$

Subtract x_{t-1} from both sides in (2)

$$\nabla x_t = (\phi_1 + \phi_2 - 1)x_{t-1} - \phi_2(x_{t-1} - x_{t-2}) + w_t = \gamma x_{t-1} - \phi_2 \nabla x_{t-1} \tag{3}$$

To test the hypothesis whether the Φ has a unit root at , we can test $H_0 : \gamma = 0, H_1 :\neq> 00$ by estimating γ in the regression of ∇x_t on x_{t-1} ,∇x_{t-1} and forming a Wald test based on $t = \frac{\hat{\gamma}}{se(\gamma)}$

Table 1: Preset AR(2) Model Causality Test

	Estimate	Std. Error	t value	Pr(>|t|)
x_{t-1}	0.112	0.018	6.272	0.000
∇x_{t-1}	0.478	0.048	9.973	0.000

In Table 1, we denote that $\hat{\gamma}$ is significantly nonzero and therefore the original sunspots data is stationary. Besides, we implemented augmented Dickey-Fuller test(ADF) at lag order $(M_1 - 1)^{\frac{1}{3}}$, say 6, and its p-value yields 0.01 which also confirm the stationarity of the sunspot series.

Augmented Dickey-Fuller Test

data: Sunspots

Dickey-Fuller = -4.7656, Lag order = 6,
p-value = 0.01
alternative hypothesis: stationary

1.2 Model Selection

Since we have observed the obvious seasonal pattern, we implement the *seasonal autoregressive moving average model*, say, ARMA$(P,Q)_s$, the model form is

$$\Phi_P(B^s)x_t = \Theta_Q(B_s)w_t \tag{4}$$

where the operators are

$$\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_p B^{Ps} \tag{5}$$

and

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Ps} \tag{6}$$

Usually, we combine the seasonal and nonseasonal operators into a mixed model, denoted by ARMA$(p,q)\times(P,Q)_s$, and write as,

$$\Phi_p(B^s)\phi(B)x_t = \Theta_Q(B^s)\theta(B)w_t \tag{7}$$

In this sunspot case, we assume the cycle of $s = 11$ years in sunspot(See in reference).

In Figure 2,Figure 3, the ACF tails off at points $= 1s, 2s, \cdots$, with repeated peaks and PACF cuts off after lag s, these results imply the model SAR(1),P=1,Q=0, in the seasonal model. However, after inspecting the ACF and PACF in a lower lag, it appears they both are tails off, suggesting an ARMA$(p,q)(0 < p,q < s)$ model within the seasonal model.

Thus, we preform fit ergodic models, $ARMA(p,q)\times(1,0)_{11}, 0 < p < 11, 0 < q < 11$, with autocorrelated errors in (8).

$$\begin{aligned}
&ARMA(0,1)\times(1,0)_{11} \Rightarrow x_t = \delta + \Phi x_{t-11} + w_t + \theta w_{t-1}\\
&ARMA(1,0)\times(1,0)_{11} \Rightarrow x_t = \delta + \Phi x_{t-11} + w_t + \phi x_{t-1}\\
&\cdots\\
&ARMA(p,q)\times(1,0)_{11} \Rightarrow x_t = \delta + \Phi x_{t-11} + w_t + \sum_{i=1}^{p}\phi_i x_{t-p} + \sum_{j=1}^{q}\theta_j w_{t-q}
\end{aligned} \tag{8}$$

To select optimal p and q, We computed AIC, BIC, ESS and R^2 in its according model and compared its value.

The definition of Akaike' s Information Criterion (AIC)[3], Bayesian Information Criterion (BIC)[2], Error Sum of Squares(ESS) and R square(R_2) are denoted in (9),

$$
\begin{aligned}
AIC &= \log \hat{\sigma}_p^2 + \frac{n+2p}{n} \\
BIC &= \log \hat{\sigma}_p^2 + \frac{p \log n}{n} \\
\hat{\sigma}_p^2 &= \frac{1}{n-(p+1)} \sum_{t=1}^{n} (x_t - \sum_{i=1}^{p} \hat{\phi}_i x_{t-i})^2 \\
ESS &= \sum_{t=1}^{n} (x_t - \hat{x}_t)^2 = \sum_{t=1}^{n} [x_t - (\hat{\delta} + \hat{\Phi} x_{t-11} + w_t + \sum_{i=1}^{p} \hat{\phi}_i x_{t-p} + \sum_{j=1}^{q} \hat{\theta}_j w_{t-q})]^2 \\
SS &= \sum_{t=1}^{n} (x_t - \bar{x}_t)^2 = \sum_{t=1}^{n} (x_t - \frac{1}{n} \sum_{i=1}^{n} x_i)^2 \\
R^2 &= 1 - \frac{ESS/(n-(p+q+2))}{SS/(n-1)}
\end{aligned}
\tag{9}
$$

In Figure 4, we summarized the previous statistics values in each respective model with the optimal lag framed by red square.

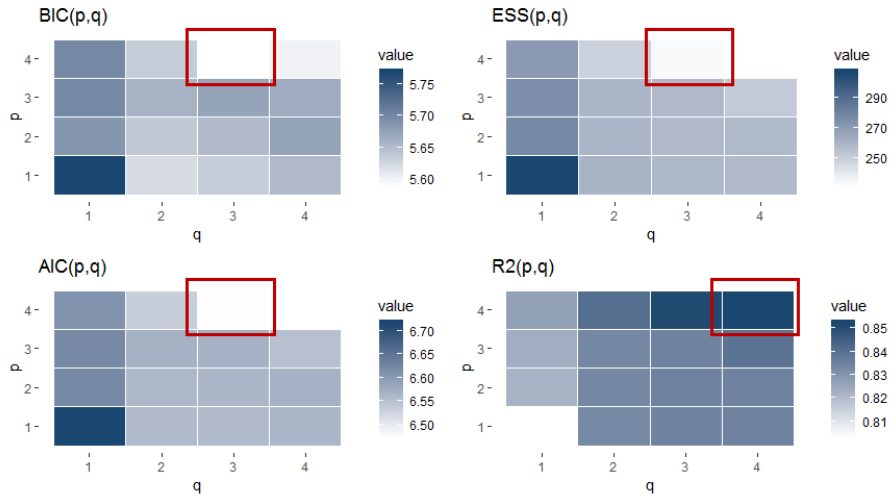

Figure 4. Statistics values for all $1 \leq p,q \leq 4$

Most information criteria prefer the $ARIMA(3,0,4) \times (1,0,0)_{11}$ model, which is displayed in (10)

$$
(1-\Phi_1 B^{11})(1-\sum_{i=1}^{3} \phi_i B^i) x_t = (1+\sum_{j=1}^{4} \theta_j B^j) w_t \tag{10}
$$

Besides, we also show the residual diagnosis and estimated coefficients in Figure 5 and Table 2). In Figure 5, except for one or two outliers, the model fit quite well showing no obvious departure of the residuals from whiteness.

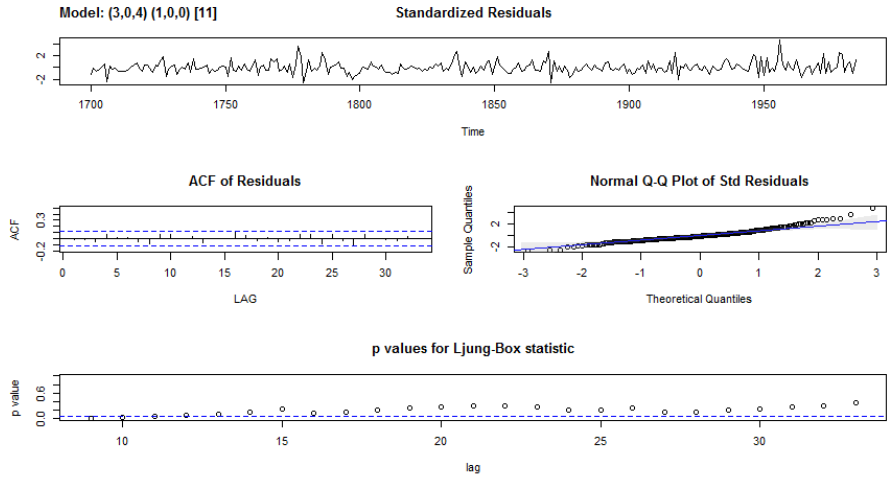

Figure 5. Residual analysis for the $ARMA(3,0,4) \times (1,0,0)_{11}$ fit to the sunspot data set

Furthermore, in Table 2, most estimators show significant p-value except for the θ_3 with light-blue highlight. In Figure 6, we plot our model prediction with the original data to assess the accuracy. Our forecast and the original data are matched up with trend, minimum and maximum.

Table 2: Estimated Coefficients

	Estimate	SE	t.value	p.value
ϕ_1	2.473	0.067	36.743	0.000
ϕ_2	-2.311	0.112	-20.669	0.000
ϕ_3	0.793	0.066	12.100	0.000
θ_1	-1.289	0.092	-14.043	0.000
θ_2	0.361	0.103	3.510	0.000
θ_3	0.054	0.096	0.559	0.576
θ_4	0.121	0.078	1.546	0.123
Φ_1	0.101	0.070	1.447	0.149
δ	48.966	5.326	9.194	0.000

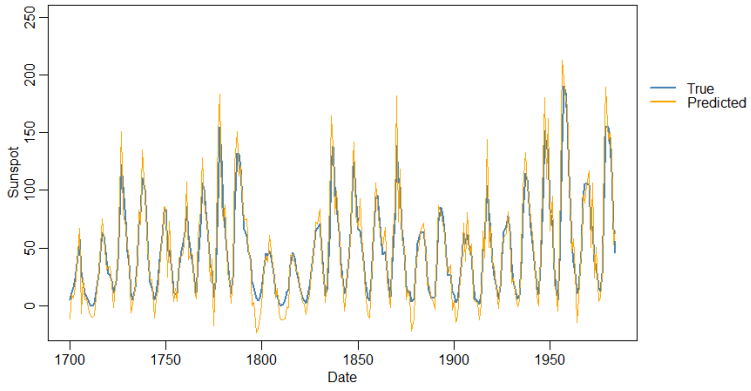

Figure 6. Real data and predicted data for $\text{ARMA}(3,0,4) \times (1,0,0)_{11}$ for the sunspot data set

1.3 Forecast and Interpretation

The details of the selected model is below in (11)

$$
\begin{aligned}
x_t = & \underbrace{48.966}_{\sigma} \\
& + \underbrace{0.101 x_{t-11}}_{\Phi_1} \\
& + \underbrace{2.473 x_{t-1} - 2.311 x_{t-2} + 0.793 x_{t-3}}_{\phi_{1,2,3}} \\
& \underbrace{-1.289 w_{t-1} + 0.361 w_{t-2} + 0.054 w_{t-3} + 0.121 w_{t-4}}_{\theta_{1,2,3,4}}
\end{aligned}
\tag{11}
$$

Lastly, we forecast the sunspots out 4 years(1985-1988),and the results are shown in Figure 7(only the last 100 observations are plotted in the graphic) with its coefficients, standard deviation and 95% Confidence Interval in Table 3

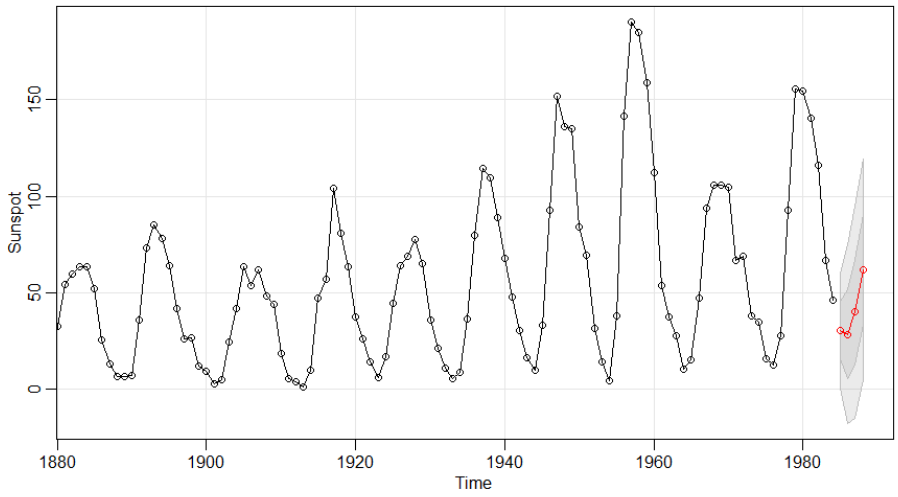

Figure 7. Forecast out 4 years for ARMA$(3,0,4) \times (1,0,0)_{11}$ for the sunspot data set

Table 3: True and Prediction sunspot(1985-1988) with s.d. and 95%CI

Year	True	Forecast	S.d.	Up Bounds	Low Bounds
1985	17.900	30.528	14.958	59.845	1.210
1986	13.400	28.472	23.189	73.921	-16.978
1987	29.200	39.963	27.434	93.733	-13.808
1988	100.200	62.022	28.570	118.019	6.026

In Figure 8, we compare our model prediction with the real data in *sunspot2.dat*

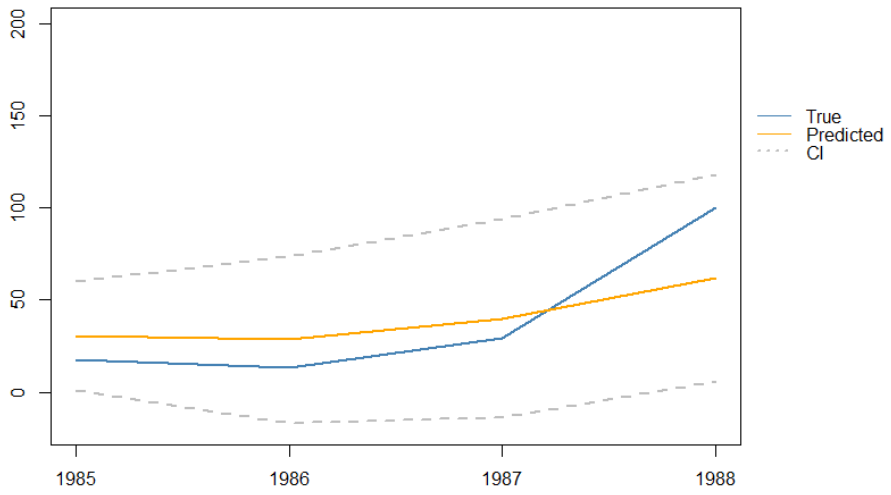

Figure 8. Forecast out 4 years for ARMA$(3,0,4) \times (1,0,0)_{11}$ compared with Real data

From our above comparison, we confirm the validity of our model in the following aspects

1. The trend of our forecast is matched with the true data series.
2. However, it appears little departure in 1988. All true data lie within the 95% Confidence Interval of our forecast which means $(\hat{\mu}_t - 1.96\hat{\sigma}_t, \hat{\mu}_t + 1.96\hat{\sigma}_t)$ could cover the true data with 0.95 probability.

2 3.35

2.1 (a)

Discuss our model fitting in a step-by-step fashion

2.1.1 Initial Examination

First, we plot the sales, S_t, data set in Figure 9 for overall view.

2.1 (a)

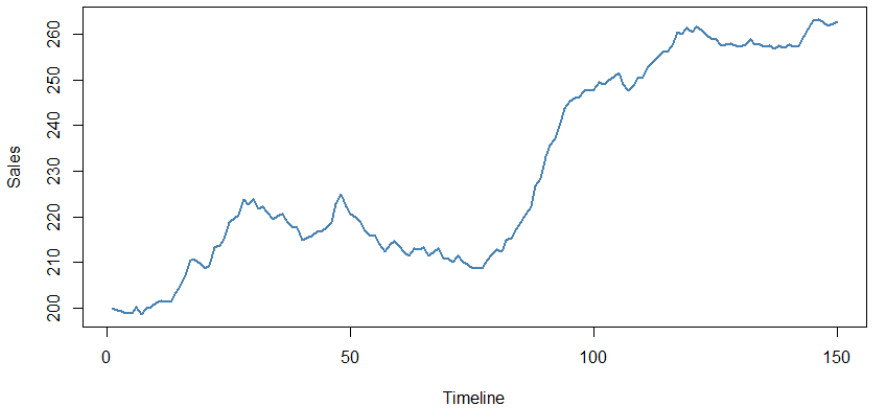

Figure 9. Plot of monthly Sales S_t data set

In Figure 9, it is pretty obvious that the original data set is non-stationary since it keeps increasing. Using theoretical confirmation, we conducted the Augmented Dickey-Fuller test, computing the test statistics value of the lag order $(M-1)^{\frac{1}{3}}$, M is length of the data set.

Augmented Dickey-Fuller Test

data: S_t
Dickey-Fuller = -2.1109, Lag order = 5,
p-value = 0.5302
alternative hypothesis: stationary

Since the p-value is 0.5302, we could not reject the null hypothesis and accept that S_t is non-stationary.

2.1.2 Transformation

We suppose the model of S_t as

$$S_t = \mu_t + x_t + w_t \tag{12}$$

where x_t is a stationary process and μ_t denotes the trend, w_t is the white noise.First, we suggest a straight line could be useful for detrending the data

$$\begin{aligned} S_t &= \mu_t + x_t + w_t \\ S_t &= \beta t + \delta + x_t + w_t \end{aligned} \tag{13}$$

2.1 (a)

We could write the regression model as (14)

$$S_t - \hat{\beta}t - \hat{\delta} = \hat{x}_t \tag{14}$$

In this case, we estimate the trend using ordinary least square and found

$$\hat{\mu}_t = 0.447t + 196.232 \tag{15}$$

And then we simply subtract $\hat{\mu}_t$ from S_t to obtain the detrended time series x_t. In Figure 10, we show the detrend time series plot with its according ACF and PACF.

$$\hat{x}_t = S_t - \hat{\beta}t - \hat{\delta} = S_t - 0.447t - 196.232 \tag{16}$$

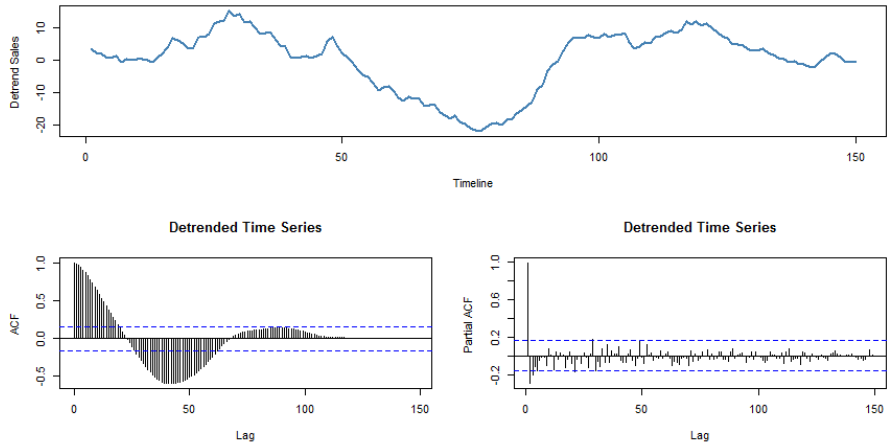

Figure 10. Plot of Detrended Sales with its ACF and PACF

Besides, we also implemented ADF unit test of lag order 5. The results are shown as,

Augmented Dickey-Fuller Test

data: x_t
Dickey-Fuller = -2.1109, Lag order = 5,
p-value = 0.5302
alternative hypothesis: stationary

However, the results show the detrended time series is still non-stationary. So we turn to the *difference operator*

$$S_t - S_{t-1} = (\mu_t + x_t + w_t) - (\mu_{t-1} + x_{t-1} + w_{t-1}) = \delta + w_t - w_{t-1} + x_t - x_{t-1} \tag{17}$$

2.1 (a)

Assume the model trend as a stochastic component using the random walk with drift model $\mu_t = \delta + \mu_{t-1} + w_t$.

Since we assume that x_t is stationary process, $x_t - x_{t-1}$ is also stationary process. Further, we denote $S_t - S_{t-1}$ as ∇S_t. Equivalently, we demonstrate the transformed series with its ACF and PACF in Figure 11 and conduct ADF test to derive the specific p-value to see whether it reject the explosive null hypothesis.

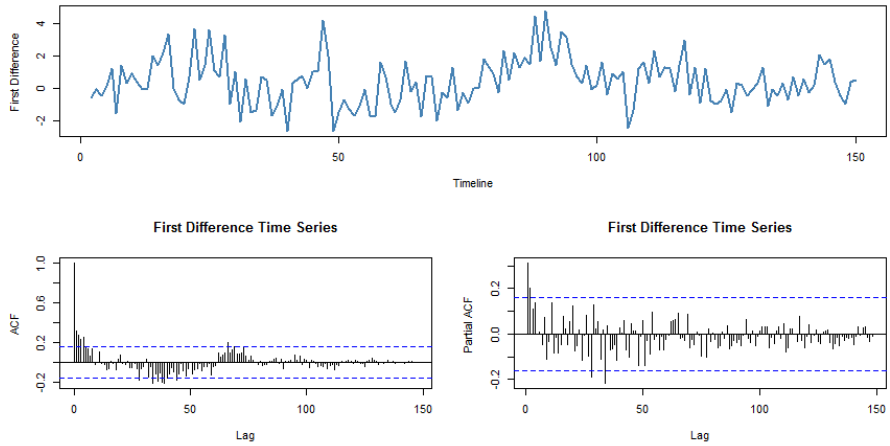

Figure 11. Plot of First Difference Sales with its ACF and PACF

Augmented Dickey-Fuller Test

data: ∇S_t
Dickey-Fuller = -3.3485, Lag order = 5,
p-value = 0.06585
alternative hypothesis: stationary

Combined the plots of ACF and PACF with the ADF test results, we perceived the ∇x_t as a stationary process

2.1.3 Model Selection

Since the ACF and PACF of ∇S_t both tail off, so we assume the fit model as $ARMA(p,q), 1 \leq p,q \leq 4$ model for ∇S_t, denoted in (18)

$$\nabla S_t = \sum_{i=1}^{p} \phi_i \nabla S_{t-i} + \sum_{j=1}^{q} \theta_j w_{t-j} + w_t \tag{18}$$

2.1 (a)

Computing the same statistical values mentioned in subsection 1.2 and, inFigure 12, select p and q with optimal lags framed by red square.

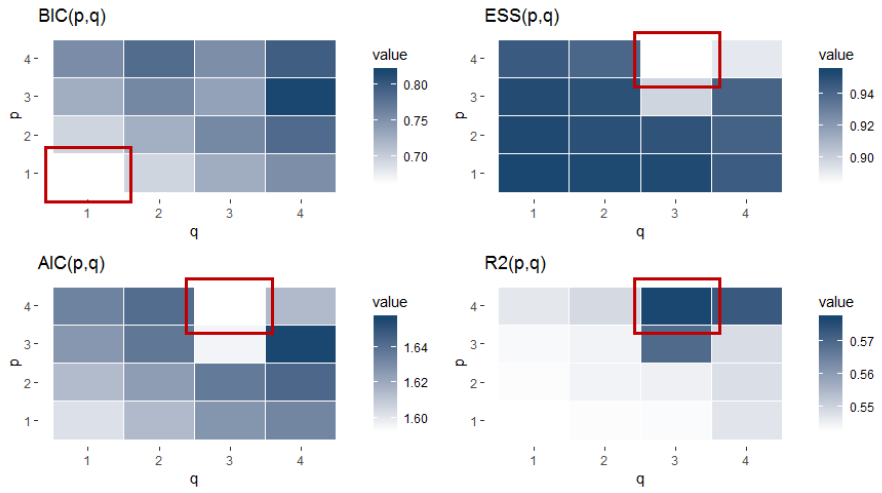

Figure 12. Statistics Values For ARMA$(p,q), 1 \leq p,q \leq 4$

Based on Figure 12, we choose $ARMA(4,3)$ as our fit model in this case.

2.1.4 Residual Diagnostics

We perform regression with autocorrelated errors of $ARIMA(4,1,3)$ for the non-stationary process S_t and the estimated $ARIMAR(4,1,3)$ model is shown in (19)

$$\begin{aligned}\nabla x_t = &\underbrace{0.1416\nabla x_{t-1} + 0.4652\nabla x_{t-2} - 0.0669\nabla x_{t-3} + 0.1409\nabla x_{t-4}}_{\phi_{1,2,3,4}} + \\ &\underbrace{0.0864 w_{t-1} - 0.2912 w_{t-2} + 0.1061 w_{t-3}}_{\theta_{1,2,3}}\end{aligned} \tag{19}$$

Figure 13 displays a plot of the standardized residuals, the ACF of the residuals, a quantile-quantile plot of the standardized residuals, and the p-values associated with the Q-statistic* at lags H = 8 through H = 20 (with corresponding degrees of freedom H -7)

*$Q = n(n+2)\sum_{h=1}^{H} \frac{\hat{\rho}_e^2(h)}{n-h}$

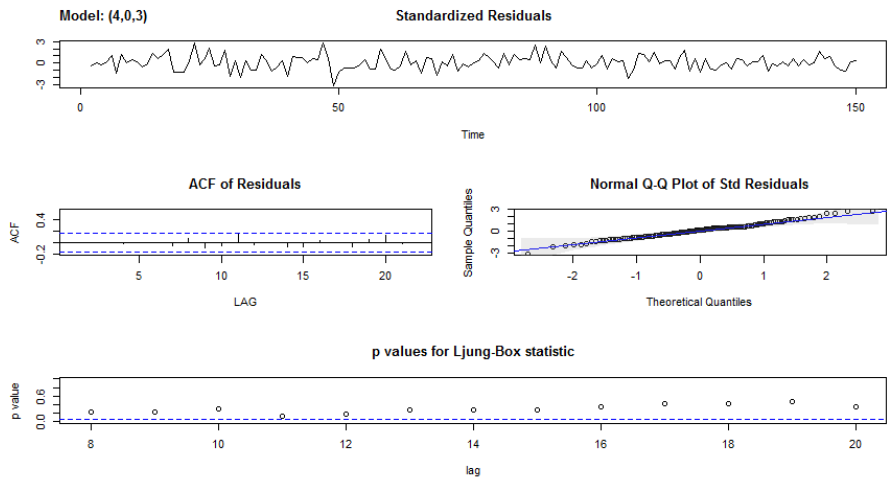

Figure 13. Diagnosis of the residual of ARIMA(4,1,3) fit on the sales data set

After inspecting the Figure 13, we notice no obvious pattern. All of the standardized residuals is within the interval of [-3,3]. The ACF of the residuals display no apparent departure from the model assumption, and the Q-statistic is never significant at the lags down. The normal Q-Q plot of the residuals shows that the assumption of normality is reasonable, with only exception of one or two.

2.2 (b)

Use the CCF and lag plots between ∇S_t and ∇L_t to argue that a regression of ∇S_t on ∇L_{t-3} is reasonable

Definition 1. *The **cross-correlation function (CCF)** between two series, x_t and y_t , is*

$$\rho_{xy}(s,t) = \frac{\gamma_{xy}(s,t)}{\sqrt{\gamma_x(s,s)\gamma_y(t,t)}} \tag{20}$$

where $\gamma_{xy}(s,t) = \mathbb{E}[(x_s - \mu_{xs})(y_t - \mu_{yt})]$.

Definition 2. *The **sample cross-covariance function** will be written using (20) as*

$$\begin{aligned} \hat{\gamma}_{xy}(h) &= n^{-1}\sum_{t=1}^{n-h}(x_{t+h} - \bar{x})(y_t - \bar{y}) \\ \hat{\rho_{xy}}(h) &= \frac{\hat{\gamma}_{xy}(h)}{\sqrt{\hat{\gamma}_x(0)\hat{\gamma}_y(0)}} \end{aligned} \tag{21}$$

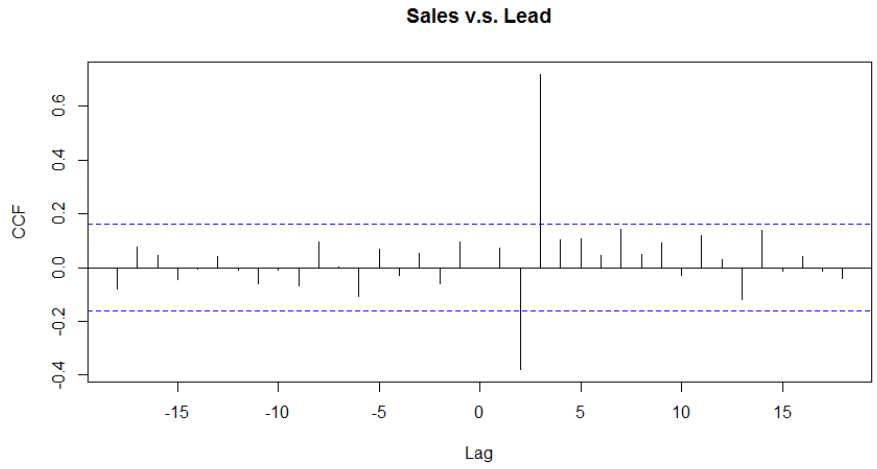

Figure 14. Sample CCF of ∇S_t and ∇L_t ;positive lags indicate sales lags lead

In Figure 14, the sample CCF following (20) shows several departure from the cyclic component of each series and there is an obvious peak at h=3($\hat{\rho_{xy}}(3) = 0.72$). This imply that ∇S_t at time $t+3$ is associated with the ∇L_t series at time t. Equivalently, we could say that the ∇L_t series leads the ∇S_t series by 3 years. Besides, the sign of the CCF is positive, leading to the conclusion that the two series move in the same direction, in another word, the increases in ∇L_t leads to increase in ∇S_t and vice versa.

Then we run lag plots(lowess fits) of ∇S_t on the lag of $\nabla L_{t-i}, i = 1, 2, \cdots, 8$ to assess the meaningfulness of the sample autocorrelations.

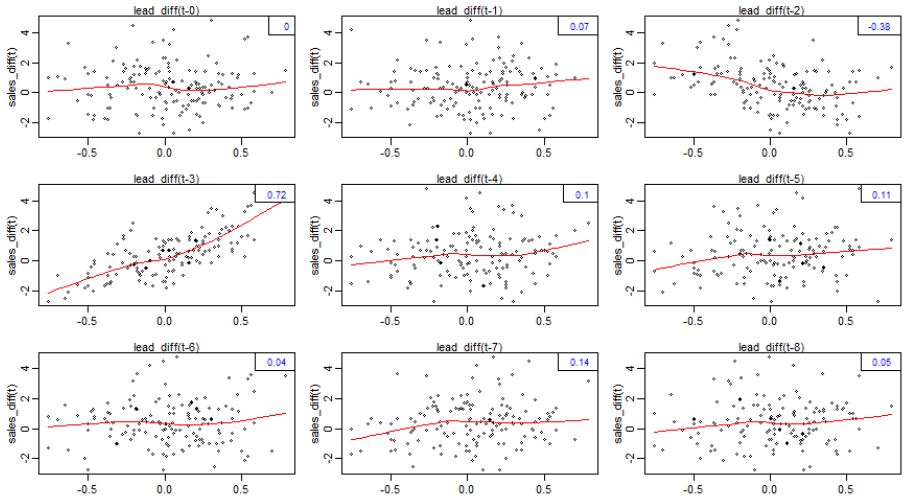

Figure 15. Scatterplot matrix of the ∇S_t and ∇L_t, on the vertical axis plotted against the ∇S_t, and ∇L_t on the horizontal axis at lags h=$0, 1, \cdots, 8$. The valus in the upper right corner are the sample cross-correlations and the lines are a lowess fit

From the Figure 15, we notice that the lowess fits are approximately linear, so that the sample autocorrelations are meaningful. Besides, we see strong positive linear relations at lag=3, that is ∇S_t associated with ∇L_{t-3}, which match up well with the CCF results noticed in Figure 14.

From above discussion, the results indicate the ∇L_t series tend to lead ∇S_t at 3 years and the positive coefficient also implied that the increases in the ∇L_t lead to increases in ∇S_t. Apart from that, the linearity of the lowess fit suggest that the behavior between ∇S_t and ∇L_t is the same both for positive and negative values of ∇L_t

2.3 (c)

In this subsection, we fit a regression model with autocorrelated errors as

$$\nabla S_t = \beta_0 + \beta_1 \nabla L_{t-3} + x_t \tag{22a}$$

We assume the error process x_t is $ARMA(p,q)$, i.e., $\phi(B)x_t = \theta(B)w_t$. Since x_t is stationary, we could transform by $\pi(B)x_t = w_t$, where $\pi(B) = \theta(B)^{-1}\phi(B)$. We then multiply both sides of (22a) by $\pi(B)$ and obtain

$$\pi(B)\nabla S_t = \pi(B)(\beta_0 + \beta_1 \nabla L_{t-3}) + \underbrace{\pi(B)x_t}_{w_t} \tag{22b}$$

We set up our goals as minimizing the error sum of squares

$$S(\phi,\theta,\beta) = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n} [\pi(B)\nabla S_t - \pi(B)(\beta_0 + \beta_1 \nabla L_{t-3})]^2 \tag{22c}$$

An easy way to tackle (22c) was first presented in Cochrane and Orcutt[1]

First, we run the ordinary least regression model of ∇S_t on ∇L_{t-3} with constant and retain the residual as,

$$\hat{x}_t = \nabla S_t - (\hat{\beta}_0 + \hat{\beta}_1 \nabla L_{t-3}) \tag{23}$$

Then we identify the ARMA models for $\hat{x}_t$ based on AIC,BIC,ESS and R^2 in Figure 17,Figure 16. Detailed procedures are mentioned in subsection 1.2.

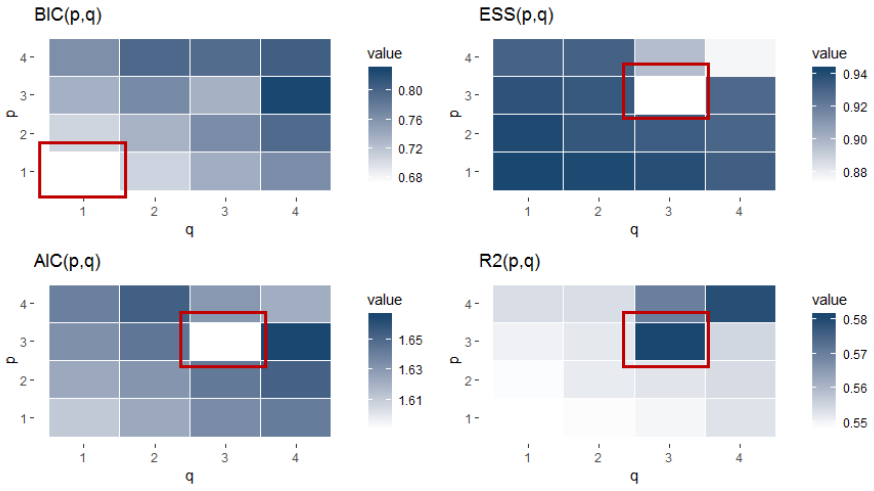

Figure 16. Models Selection of $\hat{x}_t$

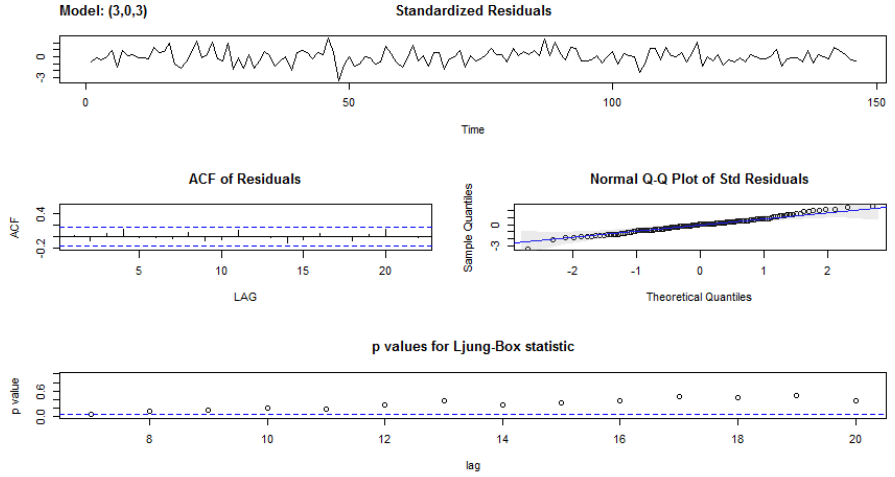

Figure 17. Residuals Analysis of ARIMA(3,0,3) of $\hat{x}_t$

Based on the statistical information of several models with its Residuals Analysis implying the fair whiteness, we choose $ARIMA(3,0,3)$ as the most optimal model in this case.

Then, we iteratively run the weight least squares on the regression model, denoted as (22c), as following procedures.

Recall, for a causal ARIMA(3,0,3) model $\phi(B)x_t = \theta(B)w_t$, that we may write

$$x_t = \pi(B)w_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} \tag{24}$$

To solve for the ψ-weights in general, we must match the coefficients in $\phi(z)\psi(z) = \theta(z)$,

$$(1-\phi_1 z - \phi_2 z^2 - \cdots)(\psi_0 + \psi_1 z + \psi_2 z^2 + \cdots) = (1 + \theta_1 z + \theta_2 z^2 + \cdots) \tag{25}$$

2.3 (c)

We could see the ψ-weights satisfy the homogeneous difference equation given by

$$\psi_j - \sum_{k=1}^{P} \phi_k \psi_{j-k} = 0, j \geq 4 \qquad \psi_j - \sum_{k=1}^{j} \phi_k \psi_{j-k} = \theta_j, 0 \geq j < 4 \tag{26}$$

The general solution depends on the roots of the AR polynomial $\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$

After that, we begin our first iterative estimation by taking the partial derivative of (22c) on β_0,β_1 and set it equal to zero

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2 \sum_{t=1}^{n} [\pi(B)\nabla S_t - (\beta_0 + \pi(B)\beta_1 \nabla L_{t-3})] = 0 \\ \frac{\partial S}{\partial \beta_1} &= -2 \sum_{t=1}^{n} \pi(B)\nabla L_{t-3}[\pi(B)\nabla S_t - (\beta_0 + \pi(B)\beta_1 \nabla L_{t-3})] = 0 \end{aligned} \tag{27}$$

We begin our first WLS estimator in (27) as $\hat{\beta}_0^{(0)}, \hat{\beta}_1^{(0)}$. Then we update the residual series x_t as $x_t^{(1)} = \nabla S_t - \hat{\beta}_0^{(0)} - \hat{\beta}_1^{(0)} \nabla L_{t-3}$ and repeat the estimate procedure in (27) until convergence appears(M times), denoted the last estimators as $\hat{\beta}_0^{(M)}, \hat{\beta}_1^{(M)}$. Finally, our model for the regression model is

$$\begin{aligned} \nabla S_t = & \underbrace{\beta_0 + \beta_1 \nabla L_{t-3}}_{Main} + \\ & \underbrace{\sum_{i=1}^{3} \phi_i x_{t-i} + \sum_{j=1}^{3} \theta_j w_{t-j} + w_t}_{Residual\ x_t} \end{aligned} \tag{28}$$

We run the estimate in R and it converged after 30 iteration, we display the estimated results below

	Estimate	SE	t.value	p.value
ϕ_1	-0.191	0.116	-1.653	0.101
ϕ_2	-0.036	0.098	-0.366	0.715
ϕ_3	0.733	0.087	8.393	0.000
θ_1	0.437	0.137	3.191	0.002
θ_2	0.384	0.135	2.853	0.005
θ_3	-0.596	0.131	-4.555	0.000
β_0	0.407	0.257	1.586	0.115
β_1	0.422	0.349	1.210	0.228

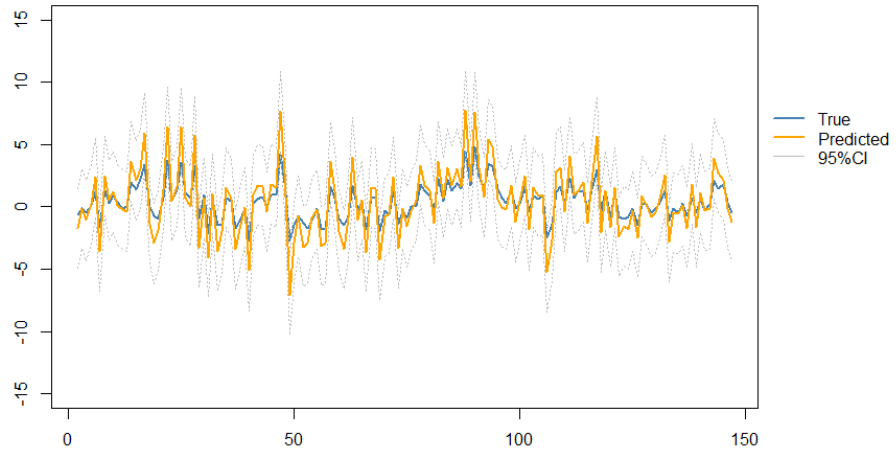

Figure 18. Real S_t series and Predicted $\hat{S}_t$ for model ARIMA(3,0,3) series with autocorrelated errors

In Figure 18, it appears that the trend of the forecast match up with the real ∇S_t series and all real series lie within the 95% confidence interval assuming the normality.

3 3.36

3.1 (a)

The data set cpg, denoted as c_t, taken from a sample of manufactures from 1980 to 2008. Plot the data using R and the below Figure 19 is the graphic figure.

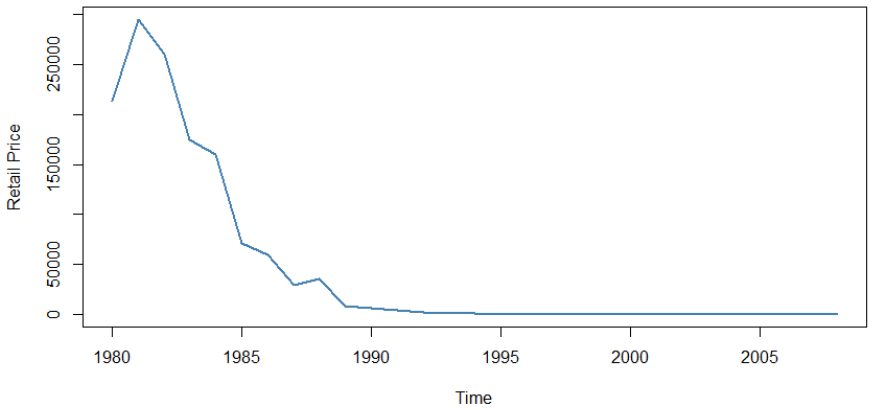

Figure 19. Plot of meadian annual retail price c_t

We note from Figure 19 the price reach its maximum at 1981 and then drastically decrease

from 1981 to 1996. After that, the price remain the low level.

3.2 (b)

First, we fit a linear regression model of $\log c_t$ on t,

$$\begin{aligned} c_t &\approx \alpha e^{\beta t} \\ \log c_t &= \log \hat{\alpha} + \hat{\beta} t = \hat{\beta}_0 + \hat{\beta}_1 t \end{aligned} \tag{29}$$

and then plot the fitted values of (29) compared with the real logged data.

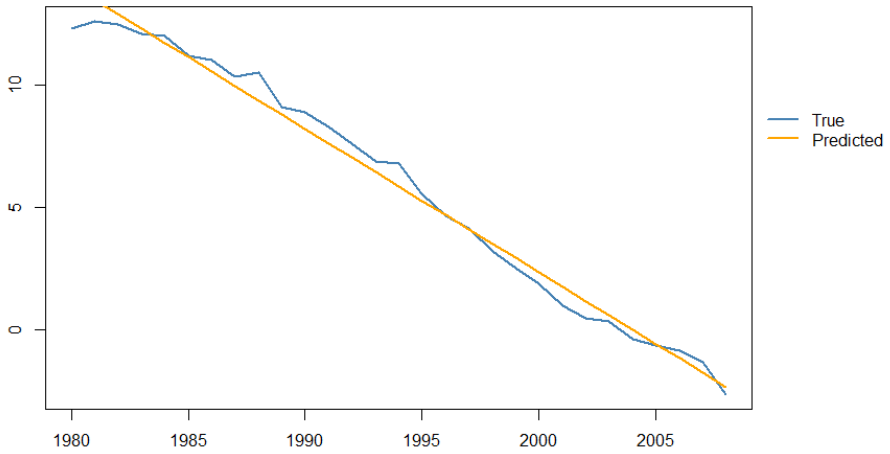

Figure 20. The fitted line and the logged data, the blue solid line represent the logged data, the orange solid line represent the linear fitted values

After the log transformation of the c_t, the transformed series appear fair linearity and the sharp decrease trend is eliminated by log transformation. So, our linear regression model (29) obtains quite optimal fitness.

3.3 (c)

In this section, we inspect the residuals of our model (29)

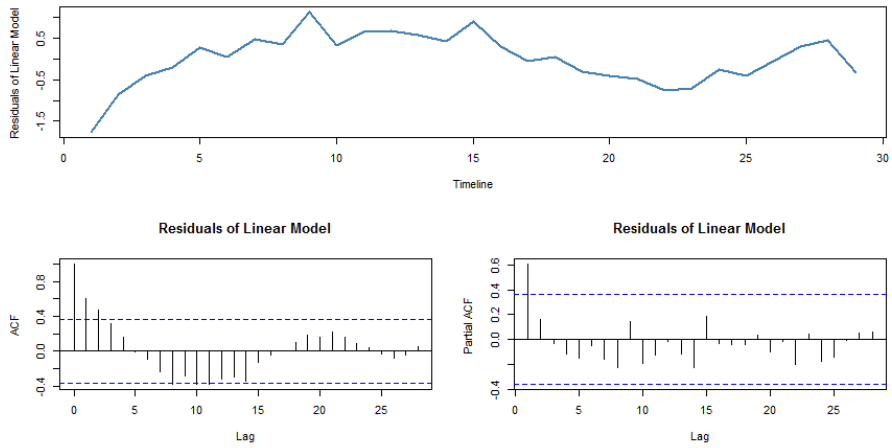

Figure 21. Residuals of the linear regression

In Figure 21, the sample ACF tails off and PACF cuts off after lag 1. So we assume the AR(1) model for the residuals of the linear regression.

3.4 (d)

In this section, we iteratively fit the regression model with autocorrelated error x_t, say AR(1).

$$\log c_t = \underbrace{\beta_0 + \beta_1 t +}_{Main} \underbrace{\phi_1 x_{t-1} + w_t}_{Residuals\ x_t} \tag{30}$$

Multiplying (30) by $(1-\phi_1 B)$ and obtain

$$(1-\phi_1 B)\log c_t = (1-\phi_1 B)(\beta_0+\beta_1 t) + (1-\phi_1 B)x_t = (1-\phi_1 B)(\beta_0+\beta_1 t) + w_t \tag{31}$$

Then we minimized the weighted least square mentioned in (32)

$$S(\phi,\beta) = [(1-\phi_1 B)\log c_t - (\beta_0+\beta_1(t-\phi_1(t-1)))]^2 \tag{32}$$

By take the partial derivative of (32) equal to zero and iteratively M times to estimate the β_0, β_1 for convergence, we obtain $\hat{\phi}^{(M)}, \hat{\beta}_0^{(M)}, \hat{\beta}_1^{(M)}$. More details are displayed below

	Estimate	SE	t.value	p.value
$\hat{\phi}^{(M)}$	0.830	0.119	6.974	0.000
$\hat{\beta}_0^{(M)}$	1113.011	73.567	15.129	0.000
$\hat{\beta}_1^{(M)}$	-0.555	0.037	-15.072	0.000

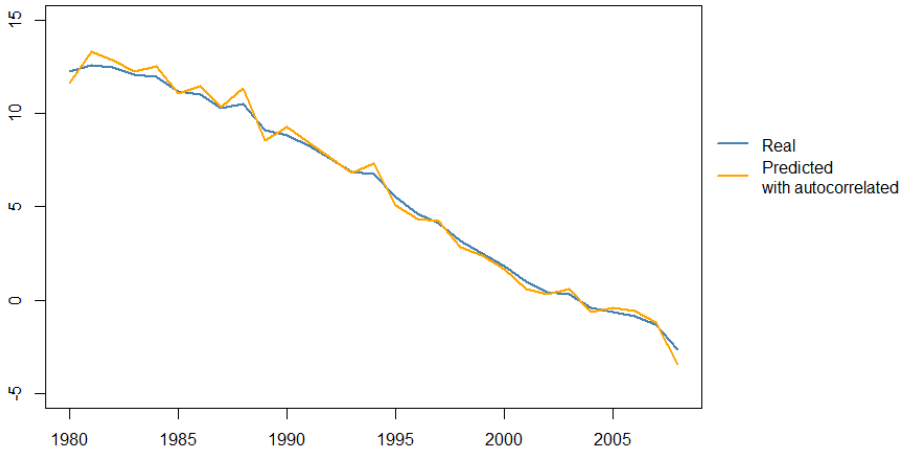

Figure 22. Real data series and Prediction with autocorrelated regression

References

1. Donald Cochrane and Guy H Orcutt. Application of least squares regression to relationships containing auto-correlated error terms. *Journal of the American statistical association*, 44(245):32–61, 1949.

2. Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

3. Nariaki Sugiura. Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. *Communications in Statistics-Theory and Methods*, 7(1):13–26, 1978.

Appendix

Set up

```
library(astsa)
library(tseries)
library(Rmisc)
library(ggplot2)
library(reshape2)
library(xtable)
```

```
my_theme <- theme_bw()+
theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.border = element_blank())
par(mai=c(0.5,0.5,0.5,1.5))
xy <- par('usr')
```

Sunspot Rcode

```
# Loading Original Data
sunspot <- read.table('sunspot.dat')
sun_ts <- ts(sunspot,
             start = 1700,end=1984,
             frequency =1)
len <- length(sun_ts)
plot.ts(sun_ts,
     col='steelblue',
     ylab='Sunspot',xlab='Date',
     lwd=2)
# compute ACF and PACF
acf(sun_ts,20,main='',lwd=2)
pacf(sun_ts,20,main='',lwd=2)
adf.test(sun_ts) # ADF test of original series

# Select p and q for seasonal ARMA(p,q)*(1,0) model
ps <- seq(4)
qs <- seq(4)
BIC <- matrix(0,nrow=4,ncol=4)
AIC <- matrix(0,nrow=4,ncol=4)
ESS <- matrix(0,nrow=4,ncol=4)
R2 <- matrix(0,nrow=4,ncol=4)
SS <- mean(var(sun_ts))
for(p in ps)
  for(q in qs)
  {
    mod <- sarima(sun_ts,p,0,q,1,0,0,11,details=FALSE)
    error <- sum(mod$fit$residuals^2,na.rm=TRUE)/(len-cad-1)
```

```
    BIC[p,q] <- mod$BIC
    AIC[p,q] <- mod$AIC
    ESS[p,q] <- error
    R2[p,q] <- 1-ESS[p,q]/SS
  }

# Define the plot_matrix function for selection
plot_matrix <- function(mat, title)
{
  frame <- melt(mat)
  p <- ggplot(frame, aes(x=Var1, y=Var2))+
    geom_tile(aes(fill=value), colour='white')+
    scale_fill_gradient(low = "white", high = "#1a476f")+
    ylab('p')+xlab('q')+ggtitle(title)+my_theme
  return(p)
}
p1 <- plot_matrix(BIC, 'BIC(p,q)')
p2 <- plot_matrix(AIC, 'AIC(p,q)')
p3 <- plot_matrix(ESS, 'ESS(p,q)')
p4 <- plot_matrix(R2, 'R2(p,q)')
multiplot(p1, p2, p3, p4, cols=2)
# Select the optimal lag p, q and Output LaTex code
p_best <- 3
q_best <- 4
mod_best_sun <- sarima(sun_ts, p_best, 0, q_best, 1, 0, 0, 11)
xtable(mod_best_sun$ttable, digits = 4)
# Obtain the Predict results
pre_ts_all <- sun_ts+mod_best$fit$residuals
# Compare the Prediction with the Real series
plot.ts(sun_ts,
        col='steelblue',
        ylab='Sunspot',
        xlab='Date',
        ylim=c(-20,250),
        lwd=2)
lines(pre_ts_all,
```

```
      col='orange')
legend(x=xy[2],
        y=xy[4]-yinch(0.8),
        legend=c('True',
        'Predicted'),
        xpd=TRUE,
        bty='n',
        col=c('steelblue','orange'),
        lty=1,
        lwd=2,
        title = '')
# Forecast the sunspots out for 4 years from 1985 to 1988
Sunspot <- sun_ts
pre_ts <- sarima.for(Sunspot ,4,p_best,0,q_best,1,0,0,11,
                     plot.all = FALSE)
# Loading real series from 1985 to 1988
sun_ts_all <- read.table('sunspot2.dat')
sun_ts_all <- ts(sun_ts_all,frequency=1,
                 start = 1700,end=1988)
real_four <- sun_ts_all[286:289]

# Compare the Forecast with Real series in plot with 95% CI
predict_four <- pre_ts$pred
predict_four_up <- pre_ts$pred+1.96*pre_ts$se
predict_four_low <- pre_ts$pred-1.96*pre_ts$se
real_predict_four <- cbind(real_four,predict_four)
plot.ts(real_predict_four[,1],
        type='l',
        col='steelblue',
        ylab='Sunspot',
        xlab='Date',
        ylim=c(-20,200),
        lwd=2,xaxt='n')
axis(side=1,at=1985:1988,labels=1985:1988)
lines(real_predict_four[,2],
      col='orange',
```

```
       lwd=2)
lines(predict_four_up,lty=2,
       col='grey',lwd=2)
lines(predict_four_low,lty=2,
       col='grey',lwd=2)
legend(x=xy[2],
         y=xy[4]-yinch(0.8),
         legend=c('True',
                  'Predicted',
                  'CI'),
         xpd=TRUE,
         bty='n',
     col=c('steelblue','orange','grey'),
     lty=c(1,1,3),
     lwd=c(1,1,2),
     title = '')

# Combine the forecast Details and Output LaTex code
pre_info_sun <- data.frame(cbind(real_four,
                                 pre_ts$pred,
                                 pre_ts$se,
                                 predict_four_up,
                                 predict_four_low))
colnames(pre_info_sun) <- c('True','Predicted',
                            'Standard Deviation','Up Bounds',
                            'Low Bounds')
xtable(pre_info_sun,digits = 3)
```

3.35 Rcode

```
lag_ADF = (length(sales)-1)^(1/3)
adf.test(sales) # ADF test of Original series
# Overall Plot
plot.ts(sales,
         col='steelblue',
         ylab='Sales',xlab='Timeline',
         lwd=2)
```

```
# Fit linear regression model
fit <- lm(sales~time(sales),na.action = NULL)
# Plot residuals and its ACF,PACF at preset layout
plot_fit <- function(fit,ylab)
{
  layout(matrix(c(1,1,1,1,1,1,1,1,2,2,3,3,2,2,3,3),ncol=4,byrow=
   TRUE))
  plot.ts(fit$residuals,
          col='steelblue',
          ylab=ylab,xlab='Timeline',
          lwd=2)
  acf(fit$residuals,150,main=ylab)
  pacf(fit$residuals,150,main=ylab)

}
# ADF of model residuals
adf.test(fit$residuals)
# Plot differenced series with its ACF and PACF
plot.ts(diff(sales),
        col='steelblue',
        ylab='First Difference',xlab='Timeline',
        lwd=2)
acf(diff(sales),150,main='First Difference Time Series')
pacf(diff(sales),150,main='First Difference Time Series')
adf.test(diff(sales)) # ADF of the differenced series
sales_diff <- diff(sales)
# Define model selection function and ggplot
statistic_test <- function(dat)
{
  ps <- seq(4)
  qs <- seq(4)
  BIC <- matrix(0,nrow=4,ncol=4)
  AIC <- matrix(0,nrow=4,ncol=4)
  ESS <- matrix(0,nrow=4,ncol=4)
  R2 <- matrix(0,nrow=4,ncol=4)
```

```
  SS <- mean(var(sales_diff))
  for(p in ps)
    for(q in qs)
    {
      mod <- sarima(dat,p,0,q,details=FALSE)
      error <- sum(mod$fit$residuals^2,na.rm=TRUE)/(len-cad-1)
      BIC[p,q] <- mod$BIC
      AIC[p,q] <- mod$AIC
      ESS[p,q] <- error
      R2[p,q] <- 1-ESS[p,q]/SS
    }
  p1 <- plot_matrix(BIC,'BIC(p,q)')
  p2 <- plot_matrix(AIC,'AIC(p,q)')
  p3 <- plot_matrix(ESS,'ESS(p,q)')
  p4 <- plot_matrix(R2,'R2(p,q)')
  multiplot(p1,p2,p3,p4,cols=2)
}

# choose the best model and Output the LaTex code
mod_best <- sarima(sales_diff,4,0,3,
                   no.constant = TRUE)
xtable(mod_best$ttable,digits=4)
lead_diff <- diff(lead)
# CCF of differenced sales and lead
ccf_value <- ccf(sales_diff,lead_diff,main='Sales v.s. Lead',
    ylab='CCF')
# the maximum value of CCF
ccf_value$lag[which.max(ccf_value$acf)]
# lag plot of differenced lead and sales
lag2.plot(lead_diff,sales_diff,8)
# Fit differenced sales of differenced lead at lag 3
lag_reg <- ts.intersect(y=sales_diff,x=lag(lead_diff,3),dframe =
   T)
fit_lag <- lm(y~x,lag_reg)
# Residual Analysis of fitted model
plot_fit(fit_lag,'Residuals Analysis')
```

```
statistic_test(fit_lag$residuals)
# Select best lag of the residuals and Output LaTex code
mod_best <- sarima(fit_lag$residuals,
                   3,0,3)
mod_all_best <- sarima(lag_reg$y,3,0,3,xreg = lag_reg$x)
xtable(mod_all_best$ttable,digits = 3)
mod_pre <- mod_all_best$fit$residuals+lag_reg$y
mod_pre_limit <- mod_pre+1.96*mod_all_best$fit$sigma2
mod_pre_low <- mod_pre-1.96*mod_all_best$fit$sigma2
# plot predict with real data and 95% CI
plot.ts(lag_reg$y,
        col='steelblue',
        lwd=2,
        xlab='Time',ylab='Values',
        ylim=c(-15,15))
lines(mod_pre,
      col='orange',
      lwd=2)
lines(mod_pre_limit,
      col='grey',
      lwd=1,lty=3)
lines(mod_pre_low,
      col='grey',
      lwd=1,lty=3)
legend(x=xy[2],y=xy[4]-yinch(0.8),
       legend=c('True',
                'Predicted',
                '95%CI'),
       xpd=TRUE,
       bty='n',
       col=c('steelblue',
         'orange',
         'grey'),
       lty=c(1,1,1),
       lwd=c(2,2,1),
       title = '')
```

3.36

```
# Plot overall Cpg
plot.ts(cpg,
        col='steelblue',
        lwd=2,
        ylab='Retail Price',
        xlab='Time')
# logged Transformation
logg_cpg <- log(cpg)
# Fit the logged series with time t
fit_logg <- lm(logg_cpg~time(logg_cpg))
fitted_cpg <- fit_logg$fitted.values
# Plot fitted values with real logged series
plot(logg_cpg,
        col='steelblue',
        lwd=2,
        ylab='Logged Retail Price',
        xlab='Time')
fitted_cpg <- ts(array(fitted_cpg),start = 1980,end=2008,
   frequency = 1)
lines(fitted_cpg,
      col='orange',
      lwd=2)
legend(x=xy[2],y=xy[4]-yinch(0.8),
       legend=c('True',
                'Predicted'),
       xpd=TRUE,
       bty='n',
       col=c('steelblue',
        'orange'),
       lty=c(1,1),
       lwd=c(2,2),
       title = '')
# Residual Analysis of fit model
plot_fit(fit_logg,'Residuals of Linear Model')
```

```
fit_residuals <- fit_logg$residuals
# Refit the model with autocorrelated error and Output the LaTex
   code
fit_all_arima <- sarima(logg_cpg,1,0,0,xreg = time(logg_cpg))
xtable(fit_all_arima$ttable,digits = 3)
# Plot the model fit with real series
plot.ts(logg_cpg,
        col='steelblue',
        lwd=2,
        ylab='Logged Retail Price',
        xlab='Time',
        ylim=c(-5,15))
lines(fit_all_arima$fit$residuals+logg_cpg,
      col='orange',
      lwd=2))
legend(x=xy[2],y=xy[4]-yinch(0.8),
       legend=c('Real',
                'Predicted \nwith autocorrelated'),
       xpd=TRUE,
       bty='n',
       col=c('steelblue',
        'orange'),
       lty=c(1,1),
       lwd=c(2,2),
       title = '')
```