Time Series Analysis

Chenyin Gao, Yuying Huang

gaochy5@mail2.sysu.edu
1104434805@qq.com

Sunspot

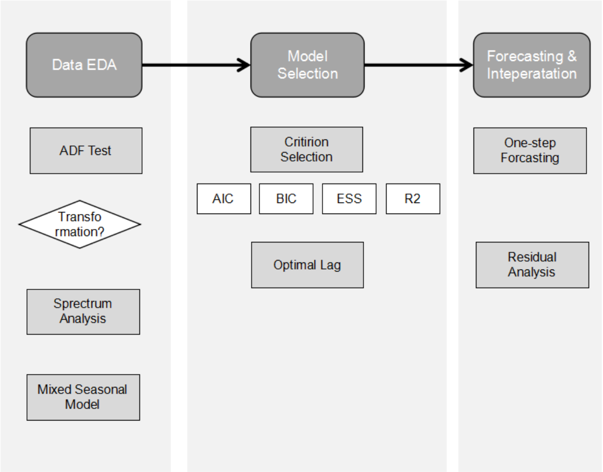


Figure 1: Flowchart Our work Organization

EDA

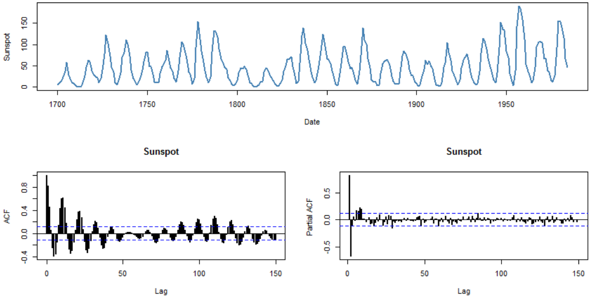


Figure 2: Yealy Sunspot from 1700 to 1984(top) with ACF and PACF(bottom)

1. Since the original series appears to contain a sequence of repeating short signals, the ACF confirms this behavior showing the repeating peaks.
2. 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.

ADF test

Augmented Dickey-Fuller Test

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

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

$$\Phi_P(B^s)x_t = \Theta_Q(B_s)w_t \tag{1}$$

where the operators are

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

and

$$\Theta_Q(B^s) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Ps} \tag{3}$$

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

$$\Phi_p(B^s)\phi(B)x_t = \Theta_Q(B^s)\theta(B)w_t \tag{4}$$

In Figure 3, we select the most significant value at frequency $\omega = 0.09028$. So we assume the cycle of $s = 11$ years in this case.

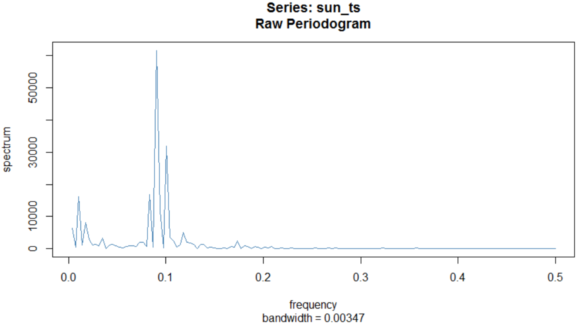


Figure 3: Periodogram value of Sunspot

In Figure 4,

1. the ACF tails off at points $= 1s, 2s, \cdots$, with repeated peaks and PACF cuts off after lag s implying the model SAR(1),$P = 1, Q = 0$, in the seasonal model.
2. From ACF and PACF, it suggests an ARMA(p, q) $(0 < p, q < s)$ model within the seasonal model in a lower lag.

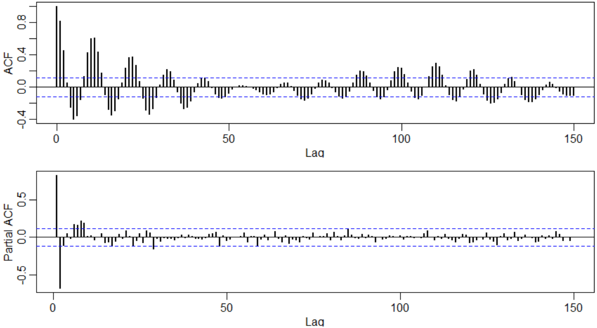


Figure 4: ACF of Sunspot(top) and PACF(bottom)

Thus, we preform fit ergodic models, $ARMA(p,q) \times (1,0)_{11}$, $0 < p < 5, 0 < q < 5$.

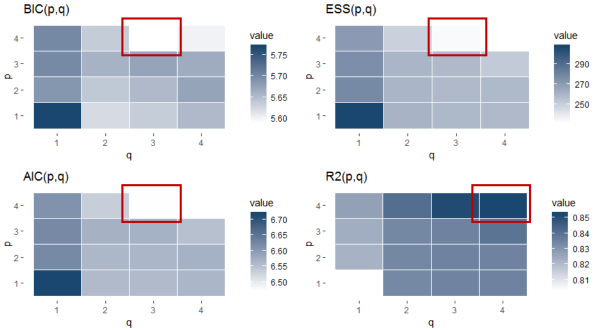


Figure 5: Statistics values for all $1 \leq p, q \leq 4$ with the optimal lag framed by red square

$$(1 - \Phi_1 B^{11})(1 - \sum_{i=1}^{4} \phi_i B^i) x_t = (1 + \sum_{j=1}^{3} \theta_j B^j) w_t \tag{5}$$

In Figure 6, except for one or two outliers, the model fit quite well showing no obvious departure of the residuals from whiteness.

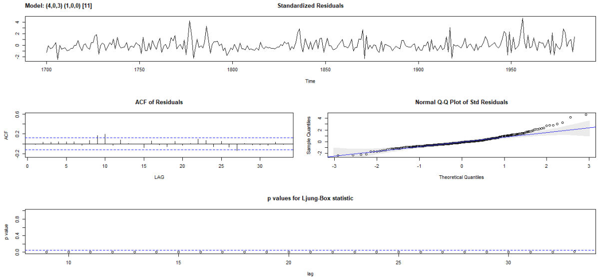
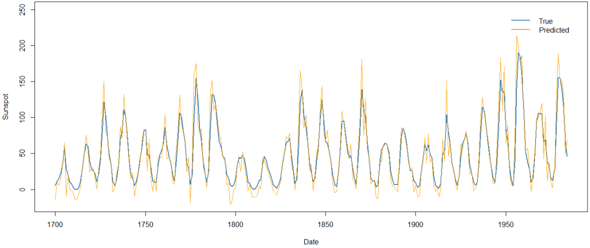


Figure 6: Residual analysis for the $ARMA(4, 0, 3) \times (1, 0, 0)_{11}$ fit to the sunspot data set

	Estimate	SE	t.value	p.value
ϕ_1	0.1073	0.1076	0.9965	0.3199
ϕ_2	0.2645	0.0780	3.3928	0.0008
ϕ_3	0.2929	0.0874	3.3513	0.0009
ϕ_4	-0.5523	0.0761	-7.2568	0.0000
θ_1	1.1421	0.1152	9.9120	0.0000
θ_2	0.6612	0.1462	4.5228	0.0000
θ_3	-0.1943	0.0955	-2.0355	0.0428
Φ_1	0.3122	0.0710	4.3950	0.0000
δ	48.5209	3.8681	12.5438	0.0000



$$
\begin{aligned}
x_t = &\underbrace{48.52}_{\sigma} \\
&+ \underbrace{0.3122x_{t-11}}_{\Phi_1} \\
&+ \underbrace{0.1073x_{t-1} + 0.2645x_{t-2} + 0.2929x_{t-3} - 0.5523x_{t-4}}_{\phi_{1,2,3,4}} \\
&+ \underbrace{1.1421w_{t-1} + 0.6612w_{t-2} - 0.1943w_{t-3}}_{\theta_{1,2,3}}
\end{aligned}
$$

Lastly, we one-step-ahead 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)

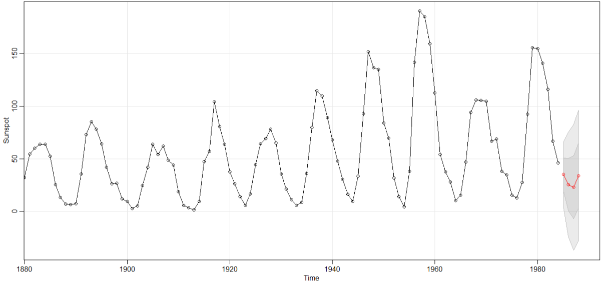


Figure 7: Forecast out 4 years for ARMA$(4,0,3)$ $\times(1,0,0)_{11}$ for the sunspot data set

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

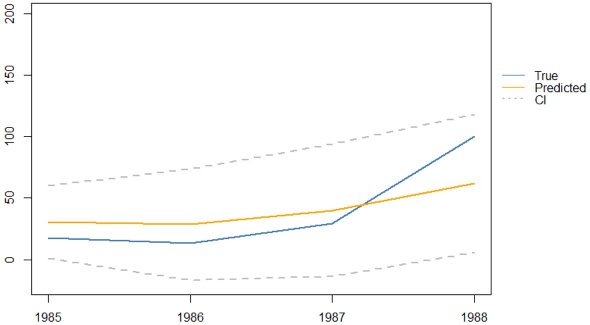


Figure 8: Forecast out 4 years for $\text{ARMA}(4,0,3) \times (1,0,0)_{11}$ compared with Real data

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.

Sales And Leads

1. it is pretty obvious that the original data set is non-stationary.
2. 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.

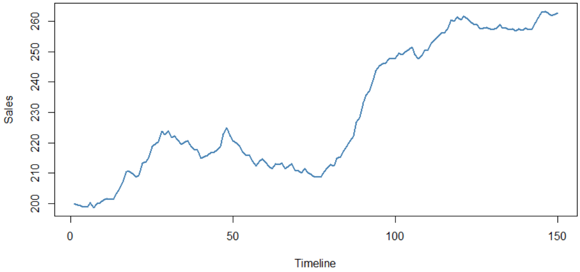


Figure 9: Plot of monthly Sales S_t data set

ADF test

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.

We assume the underlying model of S_t as

$$S_t = \underbrace{\mu_t}_{Trend} + \underbrace{x_t}_{Stationary} + \underbrace{w_t}_{Noise}$$

difference operator

$$\begin{aligned}\nabla S_t = S_t - S_{t-1} &= (\mu_t + x_t + w_t) - (\mu_{t-1} + x_{t-1} + w_{t-1}) \\ &= \delta + w_t - w_{t-1} + x_t - x_{t-1}\end{aligned}$$

where $\mu_t = \delta + \mu_{t-1}$.

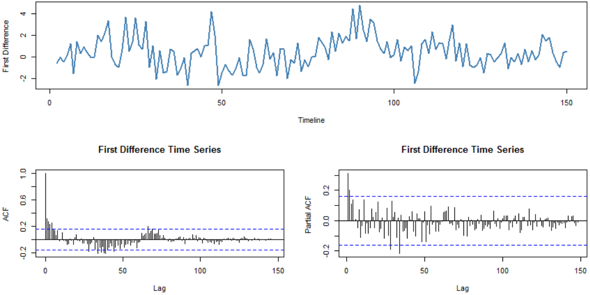


Figure 10: Plot of First Difference Sales with its ACF and PACF

Apply ADF test to see whether it reject the non-stationary null hypothesis.

ADF test

Augmented Dickey-Fuller Test

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

As a result, we perceived the ∇S_t as a stationary process

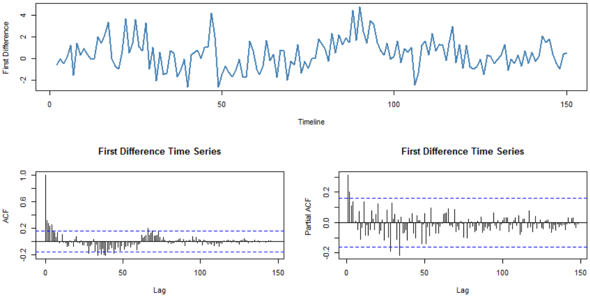


Figure 11: Plot of First Difference Sales with its ACF and PACF

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

$$\nabla S_t = \sum_{i=1}^{p} \phi_i \nabla S_{t-i} + \sum_{j=1}^{q} \theta_j w_{t-j} + w_t \tag{6}$$

Select the optimal p and q based on AIC, BIC, ESS and $R2$ criterion.

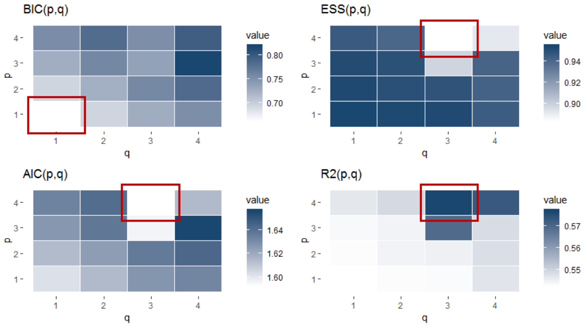


Figure 12: Statistics Values For ARMA(p, q), $1 \leq p, q \leq 4$ with optimal lags framed by red square

So, we choose $ARMA(4, 3)$ as our fit model in this case.

We perform regression of $ARIMA(4,1,3)$ for the non-stationary process S_t and the estimated $ARIMAR(4,1,3)$ model

$$\begin{aligned}\nabla x_t = &\underbrace{0.1416\nabla x_{t-1} + 0.4652\nabla x_{t-2} - 0.0669\nabla x_{t-3} + 0.1409\nabla x_{t-4}}_{\phi_{1,2,3,4}} + \\ &\underbrace{0.0864 w_{t-1} - 0.2912 w_{t-2} + 0.1061 w_{t-3}}_{\theta_{1,2,3}}\end{aligned} \tag{7}$$

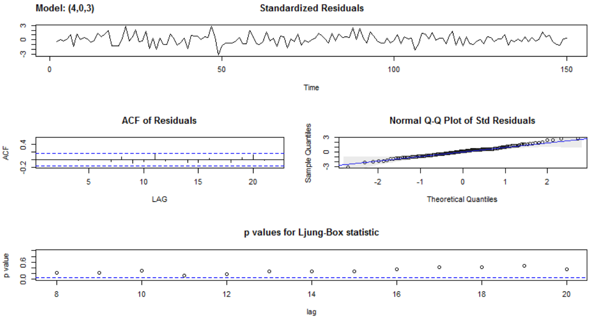


Figure 13: Diagnosis of the residual of ARIMA(4,1,3) fit on the sales data set

1. All of the standardized residuals is within the interval of [-3,3].
2. The ACF of the residuals display no apparent departure from the model assumption
3. the Q-statistic(null hypothesis $\hat{\rho}_e^2(h) = 0$) is never significant at the lags down.
4. the normal Q-Q plot of the residuals shows that the assumption of normality is reasonable, with only exception of one or two.

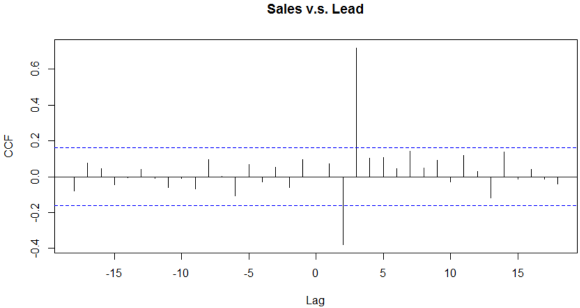


Figure 14: Sample CCF of ∇S_t and ∇L_t ;positive lags indicate sales lags lead

1. Figure 14 shows several departure from the cyclic component of each series and there is an obvious peak at h=3($\hat{\rho_{xy}}(3) = 0.72$).
2. 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.
3. 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.

Lag plots(lowess fits) of ∇S_t on the lag of $\nabla L_{t-i}, i = 1, 2, \cdots, 8$

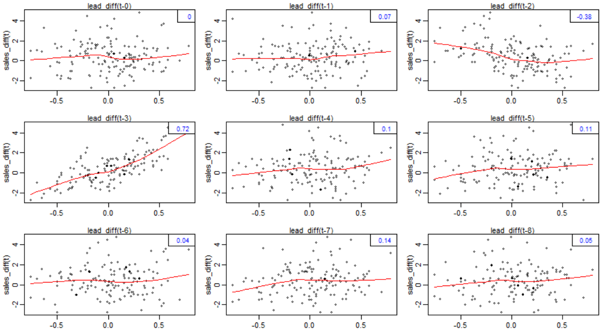


Figure 15: Scatterplot matrix of the ∇S_t and ∇L_t, on the vertical axis plotted against the ∇S_t, and ∇L_t on the horizontal axis at lags h=$0, 1, \cdots, 8$. The valus in the upper right corner are the sample cross-correlations and the lines are a lowess fit

1. In ∇S_t against ∇L_{t-3} plot, we notice that the lowess fits are approximately linear, which suggest that the behavior between ∇S_t and ∇L_t is the same both for positive and negative values of ∇L_t
2. Besides, we see strong positive linear relations at lag=3, that is ∇S_t associated with ∇L_{t-3}.

We begin to fit the regression model with autocorrelated errors as

$$\nabla S_t = \beta_0 + \beta_1 \nabla L_{t-3} + x_t \tag{8a}$$

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 (8a) by $\pi(B)$ and obtain

$$\pi(B)\nabla S_t = \pi(B)(\beta_0 + \beta_1 \nabla L_{t-3}) + \underbrace{\pi(B)x_t}_{w_t} \tag{8b}$$

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

$$S(\phi, \theta, \beta) = \sum_{t=1}^{n} w_t^2 = \sum_{t=1}^{n} [\pi(B)\nabla S_t - \pi(B)(\beta_0 + \beta_1 \nabla L_{t-3})]^2 \tag{8c}$$

An easy way to tackle (8c) 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,

$$\hat{x}_t = \nabla S_t - (\hat{\beta}_0 + \hat{\beta}_1 \nabla L_{t-3}) \tag{9}$$

Then we identify the ARMA models for $\hat{x}_t$ based on AIC,BIC,ESS and R^2 in Figure 16. So, we choose $ARIMA(3,0,3)$ as the most optimal model in this case.

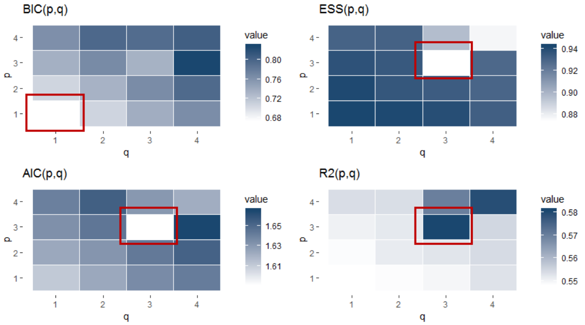


Figure 16: Models Selection of $\hat{x}_t$

Then, we iteratively run the weight least squares on the regression model as following procedures. Recall, for a causal ARIMA$(3,0,3)$ model $\phi(B)x_t = \theta(B)w_t$, that we may write

$$x_t = \pi(B)w_t = \sum_{j=0}^{\infty} \psi_j w_{t-j} \quad (10)$$

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

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

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

$$\psi_j - \sum_{k=1}^{P} \phi_k \psi_{j-k} = 0, j \geq 4$$

$$\psi_j - \sum_{k=1}^{j} \phi_k \psi_{j-k} = \theta_j, 0 \leq j < 4$$

After that, we begin our first iterative estimation by taking the partial derivative of (8c) on β_0,β_1 and set it equal to zero

$$\begin{aligned}
\frac{\partial S}{\partial \beta_0} &= -2\sum_{t=1}^{n}[\pi(B)\nabla S_t - (\beta_0 + \pi(B)\beta_1\nabla L_{t-3})] = 0 \\
\frac{\partial S}{\partial \beta_1} &= -2\sum_{t=1}^{n}\pi(B)\nabla L_{t-3}[\pi(B)\nabla S_t - (\beta_0 + \pi(B)\beta_1\nabla L_{t-3})] = 0
\end{aligned} \tag{11}$$

We begin our first WLS estimator in (11) 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 (11) 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

$$\nabla S_t = \underbrace{\beta_0 + \beta_1 \nabla L_{t-3}}_{Main} + \underbrace{\sum_{i=1}^{3} \phi_i x_{t-i} + \sum_{j=1}^{3} \theta_j w_{t-j} + w_t}_{Residual\ x_t} \tag{12}$$

Table 1: Detail Coefficient

	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

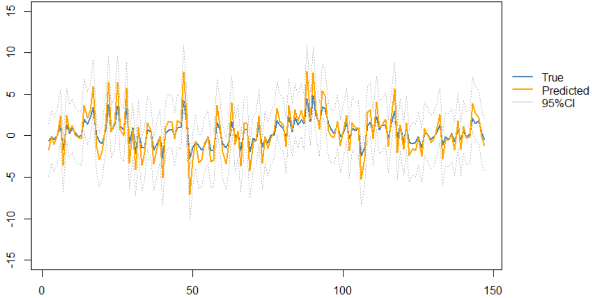


Figure 17: Real S_t series and Predicted $\hat{S}_t$ for model ARIMA(3,1,3) series with autocorrelated errors

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.

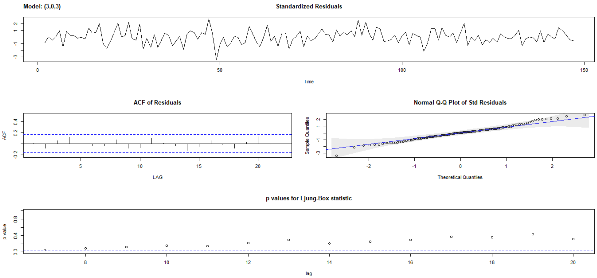


Figure 18: Residual Analysis of ARIMA(3,1,3) model with autocorrelated errors

Justify the whiteness of w_t

Hard Drives

Plot the data c_t below

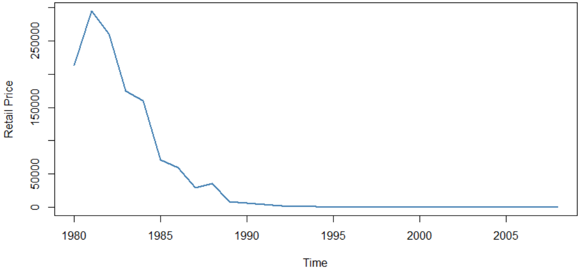


Figure 19: Plot of meadian annual retail price c_t

1. Clearly, the time series is non-stationary.
2. c_t reach its maximum at 1981 and then drastically decrease from 1981 to 1996. After that, the price remain the low level.

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

$$\begin{aligned} c_t &\approx \alpha e^{\beta t} \\ \log c_t &= \log \hat{\alpha} + \hat{\beta} t = \hat{\beta}_0 + \hat{\beta}_1 t \end{aligned} \tag{13}$$

Then plot the fitted values of (13) compared with the real logged data.

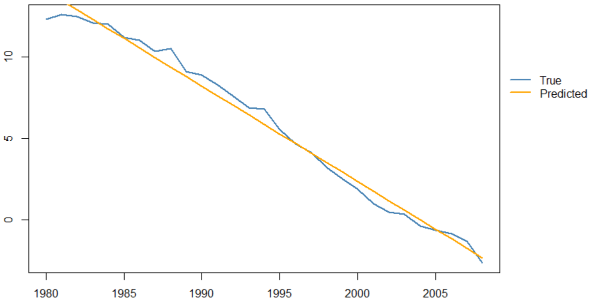


Figure 20: The fitted line and the logged data, the blue solid line represent the logged data, the orange solid line represent the linear fitted values

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 (13) obtains quite optimal fitness.

We then inspect the residuals of our model (13)

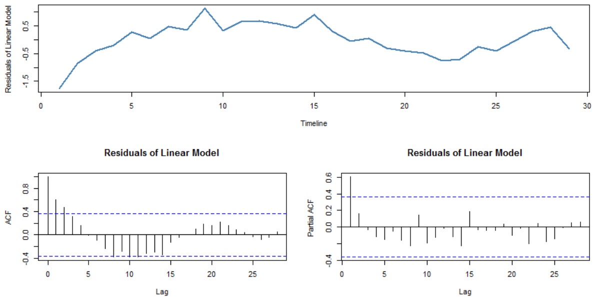


Figure 21: Residuals of the linear regression

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

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

$$\log c_t = \underbrace{\beta_0 + \beta_1 t}_{Main} + \underbrace{\phi_1 x_{t-1} + w_t}_{Residuals\ x_t} \tag{14}$$

Multiplying (14) by $(1-\phi_1 B)$ and obtain

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

Then we minimized the weighted least square

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

By take the partial derivative and set to zero. Iteratively M times to estimate the β_0, β_1 for convergence, we obtain $\hat{\phi}^{(M)}, \hat{\beta}_0^{(M)}, \hat{\beta}_1^{(M)}$.

	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

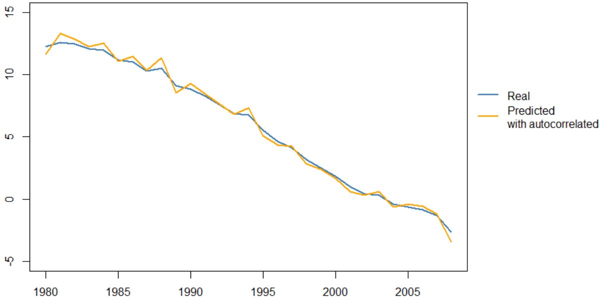


Figure 22: Real data series and Prediction with autocorrelated regression

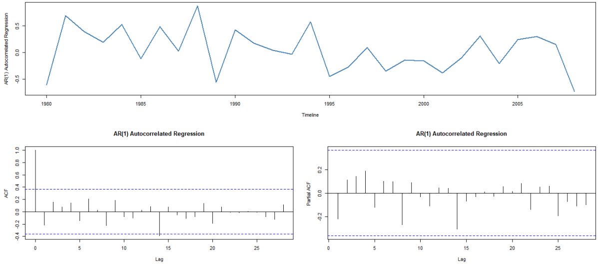


Figure 23: Residual Analysis with autocorrelated regression

From the ACF and PACF of the residual, we could justify the AR(1) residual model.

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.