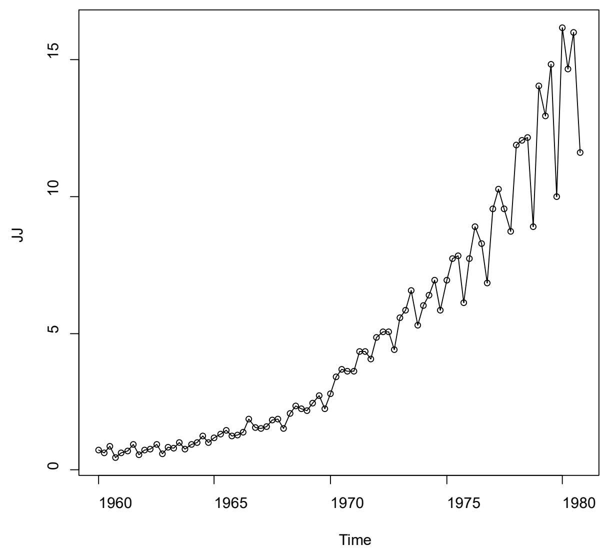
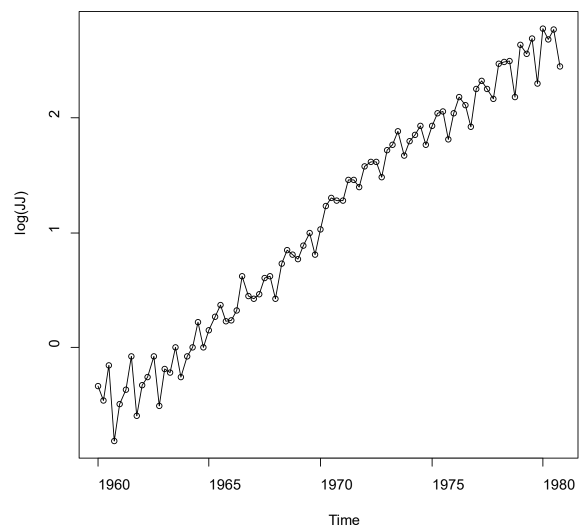
STAT 520 – Homework 6 – Fall 2023

Note: Problem 4 is mandatory for graduate students and extra credit for undergraduates.

1) The quarterly earnings per share for 1960-1980 are in the `JJ` object in the `TSA` package. Type `library(TSA); data(JJ); print(JJ)` in R to see the data set.

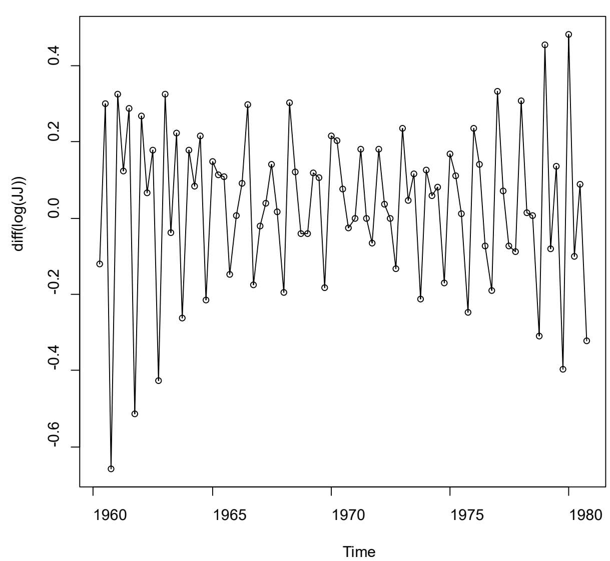
(a) Plot the original time series and the (natural) logarithm of the series. Explain why the log transformation is useful when modeling this series.





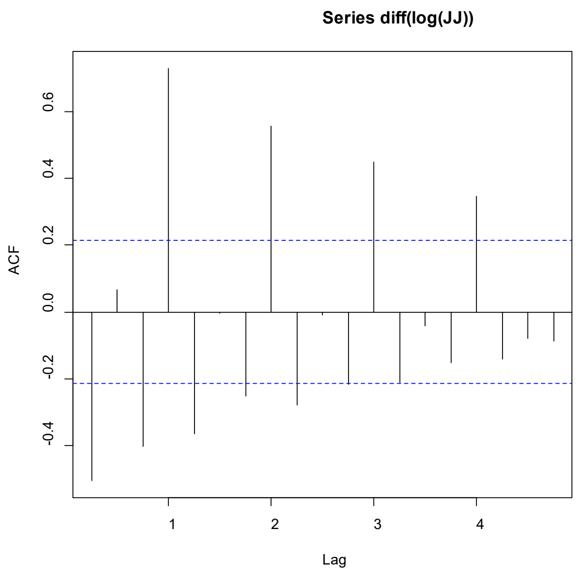
The log transformation produces a time series whose variance can be treated as constant over time.

(b) Explain why the log-transformed series is clearly not stationary. Plot the first differences of the log-transformed series. Could the differenced logged series be viewed as stationary?



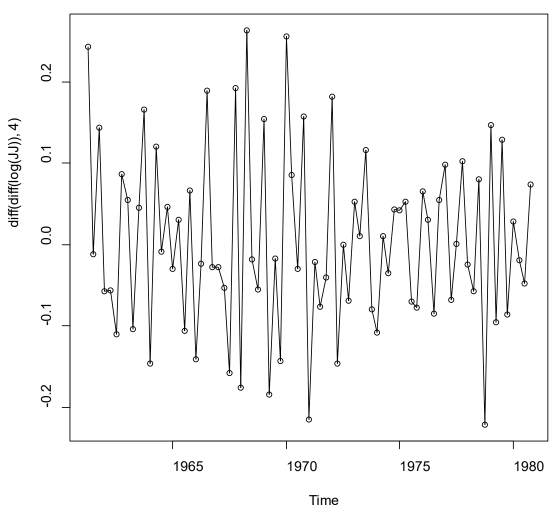
The log-transformed series is clearly not stationary since its mean function clearly increases over time. The differenced logged series is closer to stationary (its mean function appears constant), but it appears somewhat doubtful that the variance of the differenced logged series is constant over time.

(c) Graph the sample ACF of the differenced logged data. What does this plot suggest?



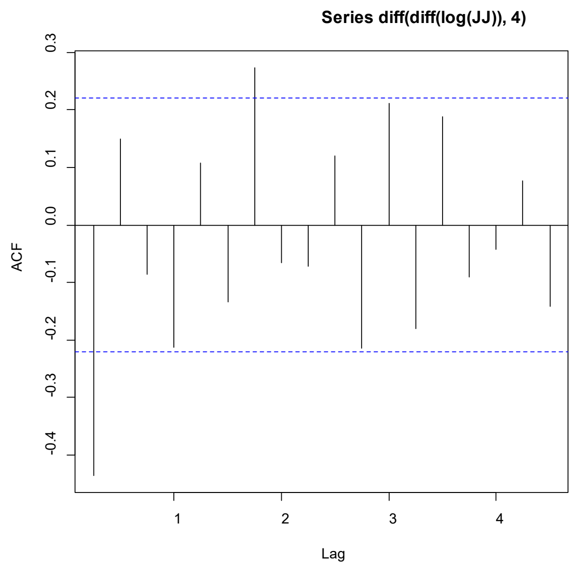
This plot suggests there is clear seasonality in the data, since the autocorrelations are strongly positive at lags 4, 8, 12, 16, …

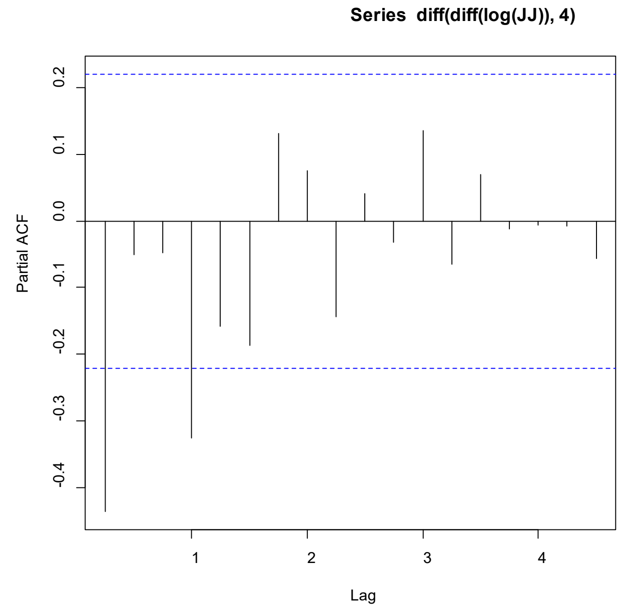
(d) Take both the first differences and the (lag-4) seasonal differences of the logged data. Plot this differenced and seasonally differenced series. Interpret what the plot indicates.



This plot resembles a stationary series a bit more than the other plots have looked at previously.

(e) Plot the sample ACF of the differenced and seasonally differenced logged series. Interpret what the plot tells you.





There is a significant ACF value at lag 1 and a nearly significant ACF value at lag 4 (the 1 year lag). The PACF values seem to die off as time goes on, but one could also view the PACFs as having significant values at lag 1 and lag 4. All in all, this would indicate possibly an ARIMA(0,1,1) × $(0,1,1)_4$ model or an ARIMA(1,1,0) × $(1,1,0)_4$ model, since we have taken first differences and seasonal differences.

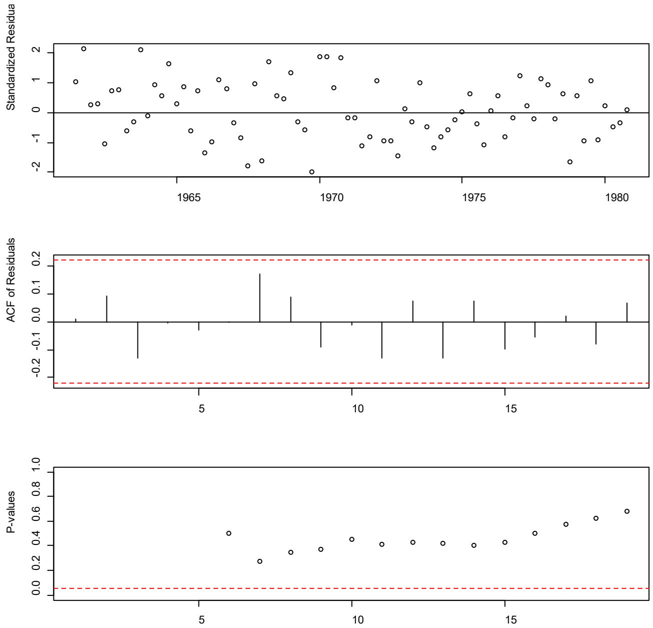
(f) Fit an ARIMA(0,1,1) × $(0,1,1)_4$ model, and assess the significance of the estimated coefficients.

```
Coefficients:
          ma1     sma1
      -0.6809  -0.3146
s.e.   0.0982   0.1070

sigma^2 estimated as 0.007931:  log likelihood = 78.38,  aic = -152.75
```

We see that both the nonseasonal MA coefficient and the seasonal MA coefficient are significant, based on their estimates and standard errors.

(g) Perform model diagnostics, based on the residuals.



There are no excessive standardized residuals; the residuals do not have significant autocorrelations, so they behave like white noise, and the Ljung-Box p-values are nonsignificant at all lags shown, indicating the residual autocorrelation are not too large as a set. The model seems to fit well.

(h) Calculate the forecasts for the next 8 quarters of the series. Plot the forecasts along with 95% prediction intervals.

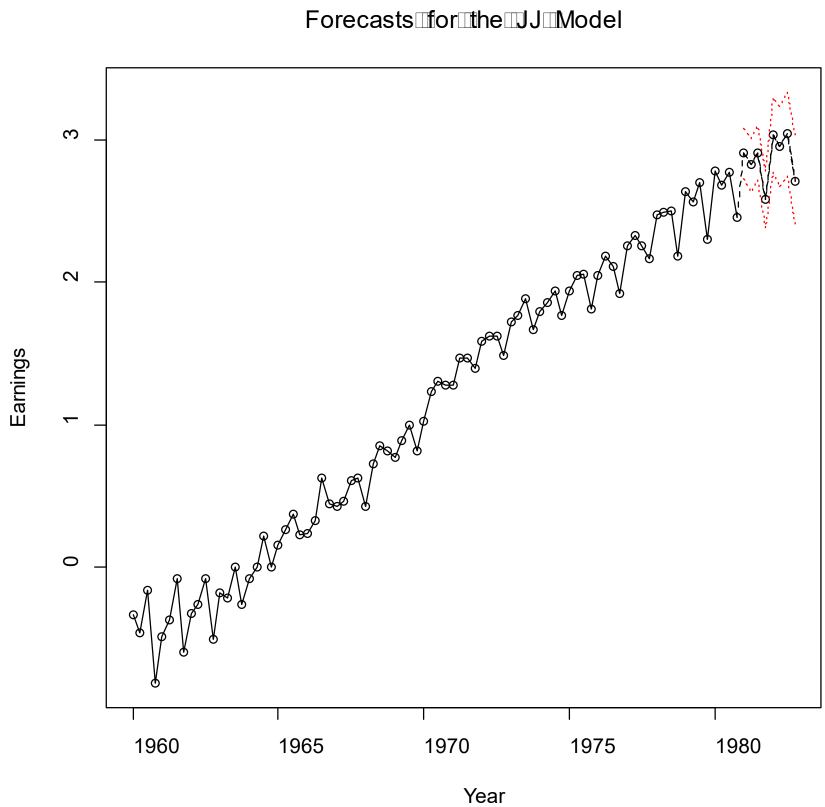
In terms of the log-transformed data:

```
$pred
         Qtr1     Qtr2     Qtr3     Qtr4
1981 2.905343 2.823891 2.912148 2.581085
1982 3.036450 2.954999 3.043255 2.712193

$se
           Qtr1       Qtr2       Qtr3       Qtr4
1981 0.08905414 0.09347899 0.09770366 0.10175307
1982 0.13548771 0.14370561 0.15147833 0.15887123
```

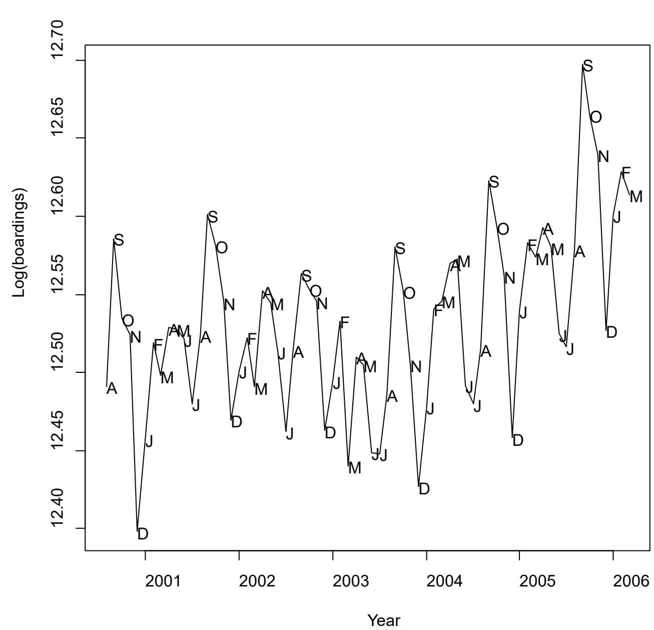
In terms of the original variable:

```
         Qtr1     Qtr2     Qtr3     Qtr4
1981 18.27151 16.84226 18.39626 13.21147
1982 20.83116 19.20170 20.97340 15.06227
```



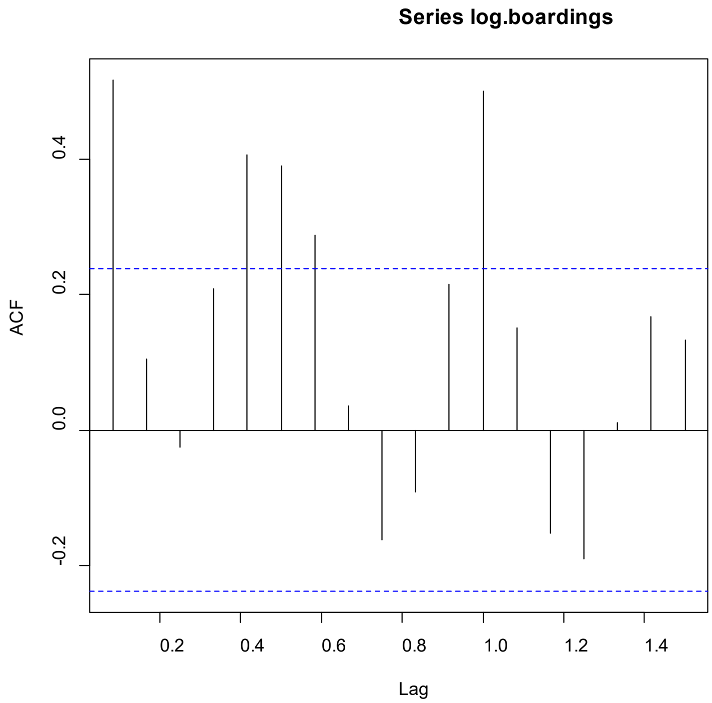
2) A data set of public transportation boardings in Denver from August 2000 through December 2005 are in the `boardings` object in the `TSA` package. These data are *already* logged: You can type `library(TSA); data(boardings); log.boardings = boardings[,1]; print(log.boardings)` in R to see the data set.

(a) Give a time series plot of these data. Include plotting symbols for the months that help you assess seasonality. Comment on the plot and any seasonality. Is it reasonable to use a stationary model for this time series?



There appears to be a seasonal pattern, with the series consistently reaching peaks in early fall (September) and mid-spring (April, May) and low valleys in mid-summer (June, July) and December … possibly when people are off work or out of town traveling. It could be reasonable to view this series as stationary, although the mean function does appear to increase slightly over the last year or so of the series.

(b) Plot the sample ACF of the series. Interpret what the plot tells you.



There appears to be major autocorrelation around 6 months, 12 months, 18 months, etc., as if there is a repeating pattern every half-year or so.

(c) Fit an ARMA(0,3) × $(1,0)_{12}$ model to the data, and assess the significance of the estimated coefficients.

```
Coefficients:
         ma1     ma2     ma3    sar1  intercept
      0.7290  0.6116  0.2950  0.8776    12.5455
s.e.  0.1186  0.1172  0.1118  0.0507     0.0354

sigma^2 estimated as 0.0006542:  log likelihood = 143.54,  aic = -277.09
```

We see that all three nonseasonal MA coefficients, and the seasonal AR coefficient, are significant.

(d) Overfit with an ARMA(0,4) × $(1,0)_{12}$ model to the data, and interpret the results.

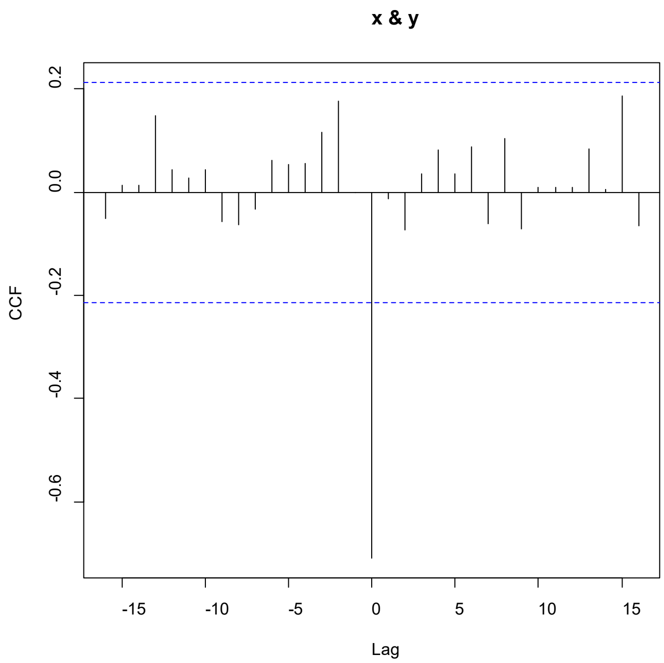
```
Coefficients:
         ma1     ma2     ma3     ma4    sar1  intercept
      0.7277  0.6686  0.4244  0.1414  0.8918    12.5459
s.e.  0.1212  0.1327  0.1681  0.1228  0.0445     0.0419

sigma^2 estimated as 0.0006279:  log likelihood = 144.22,  aic = -276.45
```

We see that the MA-4 coefficient is not significant, and the AIC is worse for this model, which is evidence that the ARMA(0,3) × $(1,0)_{12}$ model is better and may be sufficient.

3) A data set of weekly sales and prices of Bluebird Lite potato chips are in the bluebidlite object in the TSA package. The sales are *already* logged. Type library(TSA); data(bluebirdlite); print(bluebirdlite) in R to see the data set. And type data(bluebirdlite); log.sales=bluebirdlite[,1]; price=bluebirdlite[,2] to access the individual time series.

(a) Use the prewhiten function to obtain the cross-correlation function of the prewhitened version of logged sales and price. What can be discerned about the cross-correlations at the various lags?



We see strong negative contemporaneous (lag-0) cross correlation between log sales and price: As the price goes down, the (log) sales immediately tends to go up.

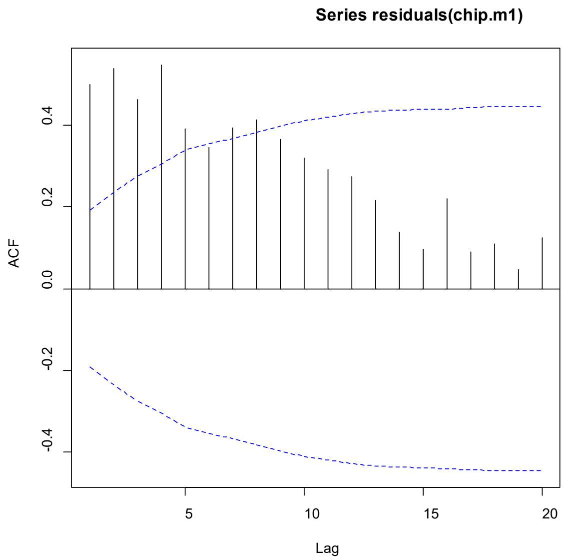
(b) Fit an ordinary least squares (OLS) regression of log sales against price.

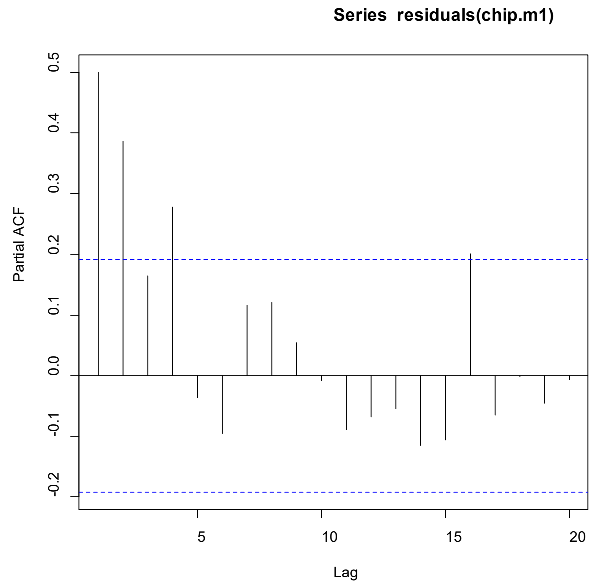
```
lm(formula = log.sales ~ price, data = bluebirdlite)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.47884 -0.13992  0.01661  0.11243  0.60085 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  13.7894     0.2345   58.81   <2e-16 ***
price        -2.1000     0.1348  -15.57   <2e-16 ***
```

(c) Plot the ACF and PACF of the residuals from the OLS fit in part (b). Also give the EACF of the residuals from the OLS fit in part (b). What do these tools tell you about the specification of the noise process?






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

The ACF values decay off gradually, while the PACF values cut off (become zero) after lag 4. This would indicate a possible AR(4) model for the noise process. The EACF possibly indicates an ARMA(1,1) noise process.

(d) Fit WLS regressions of log sales against price, specifying the following models for the noise process: ARMA(4,0); ARMA(1,1); and ARMA(4,1). What model is preferred, and why?

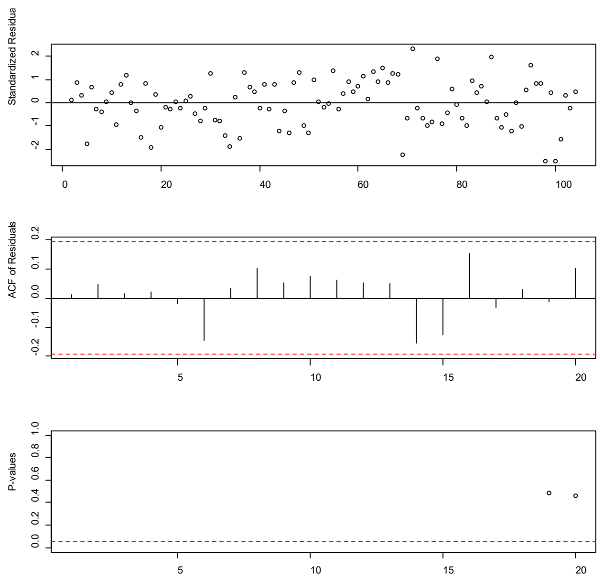
```
arima(x = log.sales, order = c(4, 0, 0), xreg = data.frame(price))
Coefficients:
         ar1     ar2     ar3     ar4  intercept    price
      0.1946  0.2189  0.1190  0.2946    13.4589  -1.9139
s.e.  0.0932  0.0938  0.0939  0.0925     0.2029   0.1093
sigma^2 estimated as 0.0207:  log likelihood = 53.52,  aic = -95.03

arima(x = log.sales, order = c(1, 0, 1), xreg = data.frame(price))
Coefficients:
        ar1      ma1  intercept    price
      0.943  -0.6700    13.4777  -1.9260
s.e.  0.037   0.0785     0.2223   0.1212
sigma^2 estimated as 0.02212:  log likelihood = 50.23,  aic = -92.45

arima(x = log.sales, order = c(4, 0, 1), xreg = data.frame(price))
Coefficients:
         ar1     ar2     ar3     ar4     ma1  intercept    price
      0.1365  0.2331  0.1373  0.3061  0.0640    13.4554  -1.9114
s.e.  0.2370  0.1077  0.1162  0.0994  0.2416     0.2031   0.1096
sigma^2 estimated as 0.02069:  log likelihood = 53.55,  aic = -93.1
```

We see the AIC is best for the ARMA(4,0) model. In particular, when we overfit with the ARMA(4,1), the MA-1 coefficient is nonsignificant, and the AIC becomes worse. We prefer the ARMA(4,0) model. Note that price has a significantly negative effect on log(sales) under this model.

(e) Look at the model diagnostics based on the residuals from your preferred model from (d). What do the diagnostics indicate?

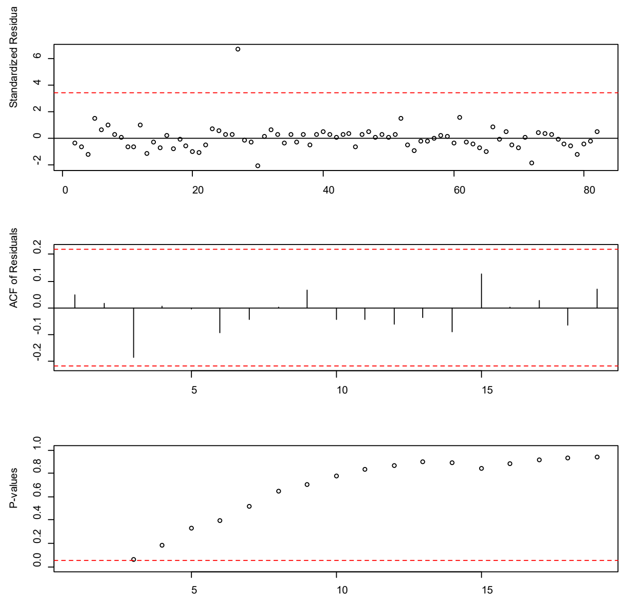


There are no excessive standardized residuals, and there are no significant autocorrelations. The residuals resemble white noise, indicating that the chosen model is a good one.

4) A data set of 82 measurements from a machining process are in the `deere1` object in the `TSA` package. Type `library(TSA); data(deere1); print(deere1)` in R to see the data set.

(a) Fit an AR(2) model to the full data set. Plot the standardized residuals from this model, and the sample ACF of the residuals. What do these diagnostics tell you?

```
arima(x = deere1, order = c(2, 0, 0))
Coefficients:
         ar1     ar2  intercept
      0.0269  0.2392     1.4135
s.e.  0.1062  0.1061     0.6275
sigma^2 estimated as 17.68:  log likelihood = -234.19,  aic = 474.38
```



There is no significant autocorrelation among the residuals, but there is one substantial outlier with a very large standardized residual.

(b) Detect either additive outliers and/or innovative outliers from the model in (a). What is your conclusion?

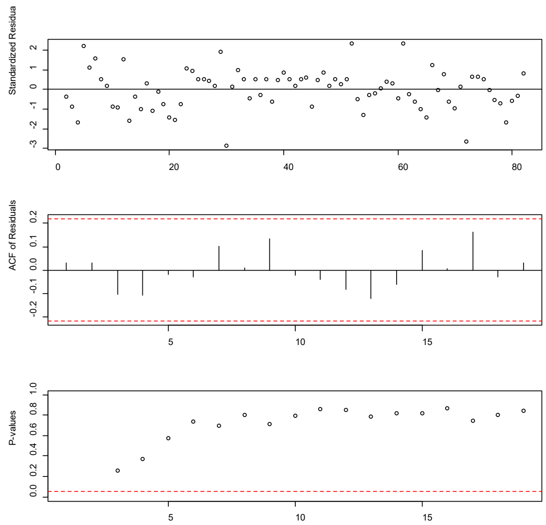
```
> detectAO(m1.deere)
              [,1]
ind      27.000000
lambda2  8.668582
> detectIO(m1.deere)
              [,1]
ind      27.000000
lambda1  8.816551
```

Observation 27 is an outlier. Since the lambda1 value exceeds the lambda2 value in magnitude, we can consider it an innovative outlier.

(c) Fit an AR(2) model that incorporates the most notable outlier into the model. Plot the standardized residuals from this model, and the sample ACF of the residuals. What do these diagnostics tell you? Compare the fitted model in part (a) to the fitted model in part (c).

```
arima(x = deere1, order = c(2, 0, 0), io = c(27))
Coefficients:
          ar1     ar2  intercept    IO-27
      -0.0143  0.2388     1.0848  27.1751
s.e.   0.1070  0.1103     0.4079   2.9114

sigma^2 estimated as 8.259:  log likelihood = -202.98,  aic = 413.95
```



The diagnostics now look fine. There is still no significant autocorrelation among the residuals, and now there are no remaining outliers. The AIC of the model in (c) in substantially better, and the estimate of the intercept term changes a great deal when we incorporate the outlier into the model.

~~5) For the intervention effect model $m_t = \delta m_{t-1} + \omega S_{t-1}^{(T)}$, find the half-life of the intervention effect when $\delta = 0.6$. Also find the half-life of the intervention effect when $\delta = 0.9$.~~

~~log(0.5)/ log(0.6) = 1.356915~~

~~log(0.5)/ log(0.9) = 6.578813~~