Chapter 12 Solutions

12.1 a. For a seasonal model the pattern would have to repeat at some regular time interval (e.g., every five years) that is predictable in advance. While many yearly series might go through increasing and decreasing cycles, it is much rarer for the timing of cycles to be on a fixed interval.

b. One example: Yearly spending on political advertising in the United States. Presidential elections happening every four years would cause a big increase in political advertising

12.2 The predicted values (and thus the residuals) are the same for the linear model, whether we use *Year* or t as the predictor. Note that for the Arctic sea ice data $t = Year - 1978$, so that $t = 1$ when $Year = 1979$. Substituting into the prediction equation for *Extent* based on t we have

$$\begin{aligned}\widehat{Extent} &= 8.008 - 0.08732Year \\ &= 8.008 - 0.08732(Year - 1978) \\ &= 8.008 - 0.08732Year + 0.08732 \cdot 1978 \\ &= 180.73 - 0.08732Year\end{aligned}$$

So the predictions from the linear function based on t are exactly the same as those from the linear function based on *Year*.

12.3 a. $Sales_t = \beta_0 + \beta_1 cos\left(\frac{2\pi t}{4}\right) + \beta_2 sin\left(\frac{2\pi t}{4}\right) + \epsilon$

b. $Price_t = \beta_0 + \beta_1 t + \beta_2 cos\left(\frac{2\pi t}{5}\right) + \beta_3 sin\left(\frac{2\pi t}{5}\right) + \epsilon$

c. $Riders_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 cos\left(\frac{2\pi t}{7}\right) + \beta_4 sin\left(\frac{2\pi t}{7}\right) + \epsilon$

d. $BP_t = \beta_0 + \beta_1 cos\left(\frac{2\pi t}{3}\right) + \beta_2 sin\left(\frac{2\pi t}{3}\right) + \epsilon$

12.4 a. $Sales_t = \beta_0 + \beta_1 Q_1 + \beta_2 Q_2 + \beta_3 Q_3 + \epsilon$

b. $Price_t = \beta_0 + \beta_1 t + \beta_2 Tues + \beta_3 Wed + \beta_4 Thur + \beta_5 Fri + \epsilon$

c. $Riders_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 Tues + \beta_4 Wed + \beta_5 Thur + \beta_6 Fri + \beta_7 Sat + \beta_8 Sun + \epsilon$

d. $BP_t = \beta_0 + \beta_1 FirstThird + \beta_2 SecondThird + \epsilon$

12.5 In general we look for extended increasing or decreasing trends to indicate that the mean is not constant over time and a regular difference is needed to help with stationarity.

a. In Series A, we see a clear decreasing trend, so a regular difference would be needed to help with stationarity.

b. In Series B, we see a clear increasing trend, so a regular difference would be needed to help with stationarity.

c. In Series C, there might be some seasonal patterns, but the overall mean level and variability look relatively constant, so we would probably not need a regular difference.

d. In Series D, we see a fairly random scatter about a mean of about 100, with no general increasing/decreasing trends, so we don't see any evidence for needing a regular difference.

e. In Series E, we see long stretches of generally increasing values, followed by decreasing trends, without evidence of movement around a constant mean, so a regular difference would be needed to help with stationarity.

12.6 For seasonal differences, we look for strong patterns that recur on a regular time interval.

a. In Series A, although there is a clear decreasing linear trend, the fluctuations above and below this trend look relatively random, without peaks and valleys occurring at regular times. This series would probably not need a seasonal difference.

b. In Series B, the general trend is increasing, but there is also a strong seasonal pattern with peaks and valleys occurring at regular intervals (actually the period is of length 12 as it would be for monthly data).

c. In Series C, there is no long-term increasing or decreasing trend, but the series appears to return to a high point at a regular pattern, with low points equally spaced in between the highs. This is likely a seasonal pattern that would benefit from a seasonal difference to help with stationarity. Although it's difficult to tell precisely from the scale on the graph, the period is of length 4 as we might see for quarterly data.

d. In Series D, we see no regular recurring patterns, so a seasonal difference would probably not be needed.

e. In Series E, we see some periods of general increasing and decreasing patterns, but not at any sort of regular time intervals, so seasonal differences would probably not be helpful.

12.7 Watch for a slow, linear decay in the early lags of an ACF as a sign that at least one regular difference is needed

a. In Series 1, we see a slow linear decay in the ACF from lags 1 through 20, so a regular difference would be needed to help with stationarity.

b. In Series 2, we also see slow linear decay in the ACF for even more lags than Series 1, so a regular difference would be needed to help with stationarity.

c. In Series 3, we see that none of the autocorrelations exceed the significance boundaries and all are between ± 0.2. No sign of a slow linear decay, so no indication that a regular difference is needed.

d. In Series 4, we see slow decay in the first few lags with some bumps that might be due to seasonality. Thus we would expect to need a regular difference (and a seasonal difference also).

e. In Series 5, we see slow linear decay, but only at seasonal intervals (positive spikes at lags 4, 8, 12, ... and negative spikes at lags 2, 6, 10, ...), not across all of the early lags. So we might need a seasonal difference, but don't have evidence here to suggest needing a regular difference.

12.8 For evidence in the ACF of needing seasonal differences we look for a slow, linear decaying pattern that occurs across lags that are a fixed difference apart.

a. In Series 1, we see a slow linear decay in the ACF but nothing evident at any seasonal lags, so we don't see any evidence there of needing a seasonal difference.

b. Series 2 has a pattern similar to Series 1, linear decay, but not seasonal pattern and so no evidence for needing a seasonal difference.

c. In Series 3, we see that none of the autocorrelations exceed the significance boundaries and all are between ± 0.2. Since there is clearly no seasonal pattern, we don't need a seasonal difference.

d. In Series 4, we see all positive autocorrelation, but the decreasing pattern is disturbed with bumps staying high at regular intervals (12, 24, 36, 48). This suggests a need for a seasonal difference at lag 12.

e. In Series 5, we see slow linear decay spread across the lags 4, 8, 12, ... and negative spikes decaying to 0 at the lags halfway between those (i.e., lags 2, 6, 10, ...). This suggests the need for trying a seasonal difference with a period of 4 (possibly quarterly data).

12.9 Let's start with the easier cases first.

The series that looks mostly random, with no consistent increasing/decreasing trends or seasonality is **Series D**, so its ACF should be the one with no significant autocorrelations, **Series 3**.

Series B is also pretty easy since it's the only one of the time series plot that shows both an inceasing trend and a clear seasonal pattern. This should correspond to the linear decay with regular seasonal bumps for the ACF plot of **Series 4**.

The only other plot with a clear seasonal trend is **Series C** with a period of around 4. This should match with the ACF for **Series 5**, which shows only a seasonal pattern, also with a period of 4.

Now for the trickiest pair. The ACFs for Series 1 and Series 2 both show a consistent linear decay with no seasonal patterns, and a somewhat more prolonged pattern for Series 2. Series A and Series E both appear to need regular differencing due to nonconstant means, but pattern is consistently

decreasing for Series A, while Series E has some different increasing and decreasing stretches. For this reason it makes sense to match the decreasing pattern of **Series A** with the longer decay of **Series 2**.

That leaves **Series E** to match with the somewhat shorter decay of **Series 1**.

12.10 Ice cream sales will increase in the summer months, when temperatures increase. Reservations at a beach hotel in a region of the country that experiences all four seasons will also increase when temperatures increase. New memberships at a gym will increase around the holidays when individuals make resolutions to get fit, and perhaps in the spring when bathing suit weather approaches. Cholesterol levels may increase as the author ages, but there will probably not be a seasonal pattern.

12.11 Household spending on gifts will increase near holidays, especially in December. The cost of a dental cleaning and prices for milk may increase, but seasonal patterns are not likely. Stock prices might fluctuate up and down, but probably not on a regular monthly pattern.

12.12 a. True, there will be day-to-day variation over short periods of time so we are definitely expecting the stock to fluctuate.

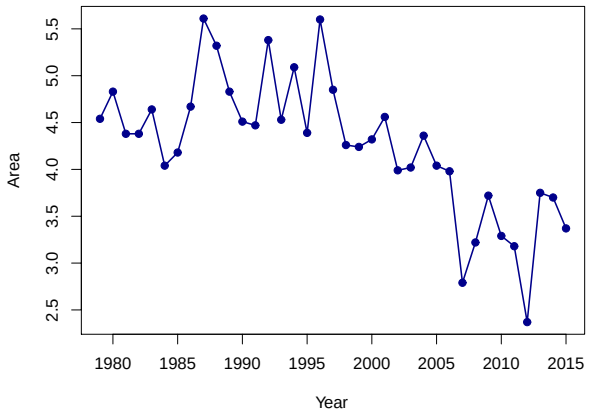
b. False, the reason we are investing money in the stock is that we are hoping to see an increase in the stock price over a long period of time.

12.13 a. False, athletic performance for a particular individual will be related to the overall health and fitness of that particular athlete.

b. False, there is no reason to believe that season influences performance. Note that performance may be influenced by extreme conditions, such as high winds or extreme temperatures (high or low).

c. True, as intensity increases, the body needs more fuel, so weight loss is likely if caloric intake is not increased.

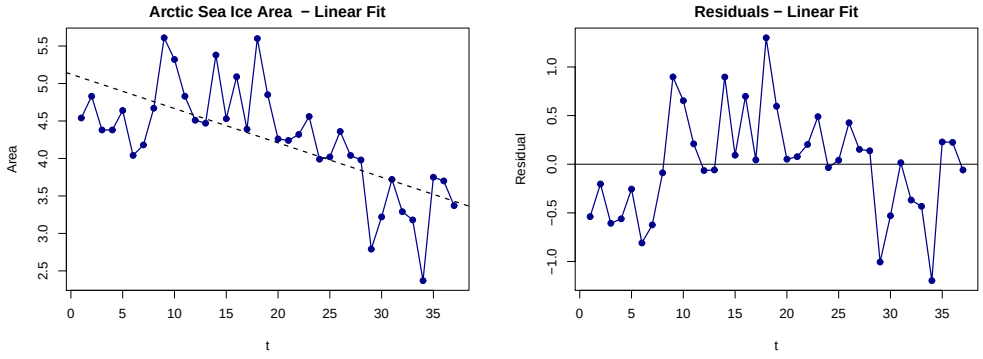
12.14 a. The following figure shows a time series for the *Area* variable in the **SeaIce** dataset. It shows a similar decreasing pattern to what was observed with the *Extent* variable.



b. Some output for fitting a linear model based on t for the sea ice *Area* follows. The prediction equation is $\widehat{Area} = 5.1246 - 0.04582t$.

```
 Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.124640   0.183255  27.965  < 2e-16
t           -0.045820   0.008408  -5.449 4.11e-06
---
Residual standard error: 0.5461 on 35 degrees of freedom
Multiple R-squared:  0.459,Adjusted R-squared:  0.4436
F-statistic:  29.7 on 1 and 35 DF,  p-value: 4.11e-06
```

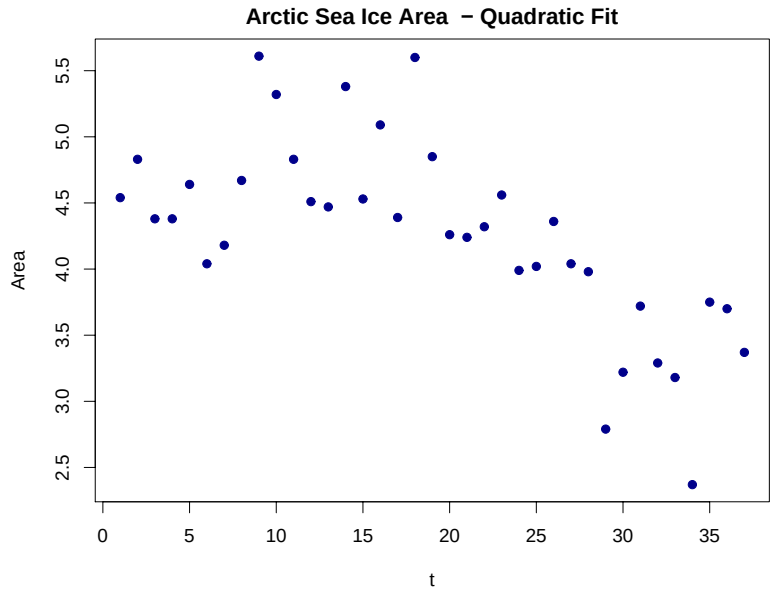
c. The plots below show the *Area* time series with a linear fit and a time series plot of the residuals. As with the *Extent* variable we see evidence of curvature with the rate of decrease increasing as time increases. The residual plot shows mostly negative residuals in early and late years, with mostly positive residuals in the middle years.



12.15 a. Some output for fitting a quadratic model based on t for the sea ice *Area* is shown as follows. The prediction equation is $\widehat{Area} = 4.45426 + 0.057315t - 0.0027141t^2$.

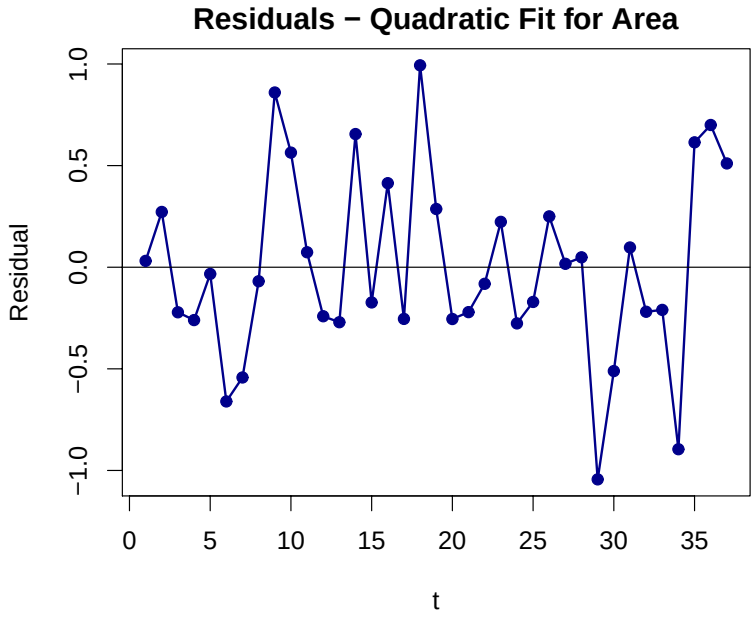
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.4542600  0.2465370  18.067  < 2e-16
t            0.0573150  0.0299179   1.916  0.06384
I(t^2)      -0.0027141  0.0007636  -3.554  0.00114
---
Residual standard error: 0.4731 on 34 degrees of freedom
Multiple R-squared:  0.6056,Adjusted R-squared:  0.5824
F-statistic:  26.1 on 2 and 34 DF,  p-value: 1.354e-07
```

Here is a plot of the *Area* time series with the quadratic fit.

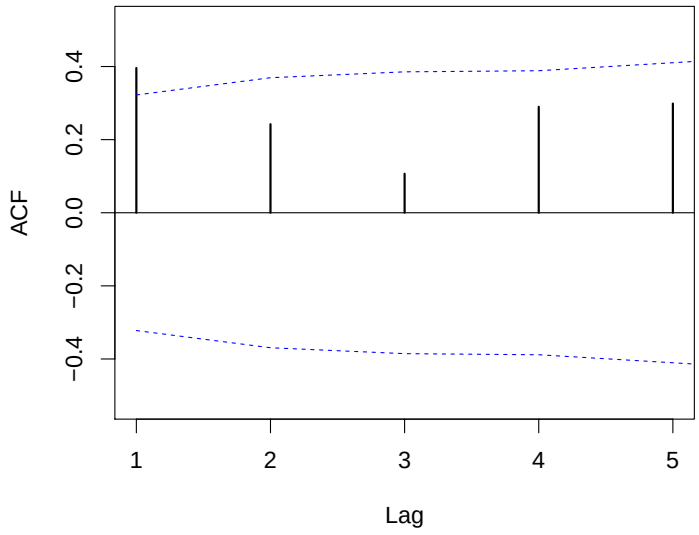


b. The P-value for testing the coefficient of the quadratic term is 0.00114, which is quite small. This indicates that the quadratic term is useful in the model. Also the R^2 increases from 45.9% in the linear model to 60.6% in the quadratic model and the adjusted R^2 improves from 44.4% to 58.2%.

c. A time series plot of the residuals from the quadratic model follows. It shows much better scatter above and below the zero line with no consistent curvature patterns.

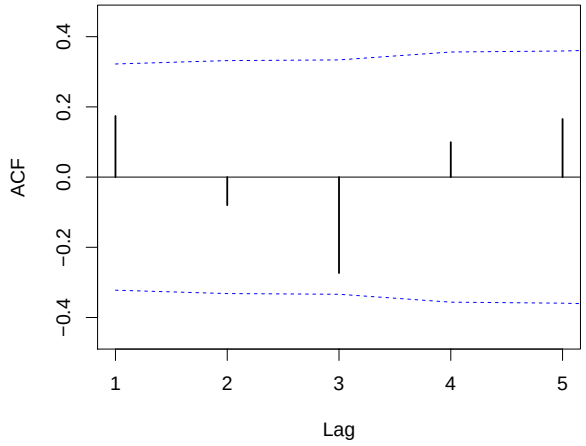


12.16 Here is an ACF plot for the first few lags of the residuals of the linear model to predict Arctic sea ice *Area*.



The lag 1 autocorrelation of the residuals, $r_1 = 0.396$, extends beyond the significance boundaries, so we have convincing evidence of a positive autocorrelation between adjacent residuals. This implies that the residuals for the linear model to predict *Area* are not independent.

12.17 Here is an ACF plot for the first few lags of the residuals of the quadratic model to predict Arctic sea ice *Area*.



The lag 1 autocorrelation of the residuals, $r_1 = 0.174$, does not go beyond the significance boundaries, so we do not have convincing evidence of lag 1 autocorrelation in the residuals. The same is true for the other autocorrelations in the plot, so we do not have concerns with the independence condition for this model.

12.18 a. Using technology for the linear model, we get the following output for forecasts and bounds for *Area* in 2017 ($t = 39$), 2019 ($t = 41$), and 2021 ($t = 43$).

```
 t       fit       lwr      upr
39  3.337648 2.163426 4.511871
41  3.246008 2.061409 4.430606
43  3.154367 1.958507 4.350227
```

In 2017, the model predicts the Arctic sea ice area to be 3.34 million km^2 and gives 95% prediction bounds from 2.16 and 4.51 million km^2. As the years get bigger, the predicted area gets smaller and the width of the forecast interval increases slightly. However, we noted problems with both the linearity and independence conditions for this model, so we should view these forecasts with some suspicion.

b. The prediction equation is $\widehat{Area} = 5.1246 - 0.04582t$. The predicted area is zero when $5.1246 - 0.04582t = 0$. This happens when $t = 5.1246/0.04582 = 111.8$ or after about 112 years from the start of the series. This happens around the year 2090. But this assumes that the linear model is appropriate (and it probably isn't) and that it will continue to hold for many years after the most recent data point, which is unlikely.

12.19 a. Using technology for the quadratic model, we get the following output for forecasts and bounds for *Area* in 2017 ($t = 39$), 2019 ($t = 41$), and 2021 ($t = 43$).

```
 t       fit       lwr      upr
39  2.561419 1.4505540 3.672284
41  2.241795 1.0648771 3.418713
43  1.900458 0.6396503 3.161267
```

In 2017, the model predicts the Arctic sea ice area to be 2.56 million km^2 and we are 95% confident it will be between 1.45 and 3.67 million km^2. As the years increase the predicted area gets smaller and the width of the forecast interval increases.

b. The prediction equation is $\widehat{Area} = 4.45426 + 0.05732t - 0.0027141t^2$. The predicted area is zero when $4.45426 + 0.05732t - 0.0027141t^2 = 0$. Using the quadratic formula this occurs when

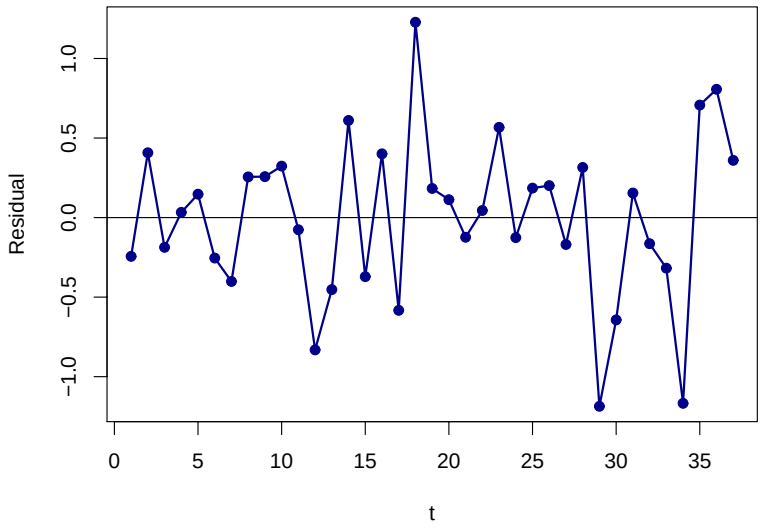
$$t = \frac{-0.05732 \pm \sqrt{0.05732^2 - 4(-0.0027141)(4.45426)}}{2(-0.0027141)} = -31.3 \text{ or } 52.4$$

The positive root is just before $t = 53$, which corresponds to the year 2031. But this assumes that the model will continue to hold for quite a few years after the most recent data point (2015), which may not be reasonable.

12.20 The fitted quadratic model is

$$\widehat{Extent} = 7.47 - 0.004622t - 0.002176t^2$$

Here is a time series plot of the residuals for this model.

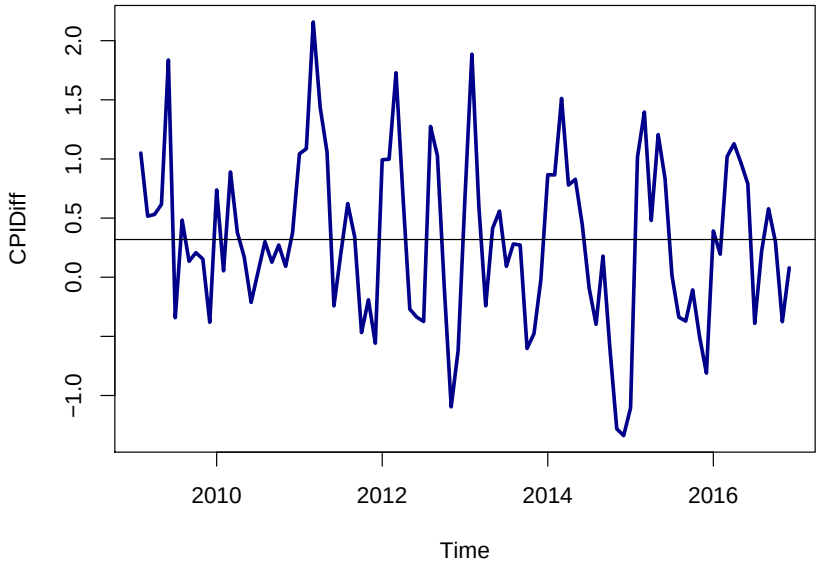


The plot shows no obvious curvature, so the quadratic model does an effective job of addressing the curvature that was apparent in the linear model.

12.21 Starting with a seasonal difference after a regular difference we have

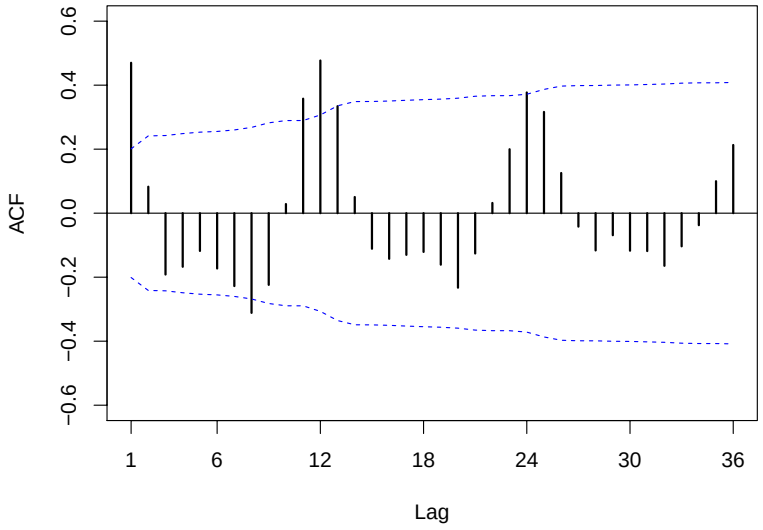
$$\begin{aligned}
\Delta_{12}(\Delta Y_t) &= \Delta_{12}(Y_t - Y_{t-1}) \\
&= (Y_t - Y_{t-1}) - (Y_{t-12} - Y_{t-13}) \\
&= Y_t - Y_{t-1} - Y_{t-12} + Y_{t-13} \\
&= Y_t - Y_{t-12} - Y_{t-1} + Y_{t-13} \\
&= (Y_t - Y_{t-12}) - (Y_{t-1} - Y_{t-13}) \\
&= \Delta_{12} Y_t - \Delta_{12} Y_{t-1} \\
&= \Delta(\Delta_{12} Y_t)
\end{aligned}$$

12.22 Here is a time series plot of the first differences (monthly changes) of the *CPI* series, $\Delta CPI_t = CPI_t - CPI_{t-1}$.



The series looks relatively stationary with no consistent increasing or decreasing trends over time. The differences appear to fluctuate around a value slightly above zero. The mean of the differences is about 0.32.

The autocorrelations for the *CPI* differences are shown below. The slow decay at early lags (seen in the ACF of the undifferenced *CPI* series) is now gone. We see a significant autocorrelation at lag 1 and some seasonal autocorrelations around lags 12 and 24.



Both of these plots are similar to the times series plot and ACF plot of the percent differences in example in the text. Stationarity looks much better than for the original *CPI* series.

12.23 a. Here is some output for fitting the ARIMA$(1, 1, 0)$ model to the *CPI* series.

```
Final Estimates of Parameters
Type          Coef  SE Coef     T      P
AR   1      0.4759   0.0912  5.22  0.000
Constant  0.16950  0.06426  2.64  0.010

Differencing: 1 regular difference
Number of observations:  Original series 96, after differencing 95
Residuals:    SS =  36.4774 (backforecasts excluded)
              MS =  0.3922  DF = 93
```

We see that the constant term is needed in the model (P-value $= 0.010$), so the fitted model is

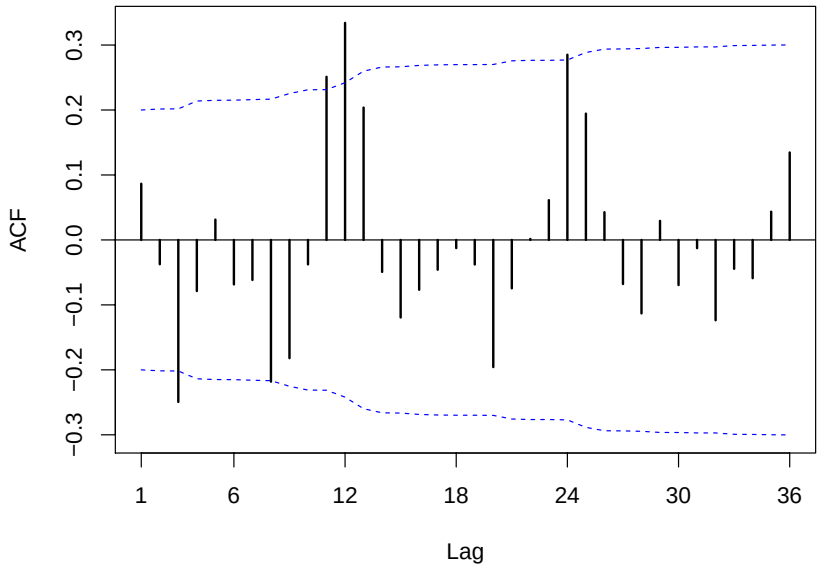
$$\Delta\widehat{CPI}_t = 0.1695 + 0.4759 \cdot \Delta CPI_{t-1}$$

We can rearrange terms to get

$$\begin{aligned} \widehat{CPI}_t - CPI_{t-1} &= 0.1695 + 0.4759(CPI_{t-1} - CPI_{t-2}) \\ \widehat{CPI}_t &= 0.1695 + 1.4759CPI_{t-1} - 0.4759CPI_{t-2} \end{aligned}$$

b. The P-value for the first order autoregressive term (ϕ_1) in the model is shown as 0.000, so this would appear to be an important term in the model.

c. Here is a plot of the autocorrelations of the residuals for the ARIMA$(1, 1, 0)$ model.



While the big spike at lag 1 that was present in the ACF plot of the *CPI* differences is gone, we still see significant autocorrelations for the residuals at lag 3 and 11, and around the seasonal lags of 12 and 24. The residuals do not appear to be independent. Perhaps we need some seasonal terms.

d. Here is some output when we add a second autoregressive term to fit an ARIMA$(2, 1, 0)$ model (with a constant term).

```
Final Estimates of Parameters
Type          Coef  SE Coef      T      P
AR   1      0.5611   0.1024   5.48  0.000
AR   2     -0.1815   0.1029  -1.76  0.081
Constant   0.20069  0.06354   3.16  0.002

Differencing: 1 regular difference
Number of observations:  Original series 96, after differencing 95
Residuals:    SS =  35.2784 (backforecasts excluded)
              MS =  0.3835  DF = 92
```

The P-value for the ϕ_2 term is 0.081, which is not significant at a 5% level, but would be significant at a 10% level. This term might be marginally helpful, but probably could be dropped.

12.24 a. Here is some output for fitting the ARIMA$(0, 1, 1)$ model to the *CPI* series.

```
Final Estimates of Parameters
Type          Coef  SE Coef      T      P
MA   1     -0.4309   0.0936  -4.61  0.000
Constant   0.32230  0.09279   3.47  0.001

Differencing: 1 regular difference
Number of observations:  Original series 96, after differencing 95
Residuals:    SS =  37.2022 (backforecasts excluded)
              MS =  0.4000  DF = 93
```

We see that the constant term is needed in the model (P-value $= 0.001$), so the fitted model is

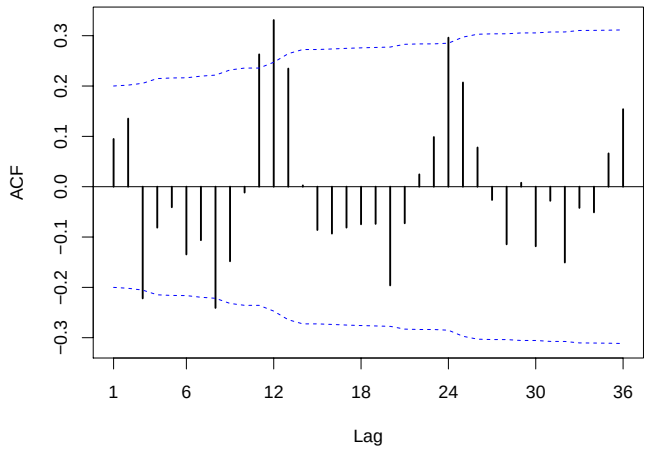
$$\Delta\widehat{CPI}_t = 0.3223 + 0.4309\epsilon_{t-1}$$

We can rearrange terms to get

$$\begin{aligned} \widehat{CPI}_t - CPI_{t-1} &= 0.3223 + 0.4309\epsilon_{t-1} \\ \widehat{CPI}_t &= 0.3223 + CPI_{t-1} + 0.4309\epsilon_{t-1} \end{aligned}$$

b. The P-value for the first order moving average term (θ_1) in the model is shown as 0.000, so this would appear to be an important term in the model.

c. Here is a plot of the autocorrelations of the residuals for the ARIMA$(0, 1, 1)$ model.



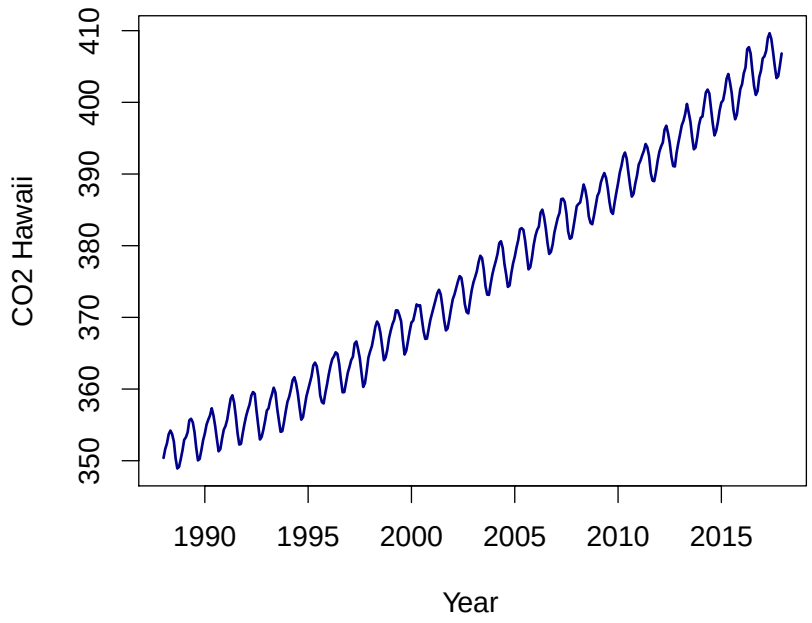
While the big spike at lag 1 that was present in the ACF plot of the *CPI* differences is gone, we still see significant autocorrelations for the residuals at lags 3, 8, and around the seasonal lags of 12 and 24. The residuals do not appear to be independent. Perhaps we need some seasonal terms.

d. Here is some output when we add a second moving average term to fit an ARIMA$(0, 1, 2)$ model (with a constant term).

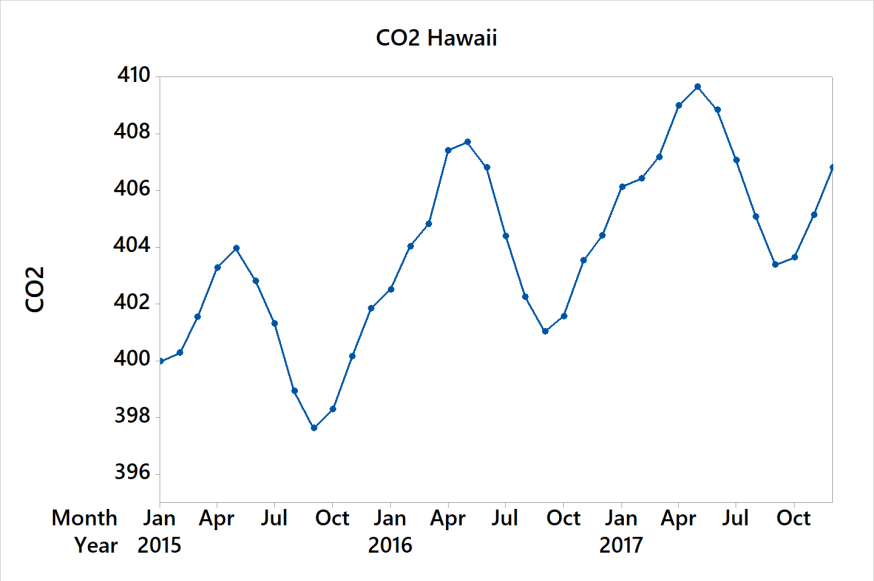
```
Final Estimates of Parameters
Type          Coef  SE Coef      T      P
MA   1    -0.5866   0.0978  -6.00  0.000
MA   2    -0.3546   0.0978  -3.63  0.000
Constant   0.3256   0.1212   2.69  0.009

Differencing: 1 regular difference
Number of observations:  Original series 96, after differencing 95
Residuals:    SS =  34.1969 (backforecasts excluded)
              MS =  0.3717  DF = 92
```

The P-value for the ϕ_2 term is reported as 0, so the second moving average term is important to keep in the model.

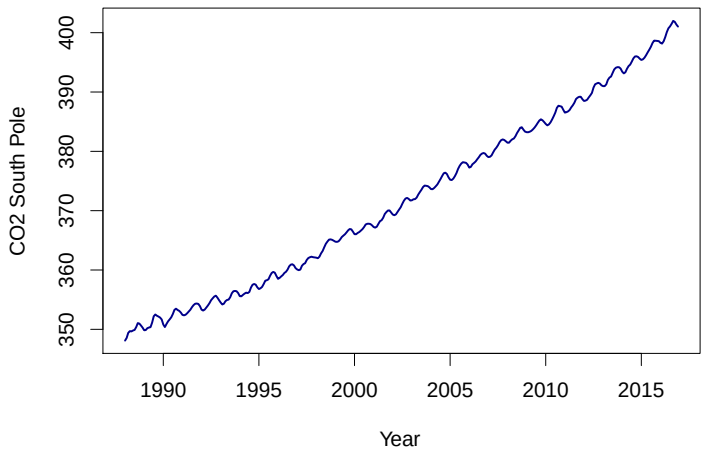
12.25 Here is a time series plot for the *CO2* variable in **CO2Hawaii**.



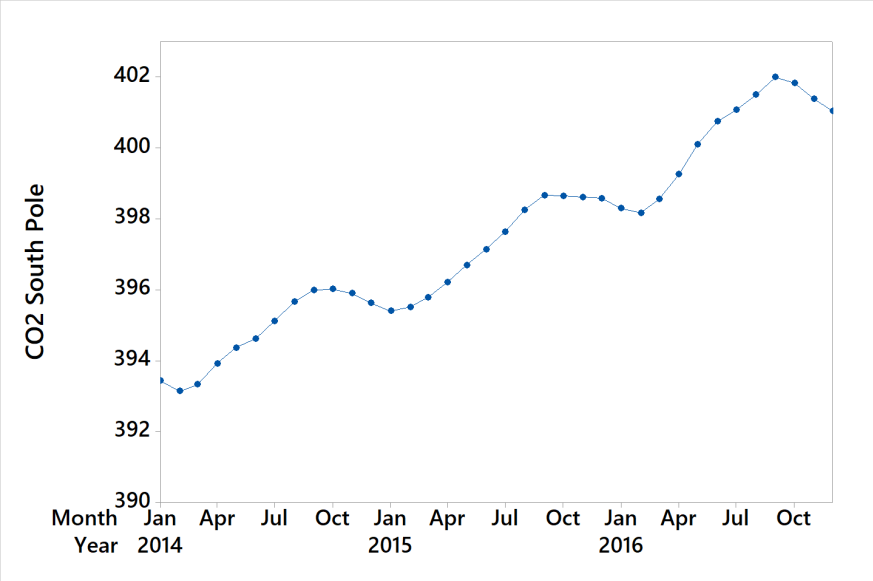
There is a clear, overall increasing trend—mostly linear with possibly a bit of upward curvature. There is also a clear seasonal pattern. The plot below for just the last three years shows the monthly pattern more clearly. The *CO2* levels tend to be highest in the late spring, then drop to the lowest values in early fall.



12.26 Here is a time series plot for the *CO2* variable in **CO2SouthPole**.



There is a clear, overall increasing trend—mostly linear with possibly a bit of upward curvature. There is also a slight seasonal pattern. The plot below for just the last three years shows the monthly pattern more clearly. The *CO2* levels tend to be highest in the early fall (September/October), then drop to the lowest values in mid-winter (January/February).



12.27 a. Here is some output for fitting the linear model using t to predict the *CO2* variable in **CO2Hawaii**.

```
  Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.474e+02  2.809e-01  1236.5   <2e-16
t           1.579e-01  1.349e-03   117.1   <2e-16
---
Residual standard error: 2.66 on 358 degrees of freedom
Multiple R-squared:  0.9745,Adjusted R-squared:  0.9745
F-statistic: 1.371e+04 on 1 and 358 DF,  p-value: < 2.2e-16
```

The fitted line is $\widehat{Y}_t = 347.4 + 0.1579t$.

b. The linear function explains a large amount of variability ($R^2 = 97.45\%$) in the *CO2* series.

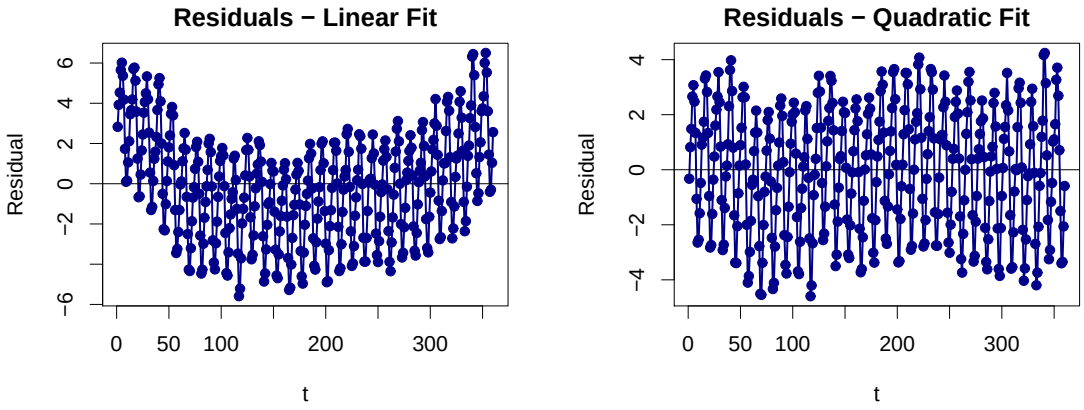
c. Here is some output for fitting the quadratic model using t to predict the *CO2* variable in **CO2Hawaii**.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.506e+02  3.574e-01  980.91   <2e-16
t           1.048e-01  4.572e-03   22.91   <2e-16
I(t^2)      1.473e-04  1.227e-05   12.01   <2e-16
---
```

```
Residual standard error: 2.248 on 357 degrees of freedom
Multiple R-squared:  0.9819,Adjusted R-squared:  0.9818
F-statistic:  9665 on 2 and 357 DF,  p-value: < 2.2e-16
```

The coefficient of the t^2 term produces a very large test statistic and small P-value, indicating that it is an important term to keep in the model. Also, the adjusted R^2 for the quadratic model (98.18%) is larger than for the linear model (97.45%). Finally, as the next part demonstrates, there is some curvature in the residuals of the linear model that is mostly gone in the quadratic residuals.

d. The residuals for both the linear and quadratic models are shown below.



Both residual plots still show a repeating seasonal pattern, but the overall trend of the quadratic residuals is centered consistently around zero, while the residuals for the linear model show some curvature.

12.28 a. Here is some output for fitting the linear model in t to the *CO2* variable in **CO2SouthPole**.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.458e+02  1.577e-01  2192.4   <2e-16
t           1.506e-01  7.833e-04   192.3   <2e-16
---
Residual standard error: 1.468 on 346 degrees of freedom
Multiple R-squared:  0.9907,Adjusted R-squared:  0.9907
F-statistic: 3.697e+04 on 1 and 346 DF,  p-value: < 2.2e-16
```

The fitted line is $\widehat{Y}_t = 345.8 + 0.1506t$.

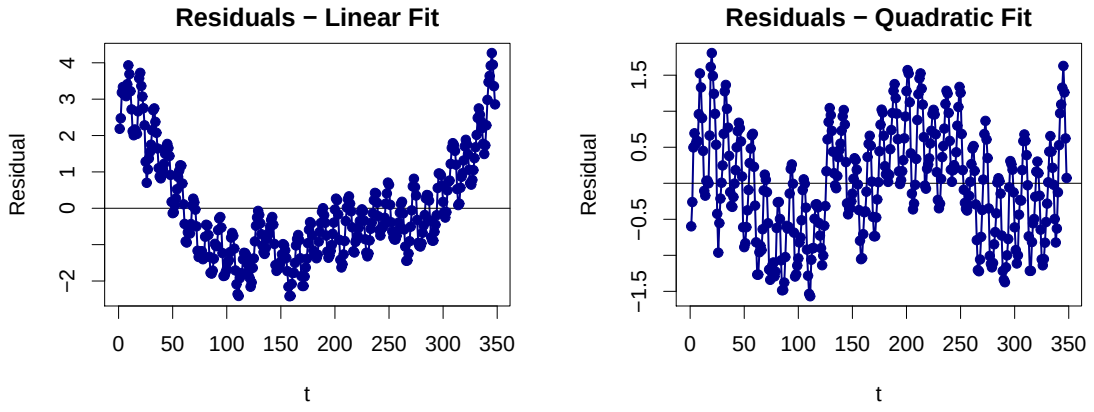
b. The linear function explains a large amount of variability ($R^2 = 99.07\%$) in the *CO2* series.

c. Here is some output for fitting the quadratic model in t to the *CO2* variable in **CO2SouthPole**.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.486e+02  1.222e-01 2851.91   <2e-16
t           1.021e-01  1.617e-03   63.10   <2e-16
I(t^2)      1.391e-04  4.488e-06   30.99   <2e-16
---
Residual standard error: 0.7557 on 345 degrees of freedom
Multiple R-squared:  0.9975,Adjusted R-squared:  0.9975
F-statistic: 7.022e+04 on 2 and 345 DF,  p-value: < 2.2e-16
```

The coefficient of the t^2 term produces a very large test statistic and small P-value, indicating that it is an important term to keep in the model. Also, the adjusted R^2 for the quadratic model (99.75%) is larger than for the linear model (99.07%). Finally, as the next part demonstrates, there is some curvature in the residuals of the linear model that is not so regular in the quadratic residuals.

d. The residuals for both the linear and quadratic models are shown below.



Both residual plots still show a repeating seasonal pattern (high and low points are relatively consistently spaced). The plot of residuals for the linear model also shows a clear curved pattern that is consistent across the graph. The residuals for the quadratic model show several milder, shorter increasing/decreasng trends. There might still be some structure to exploit in those residuals, but its not nearly as obvious as with the residuals for the linear model.

12.29 a. Here is some output for fitting the cosine plus quadratic trend to the *CO2* data in **CO2Hawaii**. The adjusted R^2 value is 99.76%.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.505e+02  1.308e-01 2679.51   <2e-16
t            1.052e-01  1.673e-03   62.89   <2e-16
I(t^2)       1.473e-04  4.489e-06   32.82   <2e-16
Xcos        -1.617e+00  6.132e-02  -26.37   <2e-16
Xsin         2.466e+00  6.134e-02   40.20   <2e-16
---
Residual standard error: 0.8227 on 355 degrees of freedom
Multiple R-squared:  0.9976,Adjusted R-squared:  0.9976
F-statistic: 3.666e+04 on 4 and 355 DF,  p-value: < 2.2e-16
```

b. Here is some output for fitting the seasonal means plus quadratic trend to the *CO2* data in **CO2Hawaii**. The adjusted R^2 value is 99.88%.

```
Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)      3.507e+02  1.349e-01 2600.647  < 2e-16
t                1.052e-01  1.166e-03   90.200  < 2e-16
I(t^2)           1.473e-04  3.128e-06   47.090  < 2e-16
factor(Month)2   6.408e-01  1.480e-01    4.328 1.97e-05
factor(Month)3   1.377e+00  1.480e-01    9.303  < 2e-16
factor(Month)4   2.561e+00  1.480e-01   17.297  < 2e-16
factor(Month)5   2.941e+00  1.480e-01   19.863  < 2e-16
factor(Month)6   2.129e+00  1.481e-01   14.379  < 2e-16
factor(Month)7   3.702e-01  1.481e-01    2.500   0.0129
factor(Month)8  -1.879e+00  1.481e-01  -12.692  < 2e-16
factor(Month)9  -3.558e+00  1.481e-01  -24.034  < 2e-16
factor(Month)10 -3.515e+00  1.481e-01  -23.740  < 2e-16
factor(Month)11 -2.254e+00  1.481e-01  -15.224  < 2e-16
factor(Month)12 -9.761e-01  1.481e-01   -6.592 1.62e-10
---
Residual standard error: 0.5734 on 346 degrees of freedom
Multiple R-squared:  0.9989,Adjusted R-squared:  0.9988
F-statistic: 2.325e+04 on 13 and 346 DF,  p-value: < 2.2e-16
```

c. For the cosine + quadratic model:

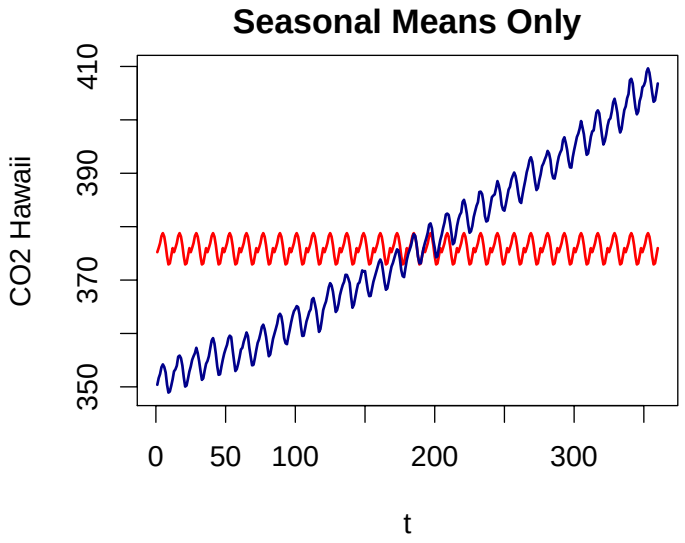
$\widehat{Y}_{370} = 350.5 + 0.1052(370) + 0.0001473(370^2) - 1.617\cos(2\pi 370/12) + 2.466\sin(2\pi 370/12) = 406.7$

For the seasonal means + quadratic model:

$$\widehat{Y}_{370} = 350.7 + 0.1052(370) + 0.0001473(370^2) + 0.3703(1) = 410.2$$

d. Although the seasonal means model has many more parameters, its adjusted R^2 is still higher than the model with the cosine trend (and both are very high). Furthermore the residual standard error for the seasonal means model (0.5734) is quite a bit less than for the cosine model (0.8227).

e. If we drop the t and t^2 terms from the seasonal means model in (b), the R^2 drops from 99.89% to 1.25% (and the adjusted R^2 becomes negative!). Clearly although the seasonal component is useful, the main pattern driving the *CO2* levels in Hawaii is the quadratic trend. The plot below shows the seasonal means only fits as a horizontal pattern that captures the month-to-month variation, but completely misses the overall increasing trend.



12.30 a. Here is some output for fitting the cosine plus quadratic trend to the *CO2* data in **CO2SouthPole**. The adjusted R^2 value is 99.87%.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.486e+02  8.991e-02 3877.33   <2e-16
t            1.019e-01  1.190e-03   85.68   <2e-16
I(t^2)       1.391e-04  3.301e-06   42.14   <2e-16
Xcos        -5.857e-02  4.214e-02   -1.39    0.165
Xsin        -7.213e-01  4.215e-02  -17.11   <2e-16
```

```
---
Residual standard error: 0.5558 on 343 degrees of freedom
Multiple R-squared:  0.9987,Adjusted R-squared:  0.9987
F-statistic: 6.498e+04 on 4 and 343 DF,  p-value: < 2.2e-16
```

b. Here is some output for fitting the seasonal means plus quadratic trend to the *CO2* data in **CO2SouthPole**. The adjusted R^2 value is 99.87%.

```
Coefficients:
                   Estimate Std. Error  t value Pr(>|t|)
(Intercept)      3.481e+02  1.325e-01 2626.702  < 2e-16
t                1.019e-01  1.186e-03   85.958  < 2e-16
I(t^2)           1.391e-04  3.290e-06   42.284  < 2e-16
factor(Month)2  -1.804e-01  1.455e-01   -1.240 0.215694
factor(Month)3  -8.945e-02  1.455e-01   -0.615 0.539042
factor(Month)4   1.347e-01  1.455e-01    0.926 0.355111
factor(Month)5   3.421e-01  1.455e-01    2.351 0.019294
factor(Month)6   5.312e-01  1.455e-01    3.651 0.000303
factor(Month)7   8.407e-01  1.455e-01    5.779 1.73e-08
factor(Month)8   1.191e+00  1.455e-01    8.189 5.66e-15
factor(Month)9   1.350e+00  1.455e-01    9.279  < 2e-16
factor(Month)10  1.253e+00  1.455e-01    8.611 2.91e-16
factor(Month)11  9.657e-01  1.455e-01    6.637 1.29e-10
factor(Month)12  4.946e-01  1.455e-01    3.399 0.000758
---
Residual standard error: 0.554 on 334 degrees of freedom
Multiple R-squared:  0.9987,Adjusted R-squared:  0.9987
F-statistic: 2.013e+04 on 13 and 334 DF,  p-value: < 2.2e-16
```

c. For the cosine + quadratic model:

$$\widehat{Y}_{360} = 348.6 + 0.1019(360) + 0.0001391(360^2) - 0.0587\cos(2\pi 360/12) - 0.7213\sin(2\pi 360/12) = 403.3$$

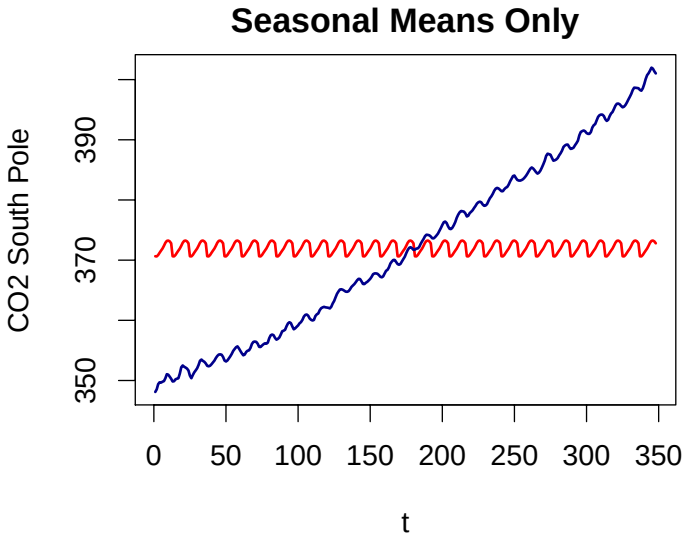
For the seasonal means + quadratic model:

$$\widehat{Y}_{360} = 348.1 + 0.1019(360) + 0.0001391(360^2) + 0.4946(1) = 403.3$$

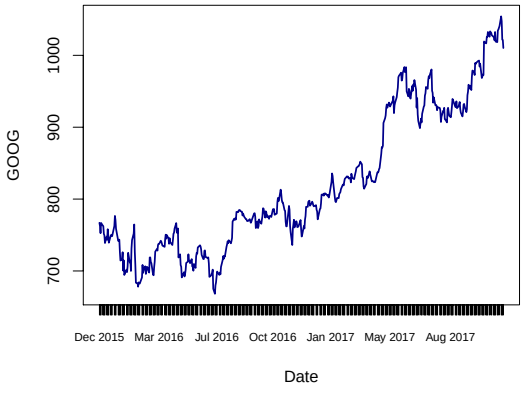
d. The seasonal means model has many more parameters, but its adjusted R^2 matches that of the model with cosine trend (at least to two decimals places as a percentage). We might also look at the residual standard error, but there's not much difference there either, 0.5558

for the cosine model and 0.554 for the seasonal means. Either model would be a reasonable choice. As the forecast in part (c) indicates, even the predicted values are very similar for the two models.

e. If we drop the t and t^2 terms from the seasonal means model in (b) the R^2 drops from 99.87% to 0.42% (and the adjusted R^2 becomes negative!). The results are very similar if you use only the sine and cosine terms from the model in (a). Clearly, although the seasonal component is useful, the main pattern driving the *CO2* levels at the South Pole is the quadratic trend. The plot below shows the seasonal means only fits as a horizontal pattern that captures the month-to-month variation, but completely misses the overall increasing trend.



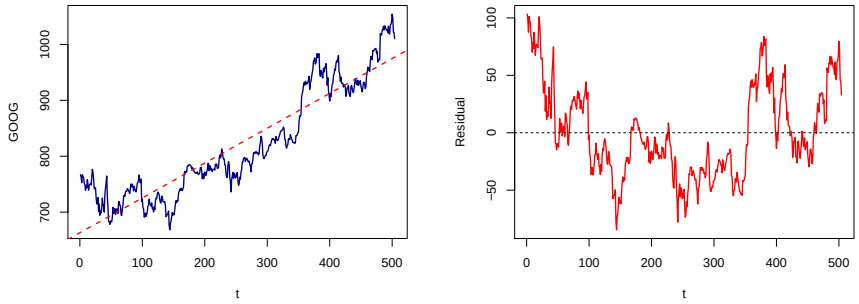
12.31 a. Here is a plot of the Google price time series. There is a general increasing trend over this two-year period.



b. Some output for fitting the linear model is shown below. The proportion of variability in the price series explained by this model is $R^2 = 84.12\%$.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 663.10410    3.52114  188.32   <2e-16
t             0.62319    0.01208   51.58   <2e-16
---
Residual standard error: 39.47 on 502 degrees of freedom
Multiple R-squared:  0.8412,Adjusted R-squared:  0.8409
F-statistic:  2660 on 1 and 502 DF,  p-value: < 2.2e-16
```

c. Here are plots of the original series with the least squares line and a time series plot of the residuals. Both plots show some curvature with points being above the line in the early days, below for most of the days in the middle, and tending above again near the end.



d. Here is some output for fitting the quadratic model. The prediction equation is

$$\widehat{y}_t = 723.5 - 0.09277t + 0.001418t^2$$

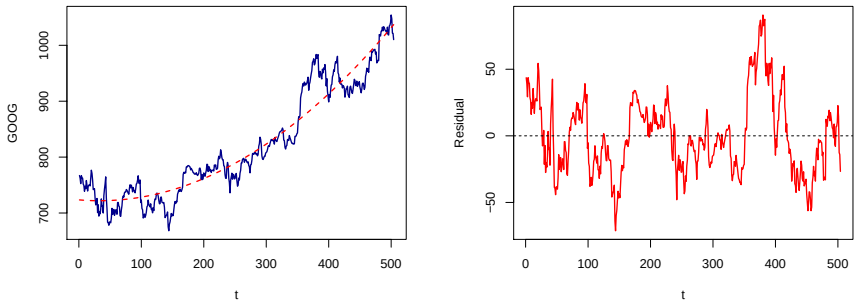
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.235e+02  3.879e+00 186.524  < 2e-16
t           -9.277e-02  3.547e-02  -2.615  0.00918
I(t^2)       1.418e-03  6.802e-05  20.843  < 2e-16
---
Residual standard error: 28.91 on 501 degrees of freedom
Multiple R-squared:  0.915,Adjusted R-squared:  0.9146
F-statistic:  2696 on 2 and 501 DF,  p-value: < 2.2e-16
```

Here is some output for fitting a linear model to the log of the Google stock prices. The prediction equation is

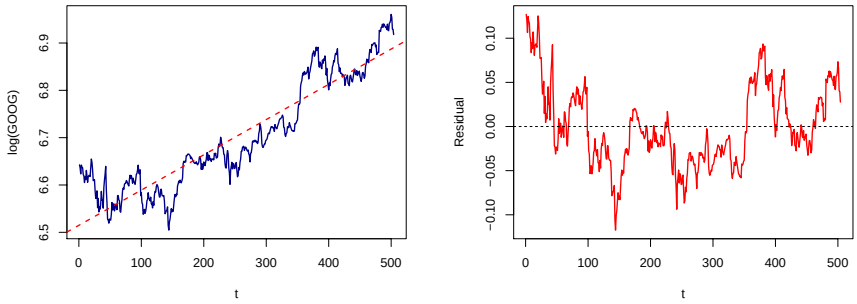
$$\widehat{log(y_t)} = 6.515 + 0.0007449t$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.515e+00  4.059e-03 1604.91   <2e-16
t           7.449e-04  1.393e-05   53.48   <2e-16
---
Residual standard error: 0.0455 on 502 degrees of freedom
Multiple R-squared:  0.8507,Adjusted R-squared:  0.8504
F-statistic:  2860 on 1 and 502 DF,  p-value: < 2.2e-16
```

e. Here are plots of the fitted values for the quadratic model with the original series and its residuals. The curvature issues from the linear model have been dealt with in the quadratic model. We see more random scatter above and below the line as the series moves along.



Here are plots of the fitted values for the linear model with the log of prices and its residuals. This model has similar issues with curvature as the linear model in the original scale, values above the line early, below in the middle, and mostly above at the end.

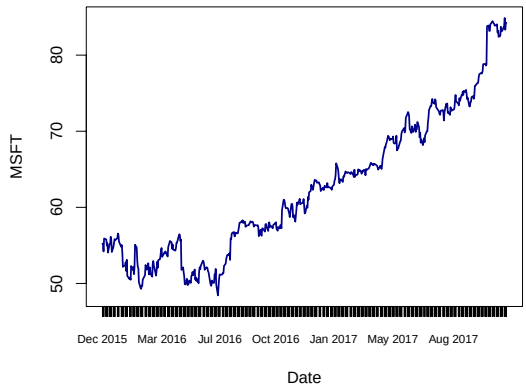


f. As we see in part (e), the quadratic model does a better job of addressing the lack of linearity in the original model, so we use that to make the forecast. Here is some output with a forecast and 95% prediction bounds for $t = 514$ (December 15, 2017).

```
        fit      lwr      upr
1 1050.36 992.9724 1107.747
```

Based on the quadratic model, we predict the Google stock price to be \$1050.36 on December 15, 2017 and are 95% sure it will be between \$992.97 and \$1107.75. Note: The actual closing price that day was \$1,064.19, which is fairly close to the forecast and easily within the prediction interval.

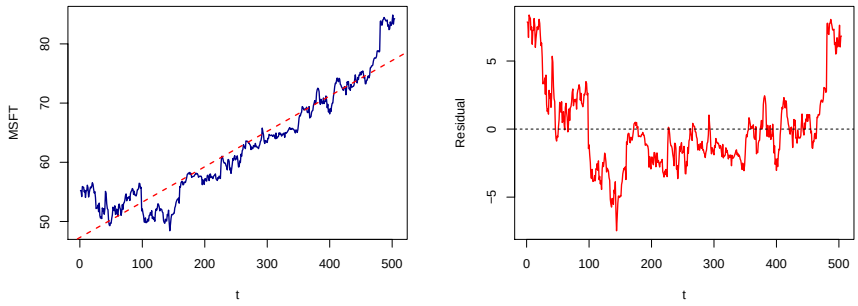
12.32 a. Here is a plot of the Microsoft price time series. There is a general increasing trend over this two-year period.



b. Some output for fitting the linear model is shown below. The proportion of variability in the price series explained by this model is $R^2 = 89.18\%$.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.729e+01  2.710e-01  174.51   <2e-16
t           5.980e-02  9.299e-04   64.31   <2e-16
---
Residual standard error: 3.037 on 502 degrees of freedom
Multiple R-squared:  0.8918,Adjusted R-squared:  0.8915
F-statistic:  4136 on 1 and 502 DF,  p-value: < 2.2e-16
```

c. Here are plots of the original series with the least squares line and a time series plot of the residuals. Both plots show some curvature with points being above the line in the early days, decreasing to below for most of the days in the middle, and increasing to above again near the end.



d. Here is some output for fitting the quadratic model. The prediction equation is

$$\widehat{y}_t = 52.55 - 0.002589t + 0.0001235t^2$$

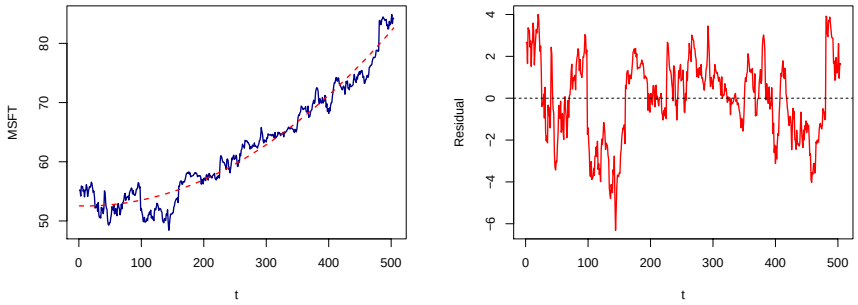
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.255e+01  2.594e-01 202.550   <2e-16
t           -2.589e-03  2.373e-03  -1.091    0.276
I(t^2)       1.235e-04  4.550e-06  27.153   <2e-16
---
Residual standard error: 1.934 on 501 degrees of freedom
Multiple R-squared:  0.9562,Adjusted R-squared:  0.956
F-statistic:  5469 on 2 and 501 DF,  p-value: < 2.2e-16
```

Here is some output for fitting a linear model to the log of the Microsoft stock prices. The prediction equation is

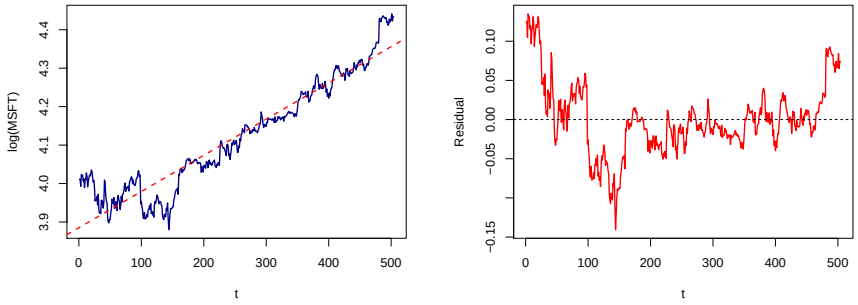
$$\widehat{log(y_t)} = 3.885 + 0.0009417t$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.885e+00  4.034e-03  963.11   <2e-16
t           9.417e-04  1.384e-05   68.03   <2e-16
---
Residual standard error: 0.04521 on 502 degrees of freedom
Multiple R-squared:  0.9021,Adjusted R-squared:  0.902
F-statistic:  4628 on 1 and 502 DF,  p-value: < 2.2e-16
```

e. Here are plots of the fitted values for the quadratic model with the original series and its residuals. The curvature issues from the linear model have been dealt with in the quadratic model. We see more random scatter above and below the line as the series moves along.



Here are plots of the fitted values for the linear model with the log of prices and its residuals. This model has similar issues with curvature, especially in the early days, as the linear model in the original scale (but not quite as severe).

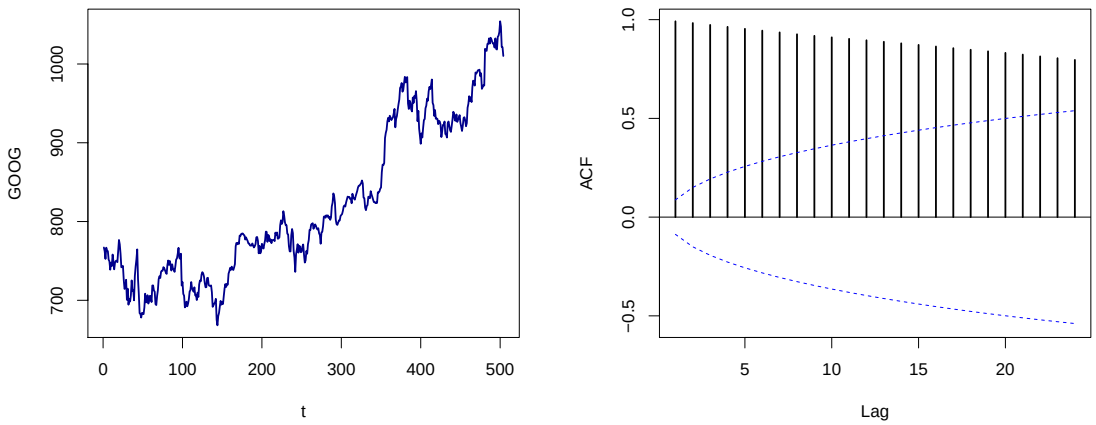


f. As we see in part (e), the quadratic model does a better job of addressing the lack of linearity in the original model, so we use that to make the forecast. Here is some output with a forecast and 95% prediction bounds for $t = 514$ (December 15, 2017).

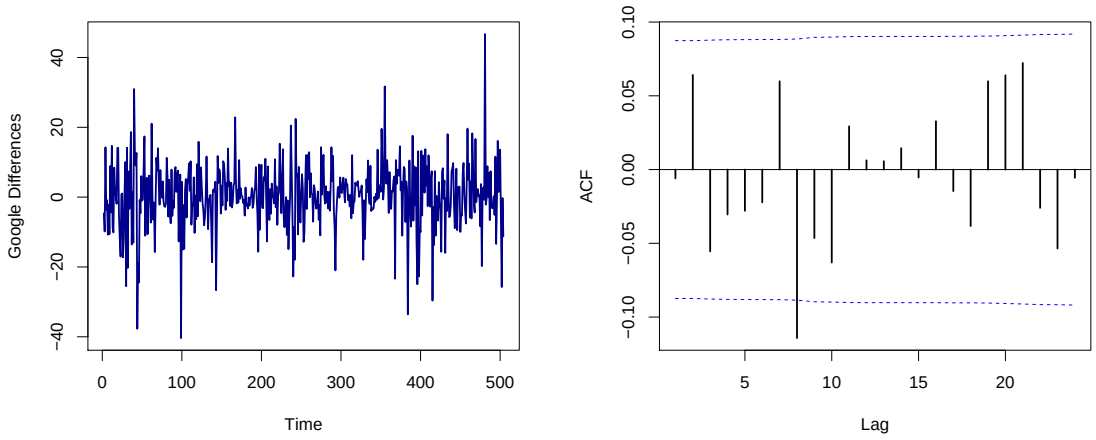
```
       fit      lwr      upr
1 83.85967 80.02105 87.69829
```

Based on the quadratic model, we predict the Microsoft stock price to be \$83.86 on December 15, 2017, and are 95% sure it will be between \$80.02 and \$87.70. Note: The actual closing price that day was \$86.85, which is within the prediction interval.

12.33 a. Here are plots of the Google price series and its ACF. There is a clear increasing trend in the prices over the two-year period and the ACF shows a slow linear decay. This is not a stationary series.



b. Here are plots of the first differences for the Google stock prices and their ACF. Stationarity looks much better now. There are no consistent increasing/decreasing trends in the time series plot and only one "significant" spike in the ACF.



c. The mean of the differences is 0.483. This means that, on average, the Google stock price increased by about 48 cents per day during this period.

d. From the ACF in (b) the only lag with an autocorrelation beyond the significance bounds is lag 8.

e. Ignoring the one spike at lag 8, the differences appear to be relatively independent already. This would imply that we don't need any autoregressive or moving average terms and should try an ARIMA$(0, 1, 0)$ model with a constant term.

$$\Delta y_t = \delta + \epsilon_t \qquad \text{or} \qquad y_t = \delta + y_{t-1} + \epsilon_t$$

This is a random walk with constant drift.

f. Here is some output for fitting the model.

```
ARIMA(0,1,0) with drift

Coefficients:
       drift
      0.4834
s.e.  0.4115
```

The forecast equation is $\hat{y}_t = 0.483 + y_{t-1}$.

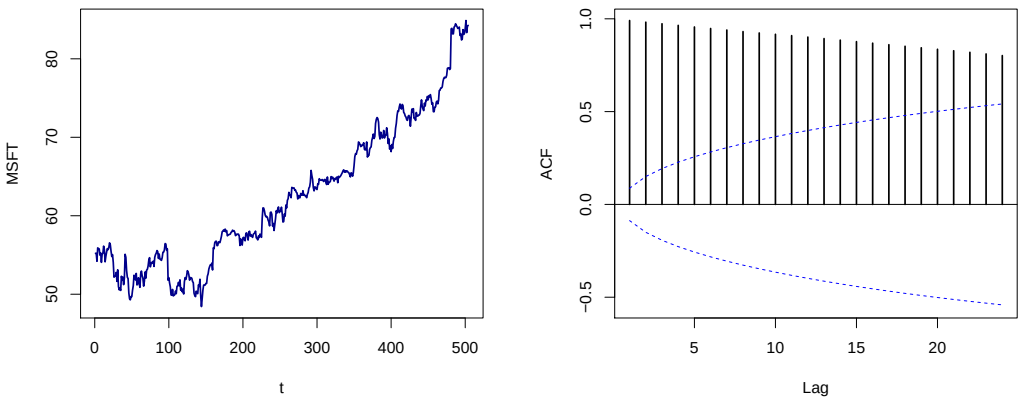
g. The constant term of the model is the same as the mean of the differences. For each new day's forecast we merely add that mean difference to the previous day's value.

h. Here are forecast and bounds for the next 10 days of the Google price series.

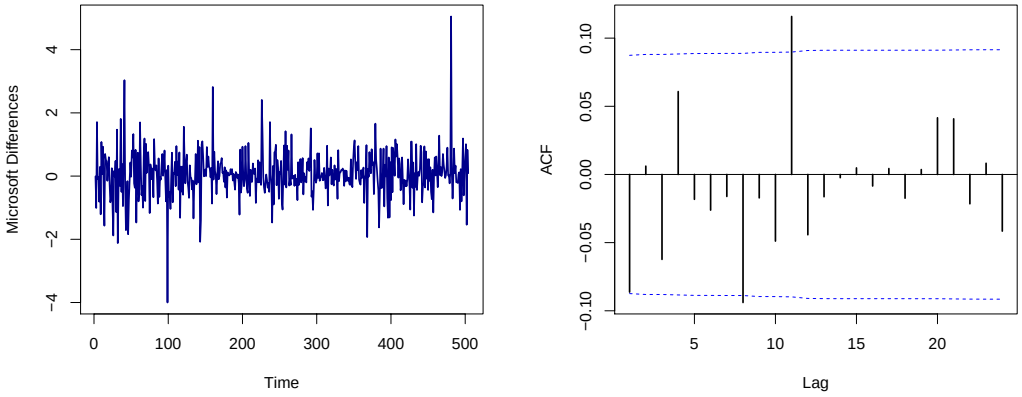
```
    Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
505       1010.653 998.8134 1022.493 992.5457 1028.761
506       1011.137 994.3925 1027.881 985.5287 1036.745
507       1011.620 991.1127 1032.127 980.2568 1042.983
508       1012.103 988.4236 1035.783 975.8882 1048.319
509       1012.587 986.1119 1039.062 972.0969 1053.077
510       1013.070 984.0684 1042.072 968.7157 1057.425
511       1013.554 982.2280 1044.879 965.6453 1061.462
512       1014.037 980.5485 1047.525 962.8208 1065.253
513       1014.520 979.0004 1050.040 960.1974 1068.843
514       1015.004 977.5624 1052.445 957.7423 1072.265
```

For December 15, 2017 ($t = 514$) the forecast price is \$1015.00 and we are 95% confident the price will be between \$957.74 and \$1072.27. Note: The actual closing price that day was \$1064.19, which is within the prediction interval.

12.34 a. Here are plots of the Microsoft price series and its ACF. There is a clear increasing trend in the prices over the two-year period and the ACF shows a slow linear decay. This is not a stationary series.



b. Here are plots of the first differences for the Microsoft stock prices and find their ACF. Stationarity looks much better now. There are no consistent increasing/decreasing trends in the time series plot and only two "significant" spikes in the ACF.



c. The mean of the differences is 0.0577. This means that, on average, the Microsoft stock price increased by about 5.77 cents per day during this period.

d. From the ACF in (b), the only lags with an autocorrelation beyond the significance bounds are lags 8 and 11.

e. Ignoring the somewhat random spikes at lag 8, the differences appear to be relatively independent already. This would imply that we don't need any autoregressive or' moving average terms and should try an ARIMA$(0, 1, 0)$ model with a constant term.

$$\Delta y_t = \delta + \epsilon_t \qquad \text{or} \qquad y_t = \delta + y_{t-1} + \epsilon_t$$

This is a random walk with constant drift.

f. Here is some output for fitting the model.

```
ARIMA(0,1,0) with drift

Coefficients:
       drift
      0.0577
s.e.  0.0322
```

The forecast equation is $\hat{y}_t = 0.0577 + y_{t-1}$.

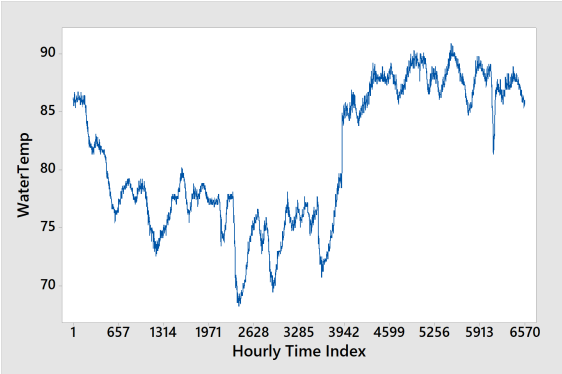
g. The constant term of the model is the same as the mean of the differences. For each new day's forecast we merely add that mean difference to the previous day's value.

h. Here are the forecast and bounds for the next 10 days of the Microsoft price series.

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
505	84.31773	83.39274	85.24273	82.90308	85.73239
506	84.37547	83.06733	85.68361	82.37484	86.37609
507	84.43320	82.83107	86.03534	81.98295	86.88346
508	84.49093	82.64095	86.34092	81.66162	87.32025
509	84.54867	82.48032	86.61702	81.38540	87.71193
510	84.60640	82.34064	86.87216	81.14122	88.07159
511	84.66414	82.21683	87.11144	80.92131	88.40696
512	84.72187	82.10559	87.33815	80.72062	88.72312
513	84.77960	82.00462	87.55458	80.53564	89.02357
514	84.83734	81.91225	87.76242	80.36380	89.31087

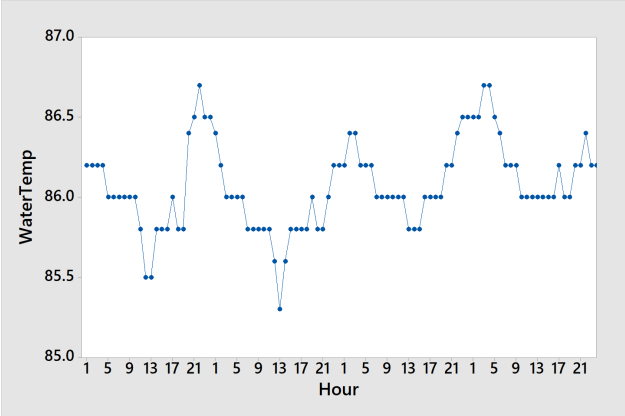
For December 15, 2017 ($t = 514$) the forecast price is \$84.84 and we are 95% confident the price will be between \$80.36 and \$89.31. Note: The actual closing price that day was \$86.85, which is within the prediction interval.

12.35 a. Here is a time series plot of the full series.



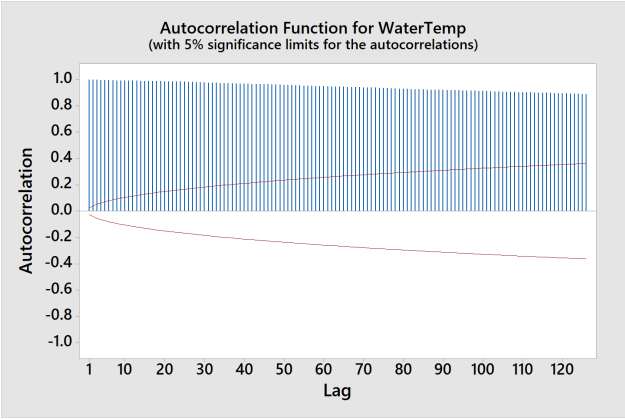
There are several interesting features in this plot. The temperatures decrease initially, which makes sense because our starting point is October 3. As we move through the winter months, the temperatures stay lower than our starting point. Moving into the summer, we see an overall increase in the temperatures. We also notice a jump in the temperatures around the hourly time index of 3942. The last temperature, October 3, 2017, is fairly close to the initial temperature on the same day in 2016. The temperature is clearly not stationary.

b. Here is a plot for the first 96 hours (four days) of the *WaterTemp* series.



A seasonal pattern is apparent within days ($S = 24$ hours), with lower water temperatures around hour 13 each day and higher temperatures roughly 12 hours after the lows.

c. Here is an ACF plot for the WaterTemp series.

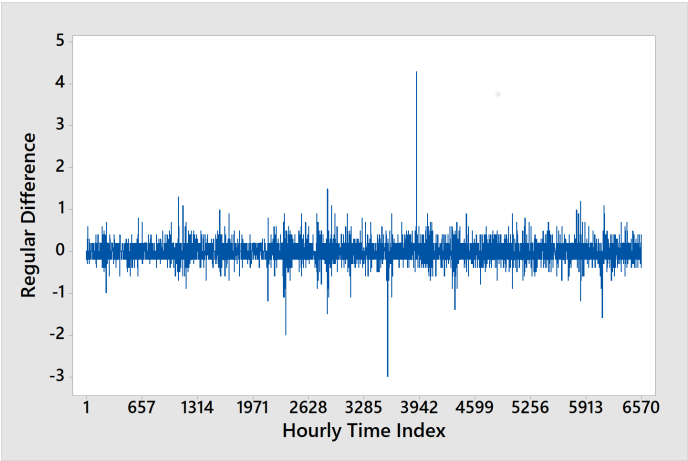


The slow steady decay indicates that the temperatures are not stationary, and we will need to consider regular difference. Since it takes a long time to change the temperature of an ocean and we are collecting data every hour, it is not surprising that we have very strong autocorrelation, even as the lag increases to over 120, only 5 days.

12.36 a. The mean of the differences is very close to zero, −0.00006.

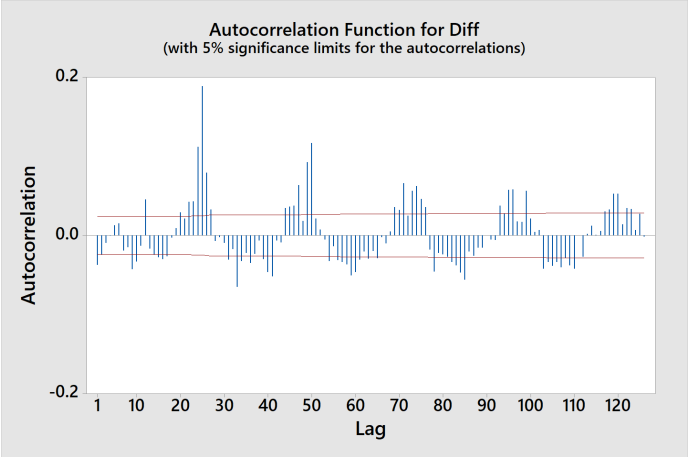
b. The standard deviation of the differences is 0.221, so 4.3 is approximately 19.5 (4.3/0.221) standard deviations above 0. This difference of 4.3 would be considered an outlier by any rule of thumb for identifying unusual observations.

c. The time series plot for the differences looks much different than the time series plot for *WaterTemp*. The series is now centered at 0 with no obvious drifts. Here is the plot.



d. Yes, the regular difference created a more stationary series that is centered at 0 with no drift.

e. The ACF plot that follows, is much different than the ACF plot for *WaterTemp*.



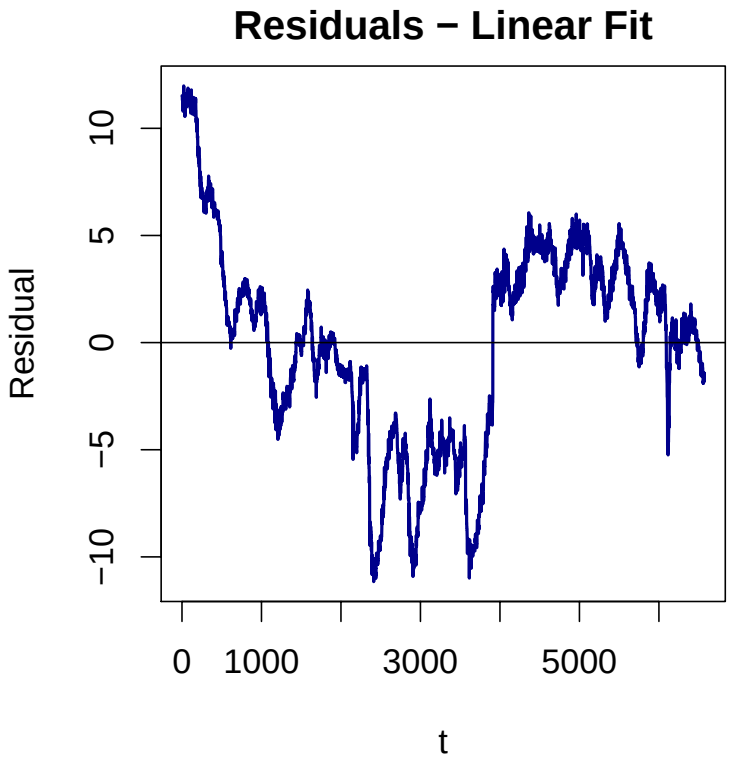
We still see some significant autocorrelations, especially at 1, 2 and near 12, 24, 36, 48, . . ., with strong autocorrelations at multiples of 24.

12.37 a. Here is some output for fitting the linear model in t.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.468e+01  1.163e-01  642.25   <2e-16 ***
t           1.940e-03  3.064e-05   63.31   <2e-16 ***
---
Residual standard error: 4.713 on 6570 degrees of freedom
Multiple R-squared:  0.3789,Adjusted R-squared:  0.3788
F-statistic:  4008 on 1 and 6570 DF,  p-value: < 2.2e-16
```

The fitted model is $\hat{y}_t = 74.68 + 0.001940t$. The P-value for the slope is very small, so that term is needed in the model. The linear model explains 37.89% of the variability in the water temperatures.

A time series plot of the residuals is shown below. It shows clear patterns so we should be able to improve on the simple linear model.

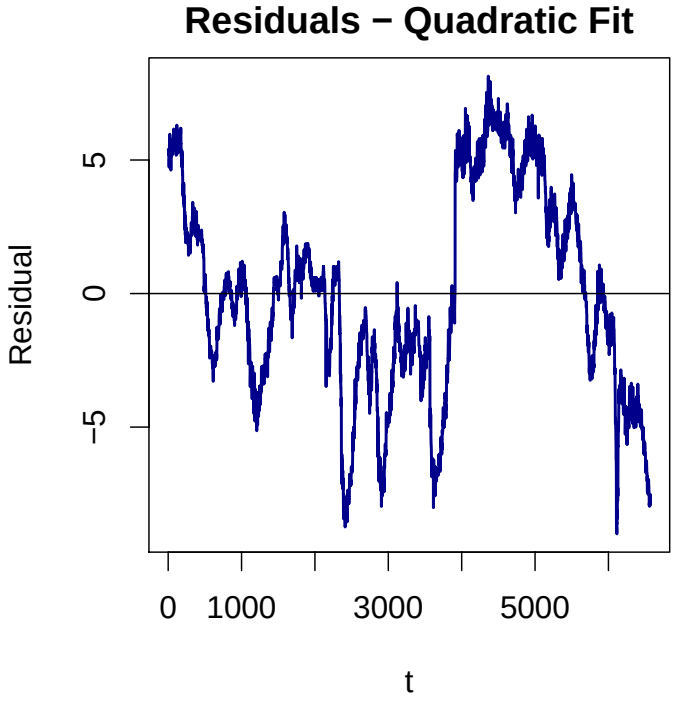


b. Here is some output for the quadratic trend model.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.081e+01  1.418e-01  569.72   <2e-16 ***
t           -3.661e-03  9.967e-05  -36.73   <2e-16 ***
I(t^2)       8.520e-07  1.468e-08   58.03   <2e-16 ***
---
Residual standard error: 3.832 on 6569 degrees of freedom
Multiple R-squared:  0.5894,Adjusted R-squared:  0.5893
F-statistic:  4715 on 2 and 6569 DF,  p-value: < 2.2e-16
```

The fitted model is $\hat{y}_t = 80.81 - 0.003661t + 0.000000852t^2$. The P-value for t^2 is very small, so that term is needed in the model. The quadratic model explains 58.94% of the variability in the water temperatures.

A time series plot of the residuals follows. The quadratic term was helpful, but the residual plot shows that there are still clear patterns in the residuals.



c. Here is some output for the cosine trend added to the quadratic model.

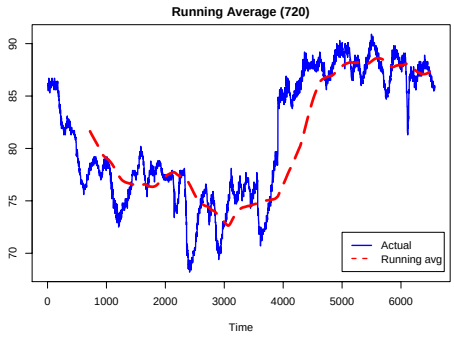
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.081e+01  1.418e-01 569.797   <2e-16 ***
t           -3.661e-03  9.966e-05 -36.733   <2e-16 ***
I(t^2)       8.521e-07  1.468e-08  58.041   <2e-16 ***
Xcos         1.267e-01  6.685e-02   1.895   0.0582 .
Xsin         3.462e-02  6.683e-02   0.518   0.6044
---
Residual standard error: 3.831 on 6567 degrees of freedom
Multiple R-squared:  0.5896,Adjusted R-squared:  0.5894
F-statistic:  2359 on 4 and 6567 DF,  p-value: < 2.2e-16
```

The fitted model is

$$\hat{y}_t = 80.81 - 0.003661t + 0.0000008521t^2 + 0.1267\cos(2\pi t/24) + 0.03462\sin(2\pi t/24)$$

Even though the P-value for $Xcos$ is somewhat small (0.0582), we are still explaining approximately 59% of the variability in water temperatures. The two additional terms in the cosine models are not helpful here.

12.38 Here are plots for the running averages of each order.



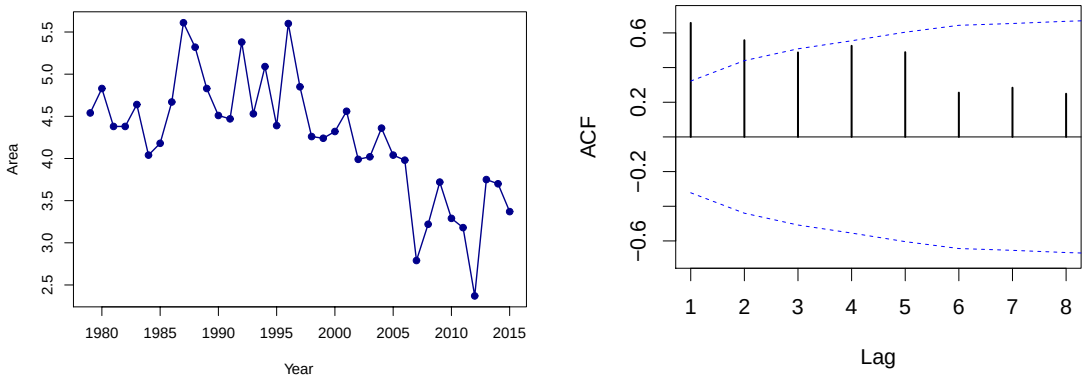
Notice that the running averages becomes flatter (smoother) as the number of hours (order) increases. There is also a delay in picking up changes in trends as the order increases. It is a lot harder to change an average of 720 values than it is to change an average of 24 values for one day. The one- and two-day models (24 and 48 hours) summarize the temperatures better than the other three models, which show too much delay in reacting to changes.

12.39 Notice that the output from these two commonly used statistical packages is different, both in format and the specific computed values. In fact, the moving average terms are very significant and have opposite signs! You must take care when interpreting output (R's model uses $+\theta_k \epsilon_{t-k}$ for moving average terms, while Minitab agrees with our book's notation for the model with $-\theta_k \epsilon_{t-k}$.

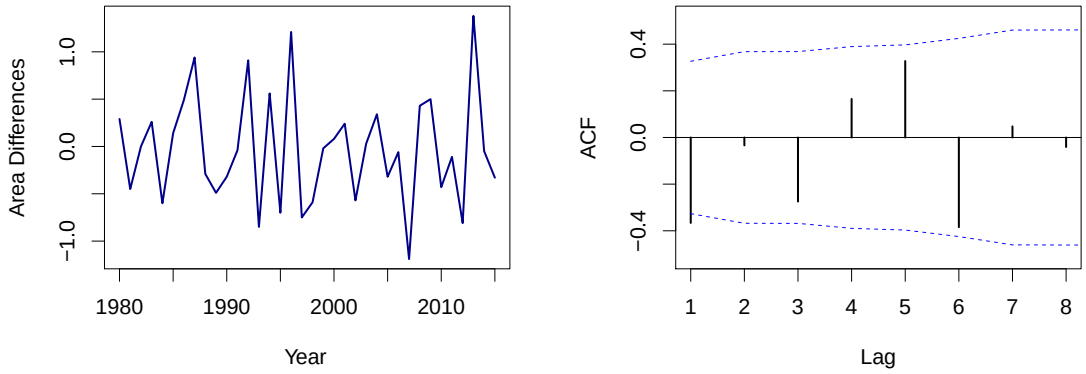
The Minitab output provides estimates, standard errors, test statistics, and P-values for the regular and seasonal moving average terms. We don't want to go too far with this output, because we have not checked any conditions, but both terms appear to be important. The estimates (and standard errors) for the regular and seasonal autoregressive terms are also provided in the R output. Minitab provides a sum of squares and mean square error, while R provides an estimate of the error variance (which is also its MSE) along with other measures of fit.

12.40 All of the coefficients are significant (bigger than $2SE$) for each of the models. The estimated error variance is smallest (0.04822) and the same for the ARIMA$(1,1,0) \times (0,1,1)$ and ARIMA$(1,1,0) \times (1,0,0)$ models and they have the same number of parameters (2). But the model with the seasonal difference uses 24 fewer data points (due to the seasonal differences) and has a smaller log likelihood (605.55 compared to 638.89), so we'll pick the ARIMA$(1,1,0) \times (1,0,0)_{24}$ as the best of these four models. In practice we should rely on more than just this output, for example, looking at ACF plots of the residuals.

12.41 As with the *Extent*, a time series plot of the Arctic sea ice *Area* shows a downward trend suggesting a difference is needed to help with stationarity. The plot of the ACF for the original series also shows a slow linear decay in the autocorrelations, providing further evidence for the need of a difference.



The plots below show the first differences as a time series and their ACF plot. Fluctautions around a constant mean look better for the differences and the ACF plot shows only a significant autocorrelation at the first lag. The first difference is useful to get to a relatively stationary series.



The autocorrelation at lag 1 in the differences of the *Area* series suggests using either one autoregressive or one moving average term in the model. Thus the initial candidates are ARIMA$(1,1,0)$ and ARIMA$(0,1,1)$. Here is some output for fitting both of those models. Note: The "drift" term is the constant in the model.

```
ARIMA(1,1,0) with drift
Coefficients:
          ar1    drift
      -0.3619  -0.0327
s.e.   0.1534   0.0670

sigma^2 estimated as 0.3127:  log likelihood=-29.2
```

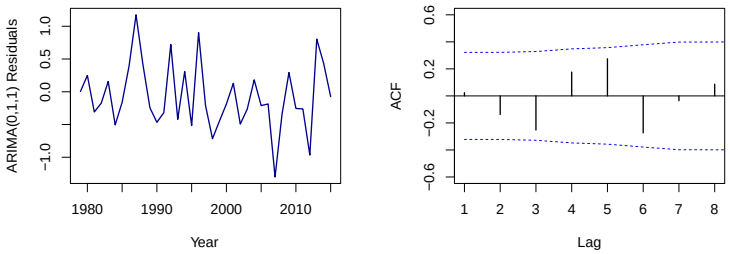
```
ARIMA(0,1,1) with drift
Coefficients:
          ma1    drift
      -0.6545  -0.0360
s.e.   0.1334   0.0308

sigma^2 estimated as 0.2695:  log likelihood=-26.73
```

In both models the constant (drift) term is not very significant (coefficient is less than 2 s.e.), so it can safely be dropped from the model. Here are the resulting outputs without a constant term.

```
ARIMA(1,1,0)
Coefficients:
          ar1
      -0.3578
s.e.   0.1537

sigma^2 estimated as 0.3058:  log likelihood=-29.32

ARIMA(0,1,1)
Coefficients:
          ma1
      -0.5965
s.e.   0.1307

sigma^2 estimated as 0.271:  log likelihood=-27.29
```

We see that the coefficient in both models is significant (more than twice its s.e. as a rough rule), but probably more so for the ARIMA$(0, 1, 1)$ model. That model also has the smaller estimated variance of the error term (0.271 compared to 0.3058). Although either model is probably adequate, these facts would give a small nod in favor of the ARIMA$(0, 1, 1)$ model with no constant term.

Checking the residuals from the ARIMA$(0, 1, 1)$ model we see no apparent pattern in the time series plot of the residuals and no significant autocorrelations in their ACF.



If we ask software for the forecasts for the next six years (after the series ends in 2015), we find the forecasts are all identical $\widehat{y}_t = 3.41$ for $t = 38, 39, \ldots$. Does it make sense that the forecasts don't change?

The fitted ARIMA$(0, 1, 1)$ prediction equation is

$$\Delta\widehat{y}_t = \widehat{y}_t - \widehat{y}_{t-1} = 0.5965\hat{\epsilon}_{t-1}$$

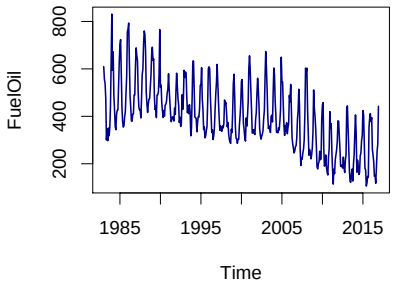
so

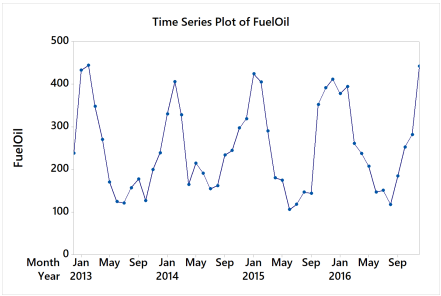
$$\widehat{y}_{38} = y_{37} - 0.5965\hat{\epsilon}_{37} = 3.37 - 0.5965(-0.0739) = 3.41$$

where 3.37 is the last value (for 2015) in the *Area* series and -0.0739 is the last residual from the fitted model.

For any forecast beyond $t = 38$, $\hat{\epsilon}_{t-1}$ is an unknown future error that is set to 0 when forecasting. Thus the rest of the future forecasts will stay fixed at 3.41.

12.42 The plot of the entire series (below left) shows a generally decreasing trend with a regular seasonal pattern. The plot showing just the last few years (below right) shows the seasonal pattern more clearly. As you might anticipate, fuel oil consumption is highest in the winter months and lowest in the summer.





From the time series plots alone we anticipate needing at least a linear t term and something seasonal, either a cosine trend or seasonal means. Here is some output for fitting both models.

Linear + Cosine trend:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 550.32912    6.49108   84.78   <2e-16
t            -0.74634    0.02751  -27.13   <2e-16
Xcos        104.00349    4.58061   22.70   <2e-16
Xsin         66.99320    4.58167   14.62   <2e-16
---
```

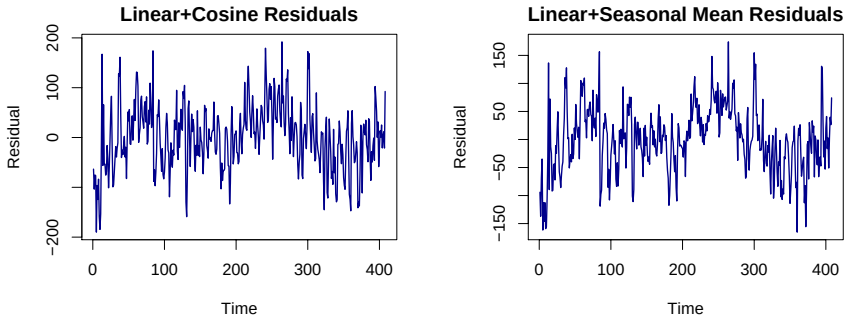
```
Residual standard error: 65.42 on 404 degrees of freedom
Multiple R-squared:  0.7852,Adjusted R-squared:  0.7836
F-statistic: 492.1 on 3 and 404 DF,  p-value: < 2.2e-16
```

Linear + Seasonal means:

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      704.54585   11.36019  62.019  < 2e-16
t                 -0.74439    0.02503 -29.735  < 2e-16
factor(Month)2   -10.91805   14.43854  -0.756   0.4500
factor(Month)3   -93.67472   14.43861  -6.488 2.61e-10
factor(Month)4  -189.80263   14.43872 -13.145  < 2e-16
factor(Month)5  -239.20939   14.43887 -16.567  < 2e-16
factor(Month)6  -241.83094   14.43906 -16.748  < 2e-16
factor(Month)7  -256.23843   14.43930 -17.746  < 2e-16
factor(Month)8  -235.14393   14.43958 -16.285  < 2e-16
factor(Month)9  -220.53445   14.43991 -15.273  < 2e-16
factor(Month)10 -189.11753   14.44028 -13.097  < 2e-16
factor(Month)11 -146.07326   14.44069 -10.115  < 2e-16
factor(Month)12  -32.84446   14.44115  -2.274   0.0235
---
Residual standard error: 59.53 on 395 degrees of freedom
Multiple R-squared:  0.8261,Adjusted R-squared:  0.8208
F-statistic: 156.3 on 12 and 395 DF,  p-value: < 2.2e-16
```

In both models we see strong evidence for including the linear term and the seasonal components. At this stage we might have a mild preference for the seasonal means model since it has a larger adjusted R^2 (82.08% compared to 78.36%) and smaller residual standard error (59.53 compared to 65.42).

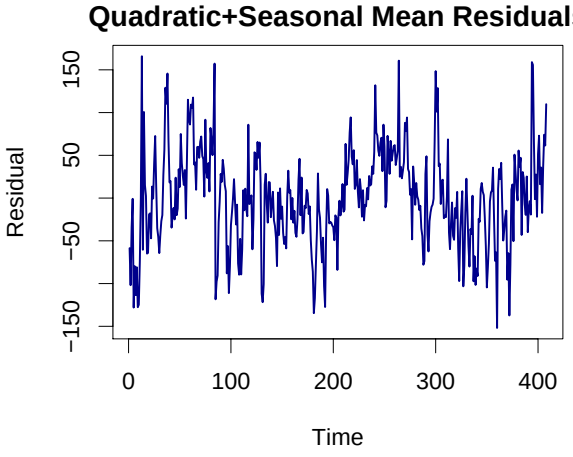
Here are plots of the residuals for both models. The flucuations about zero aren't too bad, but there might be a general increasing trend in the first 100 residuals, then somewhat decreasing between times 260 and 360. Perhaps adding the quadratic term will help with this?



Because the linear + seasonal means model looked better, we'll try adding a quadratic term to that model. Here is some output.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.685e+02  1.267e+01  52.760  < 2e-16
t               -2.164e-01  9.663e-02  -2.240   0.0257
I(t^2)          -1.291e-03  2.288e-04  -5.642 3.21e-08
factor(Month)2  -1.093e+01  1.391e+01  -0.786   0.4323
factor(Month)3  -9.370e+01  1.391e+01  -6.738 5.72e-11
factor(Month)4  -1.898e+02  1.391e+01 -13.651  < 2e-16
factor(Month)5  -2.392e+02  1.391e+01 -17.204  < 2e-16
factor(Month)6  -2.419e+02  1.391e+01 -17.393  < 2e-16
factor(Month)7  -2.563e+02  1.391e+01 -18.428  < 2e-16
factor(Month)8  -2.352e+02  1.391e+01 -16.911  < 2e-16
factor(Month)9  -2.206e+02  1.391e+01 -15.860  < 2e-16
factor(Month)10 -1.891e+02  1.391e+01 -13.600  < 2e-16
factor(Month)11 -1.461e+02  1.391e+01 -10.504  < 2e-16
factor(Month)12 -3.284e+01  1.391e+01  -2.361   0.0187
---
Residual standard error: 57.34 on 394 degrees of freedom
Multiple R-squared:  0.8391,Adjusted R-squared:  0.8338
F-statistic:   158 on 13 and 394 DF,  p-value: < 2.2e-16
```

We see that the test for the coefficient of the t^2 term has t-statistic of -5.643 and very small P-value, so that term is useful to add to the model. Note also that the adjusted R^2 (83.38%) and residual standard error (57.34) have also both improved with the inclusion of the quadratic term. Does the residual plot look better?

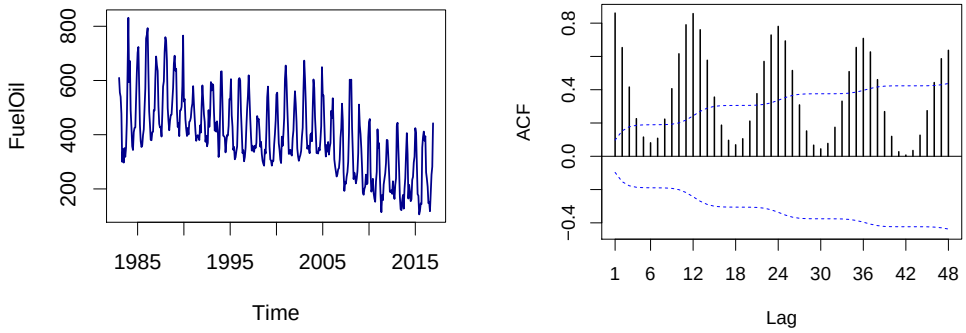


The residual plot does not look much different from the earlier two. There are still some periods with general increasing and decreasing trends. We may need the ARIMA techniques of the next exercise to help with this issue.

We'll use the quadratic trend + seasonal means model to forecast *FuelOil* consumption for each month in 2018. Output for these forecasts (with prediction intervals) follows.

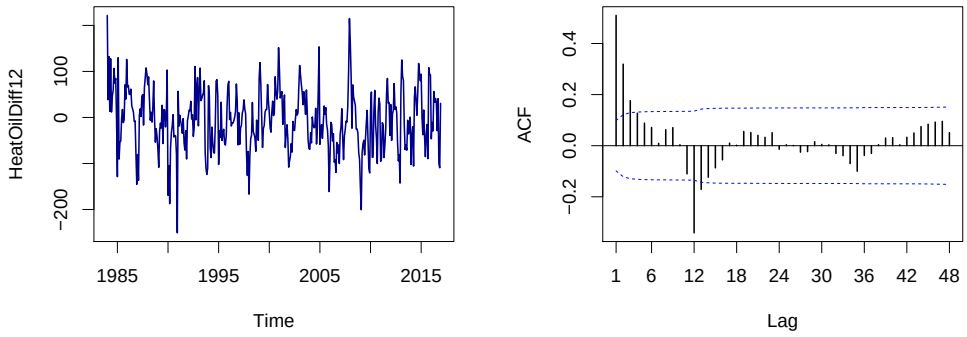
```
     fit     lwr     upr
1  364.0   248.6   479.5
2  351.8   236.3   467.3
3  267.8   152.3   383.3
4  170.4    54.8   285.9
5  119.7     4.1   235.3
6  115.8     0.2   231.4
7  100.1   -15.5   215.7
8  119.9     4.3   235.5
9  133.2    17.6   248.9
10 163.3    47.7   279.0
11 205.1    89.4   320.8
12 317.0   201.3   432.8
```

12.43 We start with a plot of the time series and its ACF.



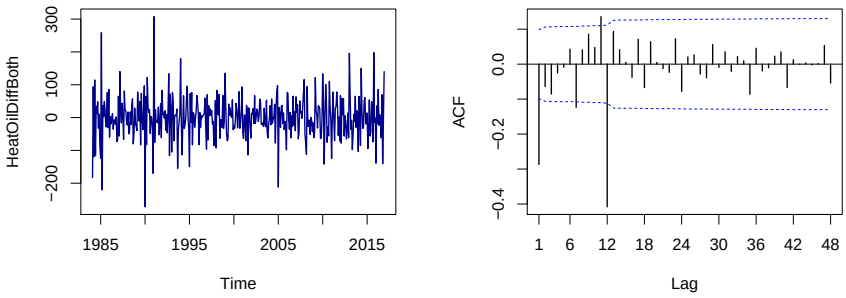
There is a strong seasonal pattern in the time series and slow decay at lags 12, 24, 36, and 48 in the ACF, suggesting the need for a seasonal difference. There is also an overall decreasing trend in the original series that might suggest the need for a regular difference. We'll try just the seasonal difference first.

Here are plots of the first seasonal differences (lag 12) and their ACF.



The decreasing trend is less pronounced and the slow seasonal decay is gone from the ACF, leaving only a single large seasonal spike at lag 12. The case for a regular difference is less clear. We see decaying positive spikes at the early lags, but it's a fairly rapid decay and we don't see a strong decreasing trend in the differenced series. We might need a regular difference, but might also be able to deal with those early spikes in the ACF with regular autoregressive or moving average terms.

Here are plots after doing a regular difference in addition to the seasonal difference.



Stationarity looks fine after taking both differences; the time series plot shows no consistent increasing or decreasing patterns and there are only large spikes at lags 1 and 12 in the ACF (and possibly a couple of random spikes at lags 7 and 11 that we will discount since they aren't at early or seasonal lags).

Summary so far: We definitely want a seasonal difference for stationarity and either a seasonal autoregressive or seasonal moving average term to address the large spike at lag 12 that is present in both ACF plots of differences. For the regular model we might try no difference and one or more AR or MA terms, or use a regular difference and a single AR or MA term for the spike at lag 1.

Here is some information for several potential models. We include constant terms only in the first four models that don't have a regular difference.

Model	Insignificant terms?	Error variance	Lags with large residual ACF?
$(1,0,0) \times (1,1,0)$	constant	2678	24
$(1,0,0) \times (0,1,1)$	none	2114	9, 12
$(0,0,1) \times (1,1,0)$	constant	2993	1, 2, 24
$(0,0,1) \times (0,1,1)$	none	2687	2 and lots others
$(1,1,0) \times (1,1,0)$	none	3127	2, 7, 12, 24
$(1,1,0) \times (0,1,1)$	none	2343	2, 4, 7, 12, 16
$(0,1,1) \times (1,1,0)$	none	2990	1, 3, 4, 5, 7, 24
$(0,1,1) \times (0,1,1)$	none	2286	3, 2, 5, 7, 12, 15

The residual ACFs look the best for the first two models without a regular difference; and the second of these has the smallest error variance. Those look the most promising if we could deal with the remaining seasonal spikes in the residual ACF. Perhaps another seasonal AR or MA term?

Model	Insignificant terms?	Error variance	Lags with large residual ACF?
$(1,0,0) \times (1,1,1)$	SMA1	1989	9
$(1,0,0) \times (0,1,2)$	none	1986	9

These appear to work almost equally well, both showing only a barely significant spike in the residual ACF at lag 9. We have no reason to suspect something unusual happening nine months apart (at least not for fuel oil use) so we can discount that one autocorrelation. All coefficients are very significant in the second model and it has a slightly smaller error variance, so our choice will be ARIMA$(1,0,0) \times (0,1,2)$ with a constant term. Note that it is often very difficult to find a model that is optimal on all criteria, so it's OK to accept a different model as workable in this situation.

Here is some output after requesting the forecasts for the next 12 months of the series (2018).

```
Forecasts from period 408
                        95% Limits
Period   Forecast   Lower   Upper
409         441.2   354.0   528.5
410         422.8   319.4   526.3
411         312.5   203.2   421.7
412         221.3   109.7   332.9
413         175.3    62.8   287.8
414         157.7    44.8   270.6
415         145.2    32.1   258.2
416         153.9    40.8   267.0
417         175.6    62.4   288.7
418         215.6   102.5   328.8
419         255.1   141.9   368.2
420         367.2   254.4   480.6
```