

1. Suppose that $\{Y_t\}$ is a stationary process with constant mean $\mu_t = E(Y_t) = \mu$ and autocorrelation function ρ_k . Define

$$\bar{Y} = \frac{1}{n} \sum_{t=1}^n Y_t$$

to be the sample mean of Y_1, Y_2, \dots, Y_n .

(a) Show that \bar{Y} is an unbiased estimator of μ ; that is, show that $E(\bar{Y}) = \mu$. Note that this result holds regardless of the values of ρ_k .

$$\begin{aligned} E(\bar{Y}) &= E\left(\frac{1}{n} \sum_{t=1}^n Y_t\right) \\ &= \frac{1}{n} \sum_{t=1}^n E(Y_t) \\ &= \frac{1}{n} \sum_{t=1}^n \mu \\ &= \mu \end{aligned}$$

(b) Prove that

$$\text{var}(\bar{Y}) = \frac{\gamma_0}{n} \left[1 + 2 \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) \rho_k \right],$$

where $\gamma_0 = \text{var}(Y_t)$.

$$\begin{aligned} \text{Var}(\bar{Y}) &= \text{Var}\left(\frac{1}{n} \sum_{t=1}^n Y_t\right) \\ &= \frac{1}{n^2} \text{Var}\left(\sum_{t=1}^n Y_t\right) \\ &= \frac{1}{n^2} \left[\sum_{t=1}^n \text{Var}(Y_t) + \sum_{t=1}^n \sum_{s=1}^{t-1} 2\text{Cov}(Y_t, Y_s) \right] \end{aligned}$$

You can see that s , the subscript for the second variable in the covariance, runs from 1 to $t-1$. Therefore, s runs through all possible lags behind t . So, we can rewrite the summation involving s using lag k notation as follows:

$$\begin{aligned} \text{Var}(\bar{Y}) &= \frac{1}{n^2} \left[\sum_{t=1}^n \text{Var}(Y_t) + 2 \sum_{t=1}^n \sum_{k=1}^{t-1} \text{Cov}(Y_t, Y_{t-k}) \right] \\ &= \frac{1}{n^2} \left[\sum_{t=1}^n \text{Var}(Y_t) + 2 \sum_{t=1}^n \sum_{k=1}^{t-1} \text{Corr}(Y_t, Y_{t-k}) \sqrt{\text{Var}(Y_t) \text{Var}(Y_{t-k})} \right] \end{aligned}$$

We can swap the order of the summations as follows:

$$\begin{aligned}
 \text{Var}(\bar{Y}) &= \frac{1}{n^2} \left[\sum_{t=1}^n \text{Var}(Y_t) + 2 \sum_{k=1}^{n-1} \sum_{t=k+1}^n \text{Corr}(Y_t, Y_{t-k}) \sqrt{\text{Var}(Y_t) \text{Var}(Y_{t-k})} \right] \\
 &= \frac{1}{n^2} \left[\sum_{t=1}^n \gamma_0 + 2 \sum_{k=1}^{n-1} \sum_{t=k+1}^n \rho_k \sqrt{\gamma_0 \gamma_0} \right] \\
 &= \frac{1}{n^2} \left[n\gamma_0 + 2\gamma_0 \sum_{k=1}^{n-1} \sum_{t=k+1}^n \rho_k \right] \\
 &= \frac{\gamma_0}{n} \left[1 + \frac{2 \sum_{k=1}^{n-1} \sum_{t=k+1}^n \rho_k}{n} \right] \\
 &= \frac{\gamma_0}{n} \left[1 + \frac{2 \sum_{k=1}^{n-1} (n-k) \rho_k}{n} \right] \\
 &= \frac{\gamma_0}{n} \left[1 + 2 \sum_{k=1}^{n-1} \left(\frac{n-k}{n} \right) \rho_k \right] \\
 &= \frac{\gamma_0}{n} \left[1 + 2 \sum_{k=1}^{n-1} \left(1 - \frac{k}{n} \right) \rho_k \right]
 \end{aligned}$$

(c) If $\{Y_t\}$ is a white noise process, what does $\text{var}(\bar{Y})$ reduce to?

In a white noise process, the observations are mutually independent. Therefore, $\rho_1 = \rho_2 = \dots = 0$. and

$$\text{Var}(\bar{Y}) = \frac{\gamma_0}{n}.$$

(d) Suppose that $\{Y_t\}$ is a **MA(1) process**; i.e.,

$$Y_t = e_t - \theta e_{t-1},$$

where θ is a fixed parameter and $\{e_t\}$ is a zero mean white noise process with $\text{var}(e_t) = \sigma_e^2$. Use the result in part (b) to find an expression for $\text{var}(\bar{Y})$.

By part (b), we know that

$$\text{Var}(\bar{Y}) = \frac{\gamma_0}{n} \left[1 + 2 \sum_{k=1}^{n-1} \left(1 - \frac{k}{n} \right) \rho_k \right].$$

So, we must find the variance and autocorrelation functions for an **MA(1) process**.

$$\begin{aligned}
 \text{Cov}(Y_t, Y_{t-k}) &= \text{Cov}(e_t - \theta e_{t-1}, e_{t-k} - \theta e_{t-k-1}) \\
 &= \text{Cov}(e_t, e_{t-k}) - \theta \text{Cov}(e_{t-1}, e_{t-k}) - \theta \text{Cov}(e_t, e_{t-k-1}) + \theta^2 \text{Cov}(e_{t-1}, e_{t-k-1}) \\
 &= \left\{ \begin{array}{ll} \text{Var}(e_t) + \theta^2 \text{Var}(e_{t-1}), & k = 0 \\ -\theta \text{Var}(e_{t-1}), & k = 1 \\ 0, & k \neq 0, 1 \end{array} \right\} \\
 &= \left\{ \begin{array}{ll} \sigma_e^2(1 + \theta^2), & k = 0 \\ -\theta \sigma_e^2, & k = 1 \\ 0, & k \neq 0, 1 \end{array} \right\}
 \end{aligned}$$

$$\begin{aligned}
 \text{Corr}(Y_t, Y_{t-k}) &= \frac{\text{Cov}(Y_t, Y_{t-k})}{\sqrt{\text{Var}(Y_t)\text{Var}(Y_{t-k})}} \\
 &= \frac{\text{Cov}(Y_t, Y_{t-k})}{\sqrt{\sigma_e^2(1 + \theta^2)\sigma_e^2(1 + \theta^2)}} \\
 &= \frac{\text{Cov}(Y_t, Y_{t-k})}{\sigma_e^2(1 + \theta^2)} \\
 &= \left\{ \begin{array}{ll} 1, & k = 0 \\ -\frac{\theta \sigma_e^2}{\sigma_e^2(1 + \theta^2)}, & k = 1 \\ 0, & k \neq 0, 1 \end{array} \right\} \\
 &= \left\{ \begin{array}{ll} 1, & k = 0 \\ -\frac{\theta}{1 + \theta^2}, & k = 1 \\ 0, & k \neq 0, 1 \end{array} \right\}
 \end{aligned}$$

It is important to note that the only non-zero correlations are at lags $k = 0, 1$. Therefore, we can reduce the result from part (b) to

$$\begin{aligned}
 \text{Var}(\bar{Y}) &= \frac{\gamma_0}{n} \left[1 + 2 \left(1 - \frac{1}{n} \right) \left(-\frac{\theta}{1 + \theta^2} \right) \right] \\
 &= \frac{\sigma_e^2(1 + \theta^2)}{n} \left[1 + 2 \left(1 - \frac{1}{n} \right) \left(-\frac{\theta}{1 + \theta^2} \right) \right] \\
 &= \frac{\sigma_e^2(1 + \theta^2)}{n} \left[1 - \frac{2\theta}{1 + \theta^2} + \frac{2\theta}{n(1 + \theta^2)} \right] \\
 &= \frac{\sigma_e^2}{n} \left[(1 + \theta^2) - 2\theta + \frac{2\theta}{n} \right]
 \end{aligned}$$

2. Consider the process $Y_t = \mu + X_t$, where $X_t = e_t - \theta e_{t-1}$ and $e_t \sim \text{iid } \mathcal{N}(0, 1)$. Let $n = 100$.

(a) Show that $E(X_t) = 0$ and $\text{var}(X_t) = \gamma_0 = 1 + \theta^2$. Because linear combinations of normal random variables are normally distributed, it follows immediately that $X_t \sim \mathcal{N}(0, 1 + \theta^2)$ and $Y_t \sim \mathcal{N}(\mu, 1 + \theta^2)$.

$$\begin{aligned} E(X_t) &= E(e_t - \theta e_{t-1}) \\ &= E(e_t) - \theta E(e_{t-1}) \\ &= 0 - \theta(0) \\ &= 0 \end{aligned}$$

$$\begin{aligned} \text{Var}(X_t) &= \text{Var}(e_t - \theta e_{t-1}) \\ &= \text{Var}(e_t) + \theta^2 \text{Var}(e_{t-1}) - \theta \text{Cov}(e_t, e_{t-1}) \\ &= 1 + \theta^2(1) - \theta(0) \\ &= 1 + \theta^2 \end{aligned}$$

(b) Find the sampling distribution of \bar{Y} .

Since we know that $Y_t = \mu + X_t$ where $X_t \sim N(0, 1 + \theta^2)$. So, by pp. 50 in the notes, we have

$$\bar{Y} \sim N\left(\mu, \frac{1 + \theta^2}{n} \left[1 + 2 \sum_{k=1}^{n-1} \left\{1 - \frac{k}{n}\right\} \rho_k\right]\right).$$

By Problem 1(d), we can rewrite the variance for \bar{Y} and obtain

$$\bar{Y} \sim N\left(\mu, \frac{1}{n} \left[(1 + \theta^2) - 2\theta + \frac{2\theta}{n}\right]\right).$$

(c) Suppose that $\theta = -0.5$. Use the R code

```
> Y.t = 10 + arima.sim(list(order = c(0,0,1), ma = 0.5), n = 100)
```

to generate a realization of this process when $\mu = 10$. Compute a 99 percent confidence interval for μ (which in this problem is known to be 10). Does your interval include 10? You can use the `mean(Y.t)` command to compute the sample mean of Y_1, Y_2, \dots, Y_{100} .

Since we know that \bar{Y} has a $N\left(\mu, \frac{1}{n} \left[(1 + \theta^2) - 2\theta + \frac{2\theta}{n}\right]\right)$ distribution, we can derive a 99 percent confidence interval for μ as follows (the sampled values used in this confidence are from my simulation, yours should be slightly different):

$$\begin{aligned}
\bar{Y} \pm z_{\frac{\alpha}{2}} \sqrt{\sigma_Y^2} &= \bar{Y} \pm z_{.005} \sqrt{\left(\frac{1}{n^2}\right) [n + (2 - 2n)\theta + n\theta^2]} \\
&= \bar{Y} \pm z_{1-\frac{\alpha}{2}} \sqrt{\left(\frac{1}{100^2}\right) [100 + (2 - 2\{100\})(-.5) + 100(-.5)^2]} \\
&= 9.928 \pm 2.576(.150) \\
&= (9.542, 10.313)
\end{aligned}$$

We are 99% confident that the true mean of Y_t is between 9.542 and 10.313. This interval includes 10.

(d) Repeat part (c) with $\theta = 0.5$. You will use the same command to simulate a new realization, except use `ma = -0.5`. R uses the convention of negating the parameter θ , as we will see in Chapter 4.

We can replicate the procedure in part (c) as follows:

$$\begin{aligned}
\bar{Y} \pm z_{\frac{\alpha}{2}} \sqrt{\sigma_Y^2} &= \bar{Y} \pm z_{1-\frac{\alpha}{2}} \sqrt{\left(\frac{1}{100^2}\right) [100 + (2 - 2\{100\})(.5) + 100(.5)^2]} \\
&= 10.015 \pm 2.576(.051) \\
&= (9.883, 10.146)
\end{aligned}$$

We are 99% confident that the true mean of Y_t is between 9.883 and 10.146. This interval includes 10.

(e) Compare the confidence intervals in parts (c) and (d) in terms of interval length. Explain why one interval should be shorter.

The length of interval in part (d) is much smaller than the one in part (c). This is because a positive θ causes the population variance to be larger, thereby increasing the size of the confidence interval. We know that the autocorrelation function for an **MA(1) process** at lag $k = 1$ is $\rho_1 = \frac{-\theta}{1+\theta^2}$. So, we can see that positive θ 's cause negative ρ_1 's and vice-versa. Therefore, we can say that positive ρ_1 's give less information about μ and negative ρ_1 's give more information about μ .

3. The TSA library contains the data set `co2`, which lists monthly carbon dioxide levels in northern Canada from 1/1994 to 12/2004. In HW1, you examined these data and fit

a straight line regression model to them. You should have concluded that this was not a very good model for these data. In this problem, we will use R to fit the model

$$\text{CO2}_t = \beta_0 + \beta_1 t + \beta_2 \cos(2\pi ft) + \beta_3 \sin(2\pi ft) + X_t,$$

where $E(X_t) = 0$. This deterministic part

$$\mu_t = \beta_0 + \beta_1 t + \beta_2 \cos(2\pi ft) + \beta_3 \sin(2\pi ft)$$

contains both linear and trigonometric trend components. Note that there are 12 observations per year, so we take the frequency $f = 1$.

(a) To fit the model, we can use the R commands:

```
> har. <- harmonic(co2,1)
> fit <- lm(co2~har. + time(co2))
> summary(fit)
```

What are the least squares estimates? Write out an equation for the fitted model. Is all of the R output relevant? Explain.

The R output is as follows:

Residuals:

Min	1Q	Median	3Q	Max
-5.1804	-1.7916	-0.1045	1.8986	5.1809

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3.332e+03	1.280e+02	-26.03	<2e-16 ***
har.cos(2*pi*t)	3.851e+00	2.868e-01	13.43	<2e-16 ***
har.sin(2*pi*t)	5.596e+00	2.874e-01	19.47	<2e-16 ***
time(co2)	1.851e+00	6.401e-02	28.91	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.329 on 128 degrees of freedom

Multiple R-squared: 0.9109, Adjusted R-squared: 0.9088

F-statistic: 436.3 on 3 and 128 DF, p-value: < 2.2e-16

The least squares estimates for the linear coefficients are $\hat{\beta}_0 = -3332$, $\hat{\beta}_1 = 3.851$, $\hat{\beta}_2 = 5.596$, and $\hat{\beta}_3 = 1.851$. The fitted regression model is

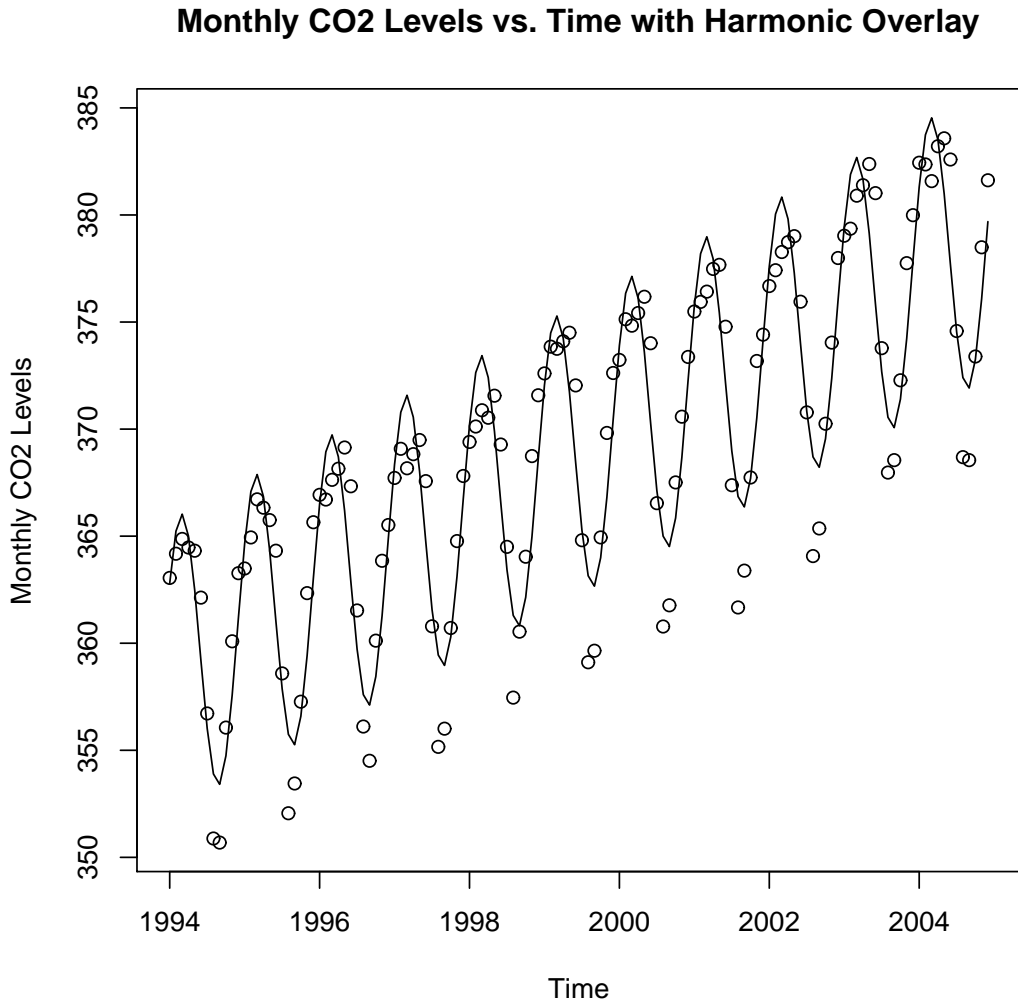
$$\mu_t = -3332 + 3.851t + 5.596 \cos(2\pi ft) + 1.851 \sin(2\pi ft).$$

For this problem, we only assumed that $E(X_t) = 0$. We did not make the traditional iid $N(0, \sigma^2)$ assumption on the error process $\{X_t\}$. Therefore, the only relevant pieces of this R output are the least squares estimates.

(b) Construct a graph which plots the points along with the fitted regression model superimposed. You can use the R commands:

```
> plot(ts(fitted(fit),freq=12,start=c(1994,1)),ylab="Monthly CO2 levels",
      type='l',ylim=range(c(fitted(fit),co2)))
> points(co2)
```

Is this model appear to fit the data better than the straight line model? How would you rate the fit overall?



This model fits the data substantially better than the straight line model. The only improvement I would like to see is for the troughs to extend a little further down. This model doesn't seem to reach the lower points. This could be due to the fact that there are many more points near the peaks, so they are more heavily weighted when considering the parameters in the model.

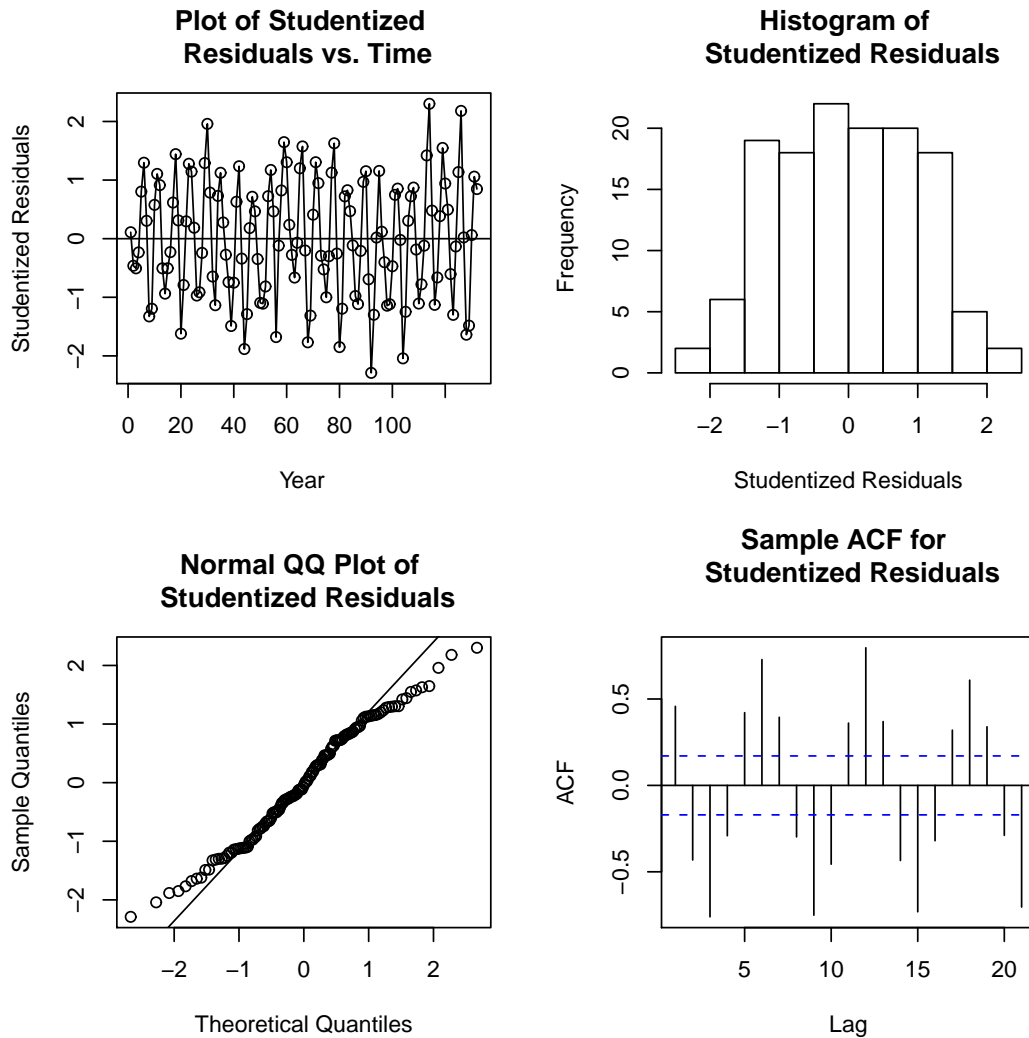
(c) Use the remaining code to perform the model diagnostics we discussed in class (Section 3.5 in the notes).

```

> plot(rstudent(fit),ylab="Std.residuals",xlab="Year",type="o")
> abline(h=0)
> hist(rstudent(fit),main="Hist. of std.residuals",xlab="Std.residuals")
> qqnorm(rstudent(fit),main="QQ plot of std.residuals")
> shapiro.test(rstudent(fit))
> runs(rstudent(fit))
> acf(rstudent(fit),main="Sample ACF for std.residuals")

```

Interpret everything and give an overall assessment of the model we have fit in this problem. Do the residuals look to resemble a stationary white noise process? Or, is there still noticeable structure left in them?



Shapiro-Wilk normality test

```

data: rstudent(fit)
W = 0.9831, p-value = 0.1003

```

Runs test

```
$pvalue
[1] 0.000152

$observed.runs
[1] 45

$expected.runs
[1] 66.98485

$n1
[1] 67

$n2
[1] 65

$k
[1] 0
```

As for the residual plot, there is significant fluctuation between negative and positive correlation between the observations. You can see this because the points are usually alternating between positive and negative values. This is much more evident if you look at the sample ACF. This is a great indicator that the cosine trend model doesn't account for the autocorrelation.

As for the histogram and the QQ plot, they don't tell us much. The residuals don't appear to be normally distributed. However, we never made this assumption in the first place. So, these plots do not really seem very helpful. The Shapiro-Wilk Test for Normality tells us that, even though the plots are "iffy", there is only mild evidence against normality of the error terms $\{X_t\}$. The runs test tells us that there is significant evidence to say that the error terms are correlated, which we already knew by looking at the residual plot and sample ACF.

So, the important information to be taken from this output is that the model does not account for all of the systematic variation in the data, normality is questionable, and independence is likely violated. For future reference, it is interesting to note that cosine trend models do not generally extract correlation from data. This task is handled much better by **ARIMA**-type models, which we will study extensively in this course.

4. Consider the ventilation data in Example 1.10 (notes, pp 11). The data are located on the course web site under the name `ventilation`. Cut and paste the data into Notepad and save the file as a `.txt` file. Then, read the data into R using the commands I give on the web site (suitably modified to the location of the `.txt` file on your computer).

(a) Fit a straight-line regression model to the data for detrending purposes.

The code used to fit the model, as well as the output obtained, is as follows:

```
> ventilation <- ts(ventilation)
> fit <- lm(time(ventilation) ~ ventilation)
> summary(fit)
```

Call:

```
lm(formula = time(ventilation) ~ ventilation)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-63.7988	-9.5768	-0.6045	9.1005	35.0251

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-46.2828	2.8183	-16.42	<2e-16 ***
ventilation	3.3663	0.0615	54.73	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.92 on 193 degrees of freedom

Multiple R-squared: 0.9395, Adjusted R-squared: 0.9392

F-statistic: 2996 on 1 and 193 DF, p-value: < 2.2e-16

(b) Do a thorough residual analysis of the detrended (residual) process using the techniques we discussed in class (Section 3.5 in the notes). Do the residuals from your straight-line model fit resemble a zero mean white noise process?

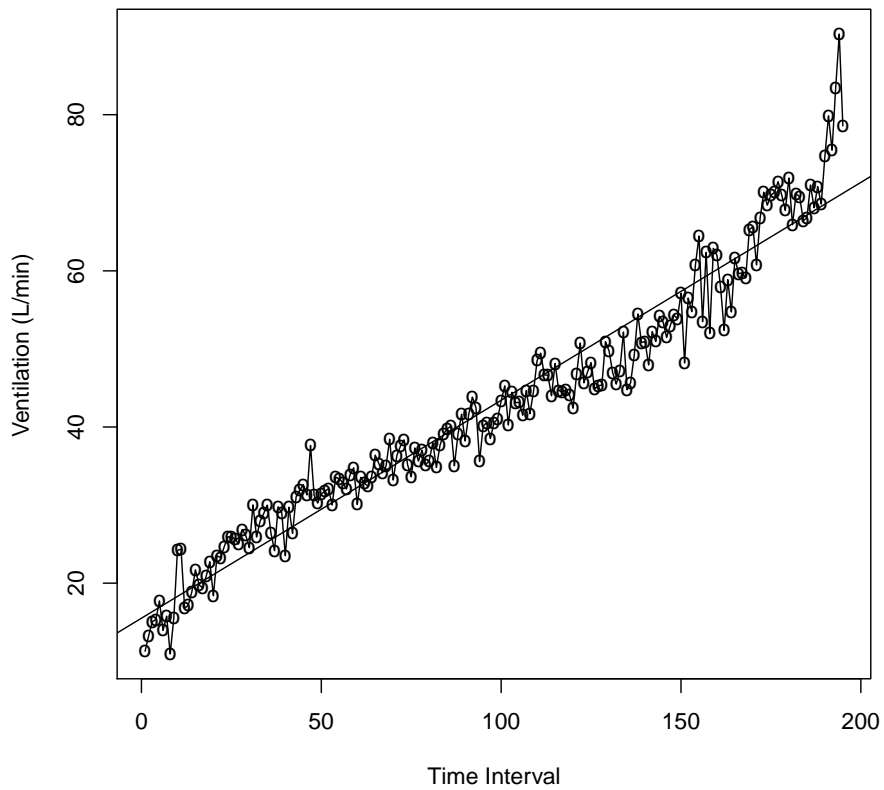
The code used to create the plots, as well as the output obtained, is as follows:

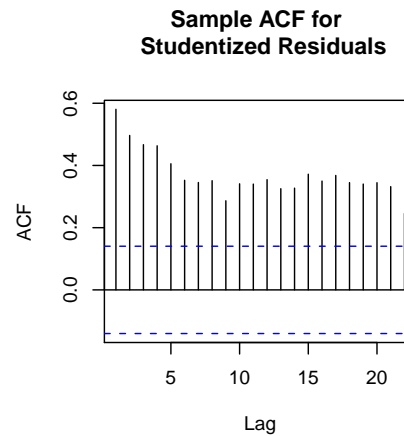
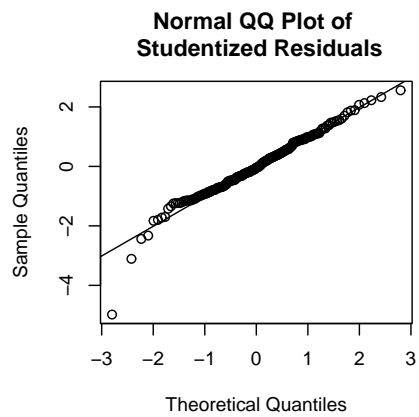
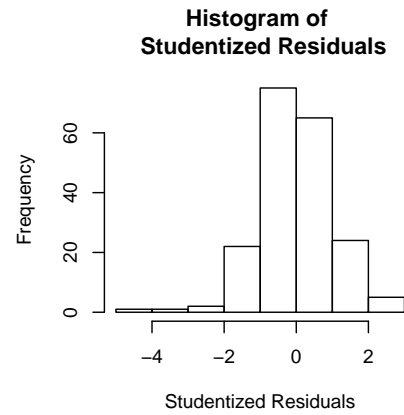
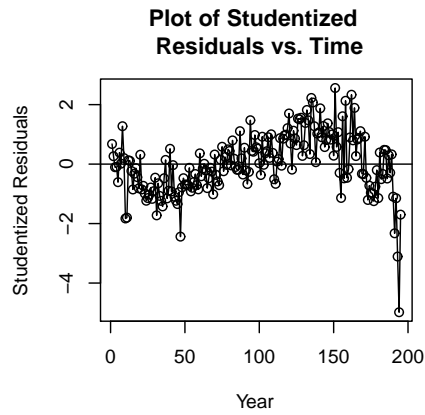
```
> plot.ts(ventilation, main = "Plot of Ventilation vs. Time
(15 second intervals)", xlab = "Time Interval", ylab =
"Ventilation (L/min)")
> points(time(ventilation), ventilation, pch = "o")
> abline(fit)

> par(mfrow = c(2,2))
> plot(rstudent(fit), ylab = "Studentized Residuals", xlab = "Year",
type = "o", main = "Plot of Studentized \n Residuals vs. Time")
> abline(h=0)
```

```
> hist(rstudent(fit), main = "Histogram of \n Studentized Residuals",  
xlab = "Studentized Residuals")  
> qqnorm(rstudent(fit),main="Normal QQ Plot of \n Studentized Residuals")  
> qqline(rstudent(fit))  
> acf(rstudent(fit),main="Sample ACF for \n Studentized Residuals")
```

Plot of Ventilation vs. Time (15 second intervals)





```
> shapiro.test(rstudent(fit))
```

Shapiro-Wilk normality test

```
data: rstudent(fit)  
W = 0.9453, p-value = 8.977e-07
```

```
> runs(rstudent(fit))

$pvalue
[1] 1.83e-07

$observed.runs
[1] 62

$expected.runs
[1] 98.37436

$n1
[1] 101

$n2
[1] 94

$k
[1] 0
```

We can see that a simple linear model might not be appropriate for these data. There is definite S-shaped curvature to the data, similar to what you would see in a cubic polynomial. This deviation from linearity is extremely evident in the residual plot. There is definite curvature and the zero mean and constant variance assumptions seem to be violated. The normality assumption seems mildly reasonable from examining the QQ plot. However, the Shapiro-Wilk Test for Normality finds significant evidence against normality at the $\alpha = .05$ level. But, this could be caused by the observation with a studentized residual of -5. Also, the ACF and runs test give reasonable evidence that the uncorrelated errors assumptions is also violated. Therefore, we can reasonably conclude that every assumption for the simple linear model is violated.

5. Consider the data set

5 7 9 4 5 6

(in this order).

(a) Compute by hand r_1 , the lag one autocorrelation coefficient. Show all calculations.

Before we do anything, we need to find the sample mean, \bar{Y} .

$$\begin{aligned}
 \bar{Y} &= \frac{1}{6} \sum_{t=1}^6 Y_t \\
 &= \frac{1}{6}(5 + 7 + 9 + 4 + 5 + 6) \\
 &= \frac{36}{6} \\
 &= 6
 \end{aligned}$$

Now, we can find the lag 1 sample autocorrelation, r_1 .

$$\begin{aligned}
 r_1 &= \frac{\sum_{t=2}^6 (Y_t - \bar{Y})(Y_{t-1} - \bar{Y})}{\sum_{t=1}^6 (Y_t - \bar{Y})^2} \\
 &= \frac{(7-6)(5-6) + (9-6)(7-6) + (4-6)(9-6) + (5-6)(4-6) + (6-6)(5-6)}{(5-6)^2 + (7-6)^2 + (9-6)^2 + (4-6)^2 + (5-6)^2 + (6-6)^2} \\
 &= \frac{(1)(-1) + (3)(1) + (-2)(3) + (-1)(-2) + (0)(-1)}{(-1)^2 + (1)^2 + (3)^2 + (-2)^2 + (-1)^2 + (0)^2} \\
 &= \frac{-1 + 3 - 6 + 2 + 0}{1 + 1 + 9 + 4 + 1 + 0} \\
 &= \frac{-2}{16} \\
 &= -\frac{1}{8}
 \end{aligned}$$

(b) Compute by hand r_2 , the lag two autocorrelation coefficient. Show all calculations.

$$\begin{aligned}
 r_2 &= \frac{\sum_{t=3}^6 (Y_t - \bar{Y})(Y_{t-2} - \bar{Y})}{\sum_{t=1}^6 (Y_t - \bar{Y})^2} \\
 &= \frac{(9-6)(5-6) + (4-6)(7-6) + (5-6)(9-6) + (6-6)(4-6)}{(5-6)^2 + (7-6)^2 + (9-6)^2 + (4-6)^2 + (5-6)^2 + (6-6)^2} \\
 &= \frac{(3)(-1) + (-2)(1) + (-1)(3) + (0)(-2)}{(-1)^2 + (1)^2 + (3)^2 + (-2)^2 + (-1)^2 + (0)^2} \\
 &= \frac{-3 - 2 - 3 + 0}{1 + 1 + 9 + 4 + 1 + 0} \\
 &= \frac{-8}{16} \\
 &= -\frac{1}{2}
 \end{aligned}$$

6. The TSA library contains the data set `prescrip`, which lists monthly prescription costs for the months August 1986 to March 1992. These data are from the State of New

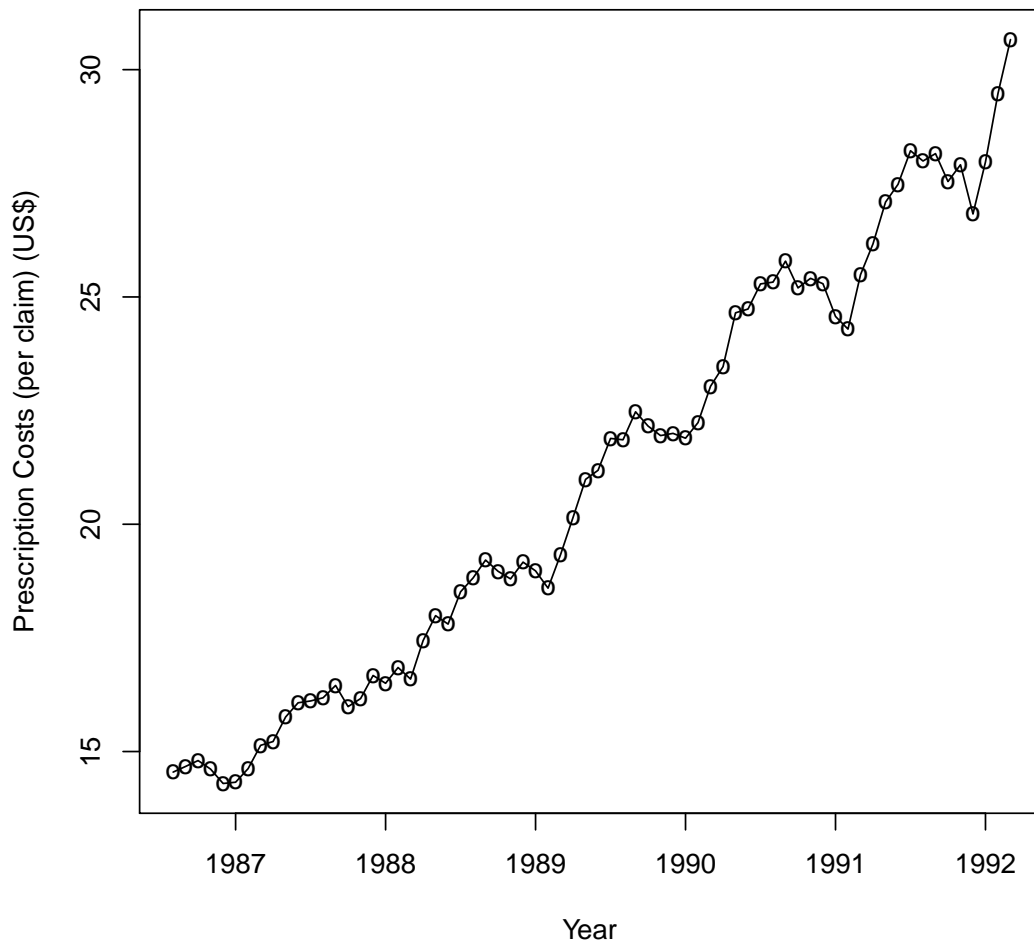
Jersey Prescription Drug Program and are the cost per prescription claim during this time period.

(a) Construct a time series plot for the data. Describe the appearance of the series.

The code to generate this plot is as follows:

```
> data(prescrip)
> plot(prescrip, main = "Plot of New Jersey Prescription Costs vs.
Year", xlab = "Year", ylab = "Prescription Costs (per claim) (US$)")
```

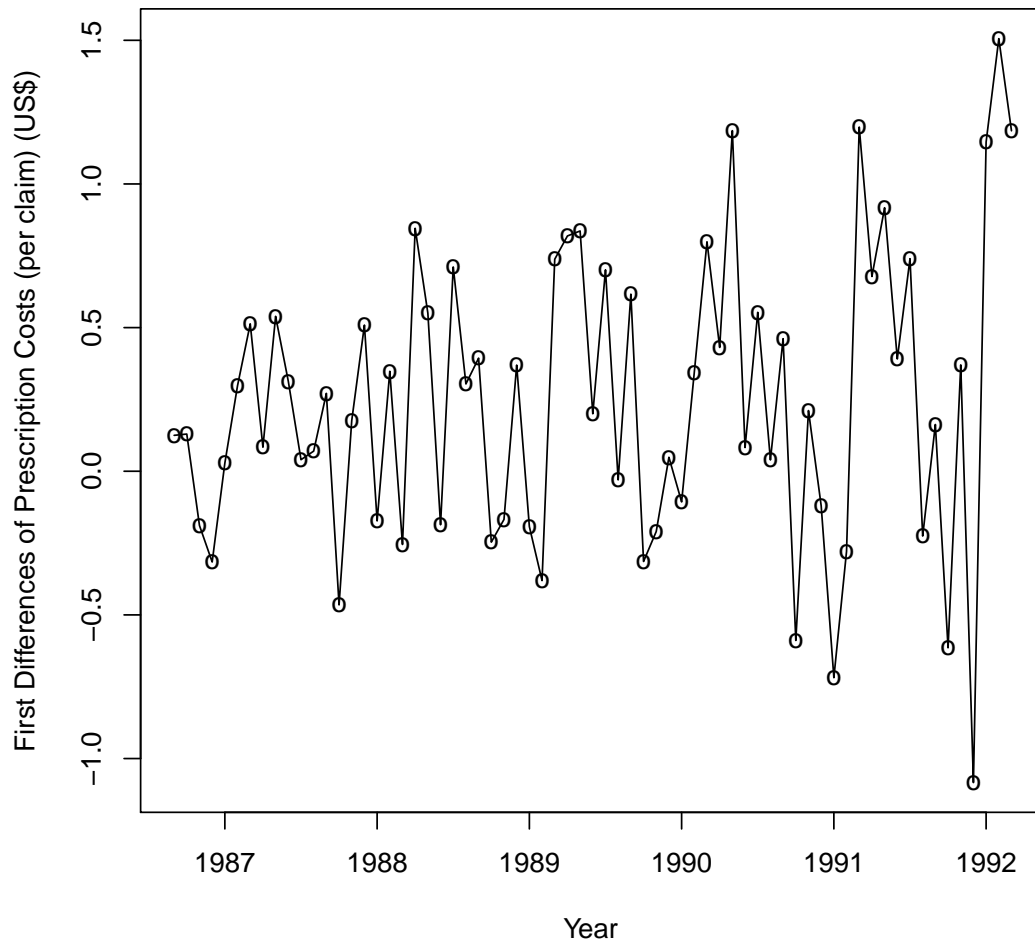
Plot of New Jersey Prescription Costs vs. Year



This series has a definite increasing trend. Also, there seems to be a seasonal effect with a wavelength of about 2 years. This series is definitely not stationary.

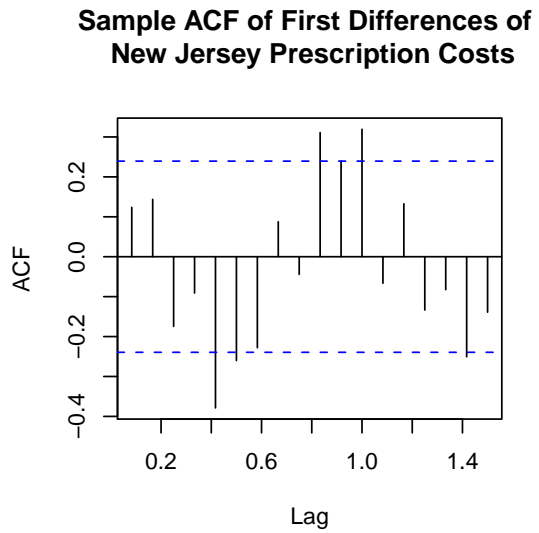
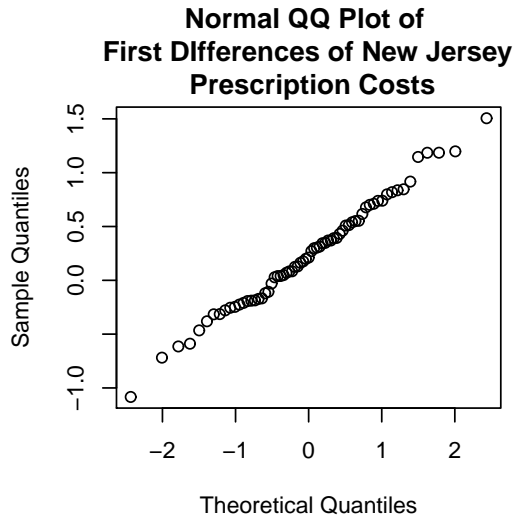
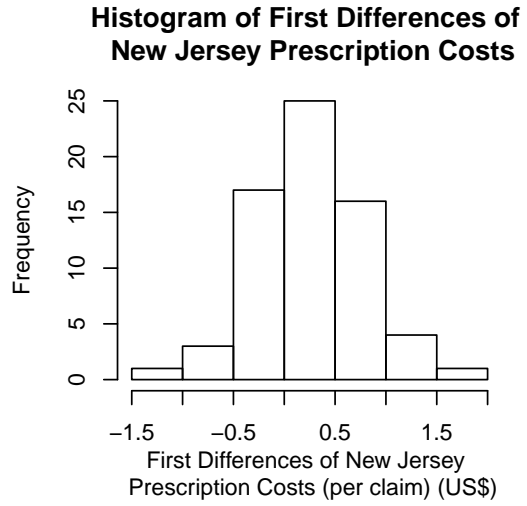
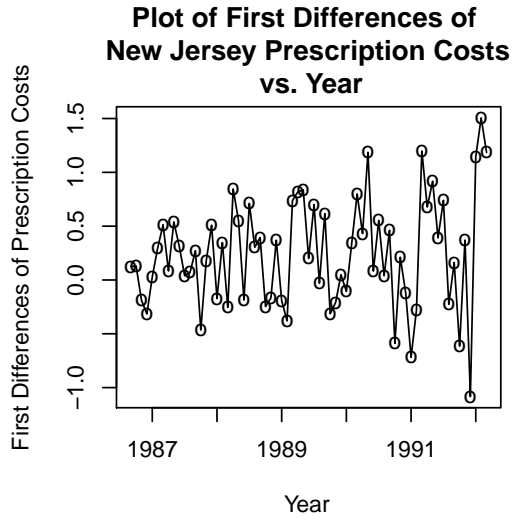
(b) Use the R command `diff.1 <- diff(prescrip)` to calculate the first differences $\nabla Y_t = Y_t - Y_{t-1}$ and plot the first differences. Describe the appearance of this plot and how it compares with the plot of the original series.

**Plot of First Differences of
New Jersey Prescription Costs vs. Year**



This first difference series appears to have zero mean. However, there is a prevalent increasing variance trend. Also, there seems to be some kind seasonal effect still, which is not surprising because first differencing can't remove seasonal effects. So, there is evidence that this first difference series is not stationary because its variance is not constant through time.

(c) Use all of the model diagnostic checks (Section 3.5 in the notes) on the difference process $\{\nabla Y_t\}$. Do the data differences resemble a normal zero mean white noise process?



Shapiro-Wilk normality test

```
data: diff.1
W = 0.9926, p-value = 0.9636
```

Runs test

```
$pvalue
```

```
[1] 0.874
```

```
$observed.runs
```

```
[1] 31
```

```
$expected.runs
```

```
[1] 29.83582
```

```
$n1
```

```
[1] 21
```

```
$n2
```

```
[1] 46
```

```
$k
```

```
[1] 0
```

As we saw in part (b), the constant variance assumption seems to be violated. However, the normality assumption holds very well, with a p-value of .96 and a supportive histogram and QQ plot. The independence assumption also looks plausible with the runs test returning a p-value of .87. Looking at the ACF, we can see the seasonal effect in the ACF. Therefore, I would be very reluctant to conclude that the first differences are a zero-mean white noise process.