

Since this is the first set of solutions for the semester, I will create a short list of expectations and advice.

1) My name is Brad Llewellyn. I am a Master's student in the Statistics Department. My e-mail address is llewellw@email.sc.edu.

2) These solutions were made after the homeworks were created. Therefore, some of the code I will use may be slightly different than the example code in the problems. If this is the case, I will tell you in my answer.

3) This should be a template for how your homeworks should look. Presentation and organization is just as important as information.

4) You will be creating a very large number of graphs in this course. EVERY graph should include a descriptive title and descriptive axis labels. You can use the `help()` feature in R if you don't know how to do this.

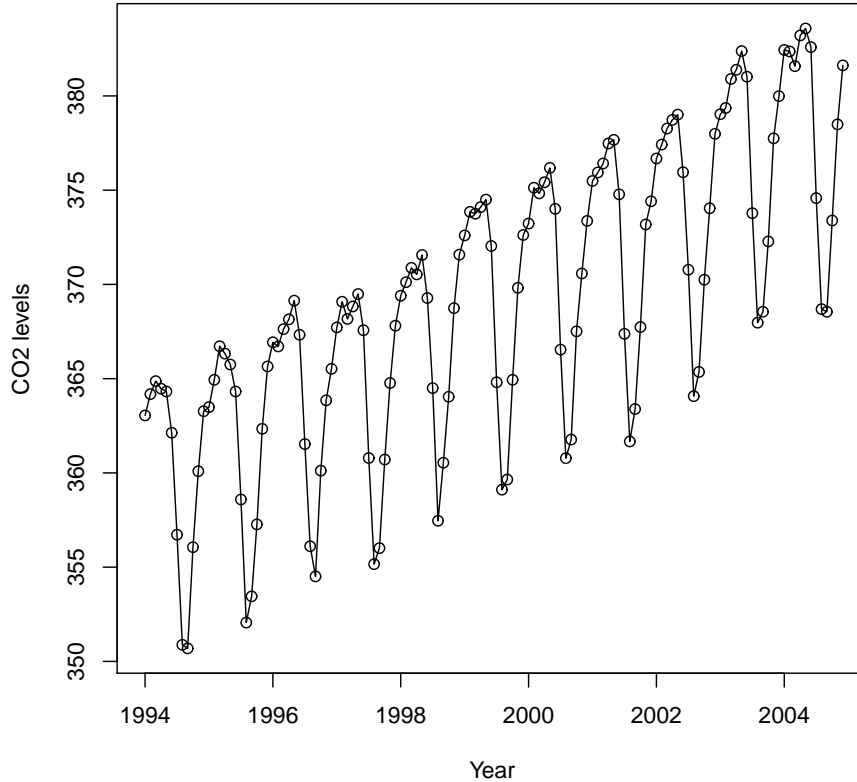
1. The TSA library contains the data set `co2`, which lists monthly carbon dioxide levels in northern Canada from 1/1994 to 12/2004. To load the data in R, remember that you need to first type

```
> library(TSA)
> data(co2)
```

(a) Construct a time series plot for these data. You can do this using the following R command:

```
> plot(co2,ylab="CO2 levels",xlab="Year",type="o")
```

Describe all systematic patterns you see in the plot.

Plot of CO₂ Levels over time

The code used to create this plot was

```
plot(co2, main = "Plot of CO2 Levels over time", ylab = "CO2 levels",
     xlab = "Year", type="o")
```

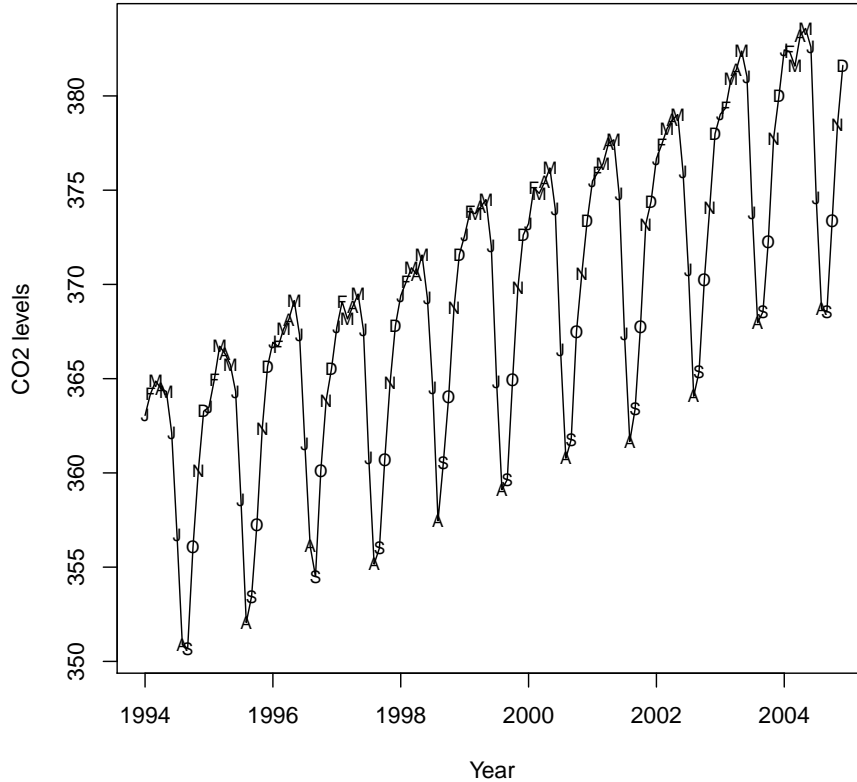
The most predominant pattern in this plot is the well-defined seasonal effect. The wavelength of this effect seems to be approximately one year. This is not surprising because CO₂ levels are definitely affected by climate, which is largely a yearly effect. The peaks of this pattern appear around May and the troughs appear around September. There is also perceptible increasing trend in this data. It seems to be increasing by about 1 unit per year.

(b) To enhance the usefulness of the plot, add monthly plotting symbols using the following R commands:

```
> plot(co2, ylab="CO2 levels", xlab="Year", type='l')
> points(y=co2, x=time(co2), pch=as.vector(season(co2)), cex=0.75)
```

Which months are consistently associated with highest CO₂ levels? the lowest? **Note:** The `cex=0.75` part controls the size of the plotting symbols specified in `pch`. Making `cex` larger increases the size of the plotting symbols.

Plot of CO2 Levels over time



This question was already answered in part (a).

(c) Consider fitting the simple straight line regression model (using the method of least squares)

$$Y_t = \beta_0 + \beta_1 t + e_t$$

to the data, where Y_t denotes the carbon dioxide level at time t . This model says that the CO_2 level is a linear function of time. You can do this using the R commands:

```
> model = lm(co2 ~ time(co2))
> summary(model)
```

What are the least squares estimates of β_0 and β_1 ? Write out the equation of the least squares regression line. In classical linear regression, what are the “usual” assumptions for the error terms e_t ?

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|-------|-------|
| -11.096 | -4.284 | 2.321 | 4.306 | 6.554 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|---------|------------|
| (Intercept) | -3127.2413 | 293.4126 | -10.66 | <2e-16 *** |
| time(co2) | 1.7486 | 0.1467 | 11.92 | <2e-16 *** |

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

Residual standard error: 5.354 on 130 degrees of freedom

Multiple R-squared: 0.522, Adjusted R-squared: 0.5184

F-statistic: 142 on 1 and 130 DF, p-value: < 2.2e-16

The least squares estimates for β_0 and β_1 are -3127.241 and 1.749, respectively.

Therefore, the equation for the least squares regression line is

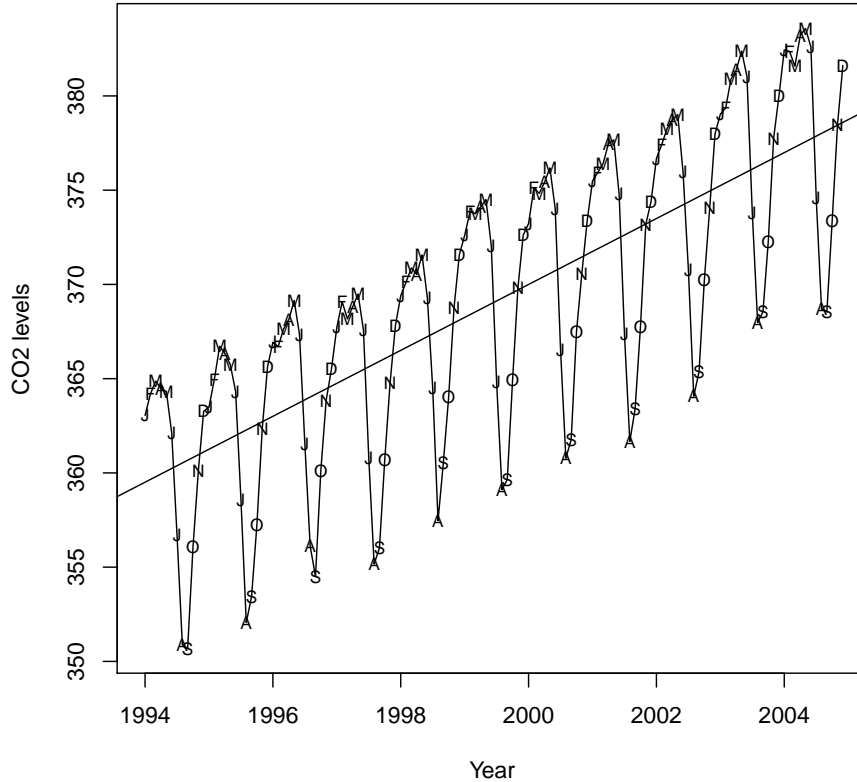
$$\text{CO}_2 = -3127.241 + 1.749 \times \text{Year}.$$

In classical linear regression, we assume that the errors are independent and identically distributed (iid) normal random variables with zero mean and a constant variance.

(d) Now construct a plot which superimposes the least squares fit over the time series plot from part (b). You can do this using the R commands:

```
> plot(co2,ylab="CO2 levels",xlab="Year",type='l')
> points(y=co2,x=time(co2),pch=as.vector(season(co2)),cex=0.75)
> abline(model)
```

Plot of CO2 Levels over time

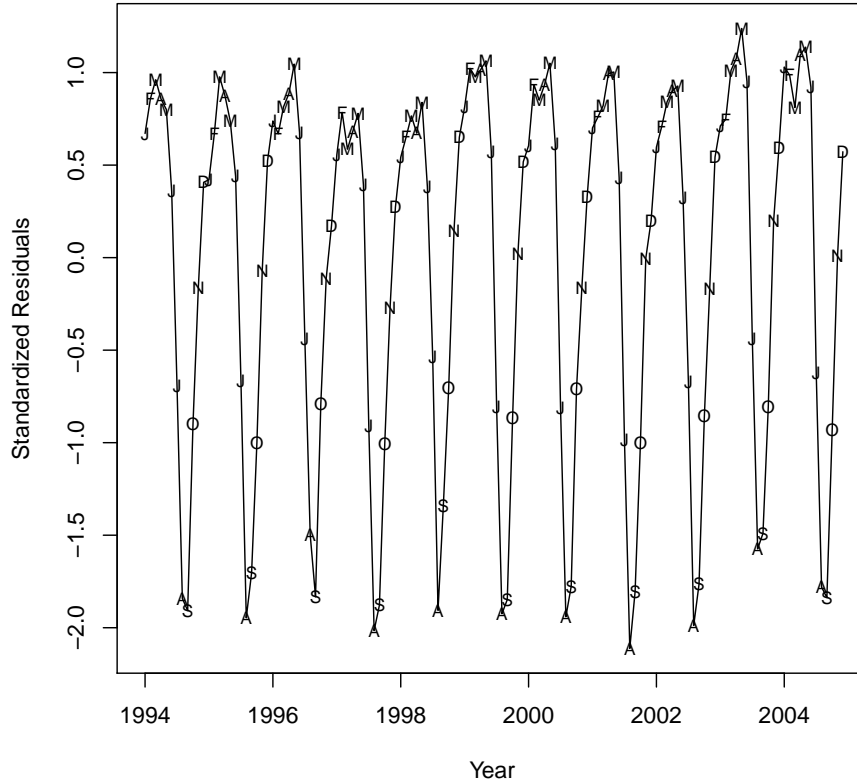


(e) As with any regression analysis, we can use the residuals to assess the quality of the model. Use the following R commands to create a plot of the residuals from the least squares fit versus time:

```
> plot(y=rstudent(model),x=as.vector(time(co2)),xlab="Year",
      ylab="Standardised residuals",type='l')
> points(y=rstudent(model),x=as.vector(time(co2)),
        pch=as.vector(season(co2)),cex=0.75)
```

Comment on the adequacy of the straight line regression model in part (c). What other types of models might do a better job in capturing the systematic part of this time series?

Plot of Studentized Residuals versus Year



There is still a significant sinusoidal pattern in these residuals. Therefore, the straight line regression model in part (a) is not appropriate. A much more appropriate model for this data would be one that shares this pattern, such as a cosine trend model.

2. Suppose that Z_1, Z_2, Z_3 are zero mean random variables with

$$\text{var}(Z_1) = 1 \quad \text{var}(Z_2) = 2 \quad \text{var}(Z_3) = 3$$

$$\text{cov}(Z_1, Z_2) = -0.5 \quad \text{cov}(Z_2, Z_3) = 2.5 \quad \text{cov}(Z_1, Z_3) = 0.$$

Calculate each of the following:

- $E(Z_1^2 - Z_2 - Z_2 Z_3)$
- $\text{var}(2Z_1 + 3Z_2 - Z_3)$
- $\text{cov}(3Z_1 - Z_2, Z_2 - 2Z_3)$
- $\text{corr}(Z_1, 2Z_2 + Z_3)$

$$\begin{aligned}
E(Z_1^2 - Z_2 - Z_2Z_3) &= E(Z_1^2) - E(Z_2) - E(Z_2Z_3) \\
&= [\text{Var}(Z_1) + E(Z_1)^2] - [E(Z_2)] - [\text{Cov}(Z_2, Z_3) + E(Z_2)E(Z_3)] \\
&= \text{Var}(Z_1) + E(Z_1)^2 - E(Z_2) - \text{Cov}(Z_2, Z_3) - E(Z_2)E(Z_3) \\
&= 1 + 0^2 - 0 - 2.5 - 0(0) \\
&= -1.5
\end{aligned}$$

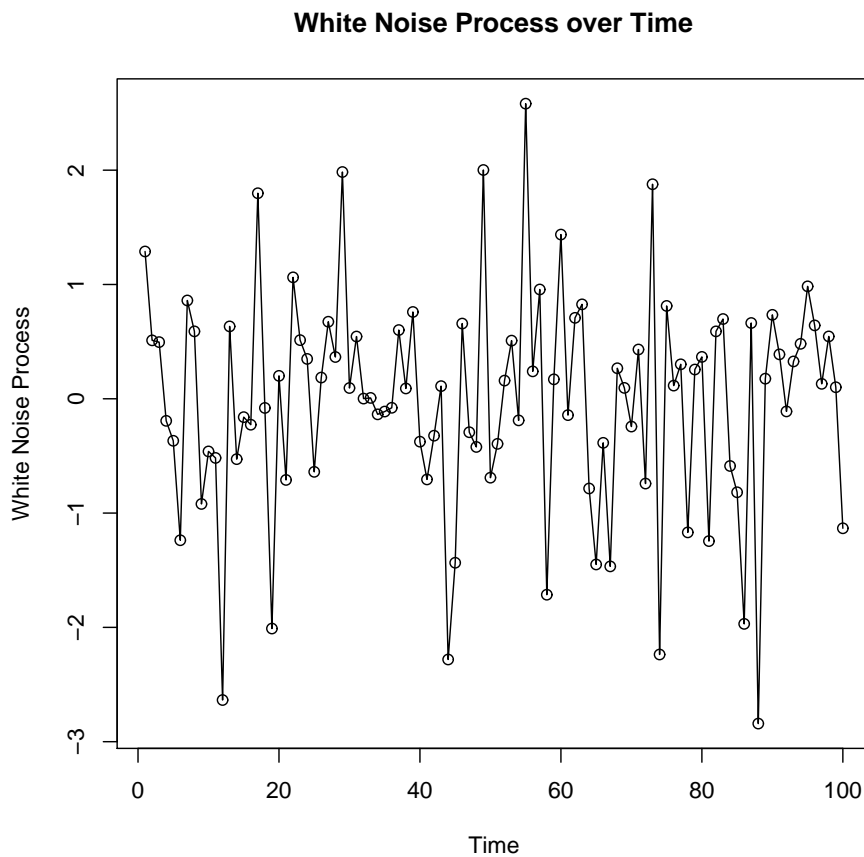
$$\begin{aligned}
\text{Var}(2Z_1 + 3Z_2 - Z_3) &= \text{Var}(2Z_1 + 3Z_2) + \text{Var}(Z_3) - 2\text{Cov}(2Z_1 + 3Z_2, Z_3) \\
&= [\text{Var}(2Z_1) + \text{Var}(3Z_2) + 2\text{Cov}(2Z_1, 3Z_2)] + \text{Var}(Z_3) - \\
&\quad 2[\text{Cov}(2Z_1, Z_3) + \text{Cov}(3Z_2, Z_3)] \\
&= 4\text{Var}(Z_1) + 9\text{Var}(Z_2) + 12\text{Cov}(Z_1, Z_2) + \text{Var}(Z_3) - \\
&\quad 4\text{Cov}(Z_1, Z_3) - 6\text{Cov}(Z_2, Z_3) \\
&= 4(1) + 9(2) + 12(-.5) + 3 - 4(0) - 6(2.5) \\
&= 4 + 18 - 6 + 3 - 15 \\
&= 4
\end{aligned}$$

$$\begin{aligned}
\text{Cov}(3Z_1 - Z_2, Z_2 - 2Z_3) &= \text{Cov}(3Z_1, Z_2 - 2Z_3) + \text{Cov}(-Z_2, Z_2 - 2Z_3) \\
&= [\text{Cov}(3Z_1, Z_2) + \text{Cov}(3Z_1, -2Z_3)] + \\
&\quad [\text{Cov}(-Z_2, Z_2) + \text{Cov}(-Z_2, -2Z_3)] \\
&= 3\text{Cov}(Z_1, Z_2) - 6\text{Cov}(Z_1, Z_3) - \\
&\quad \text{Cov}(Z_2, Z_2) + 2\text{Cov}(Z_2, Z_3) \\
&= 3\text{Cov}(Z_1, Z_2) - 6\text{Cov}(Z_1, Z_3) - \\
&\quad \text{Var}(Z_2) + 2\text{Cov}(Z_2, Z_3) \\
&= 3(-.5) - 6(0) - 2 + 2(2.5) \\
&= -1.5 - 2 + 5 \\
&= 1.5
\end{aligned}$$

$$\begin{aligned}
\text{Corr}(Z_1, 2Z_2 + Z_3) &= \frac{\text{Cov}(Z_1, 2Z_2 + Z_3)}{\sqrt{\text{Var}(Z_1)\text{Var}(2Z_2 + Z_3)}} \\
&= \frac{\text{Cov}(Z_1, 2Z_2) + \text{Cov}(Z_1, Z_3)}{\sqrt{\text{Var}(Z_1)[\text{Var}(2Z_2) + \text{Var}(Z_3) + 2\text{Cov}(2Z_2, Z_3)]}} \\
&= \frac{2\text{Cov}(Z_1, Z_2) + \text{Cov}(Z_1, Z_3)}{\sqrt{\text{Var}(Z_1)[4\text{Var}(Z_2) + \text{Var}(Z_3) + 4\text{Cov}(Z_2, Z_3)]}} \\
&= \frac{2 \times -.5 + 0}{\sqrt{1[4(2) + 3 + 4(2.5)]}} \\
&= \frac{-1}{\sqrt{8 + 3 + 10}} \\
&= -\frac{1}{\sqrt{21}}
\end{aligned}$$

3. (a) Simulate and plot a white noise process $e_t \sim \text{iid } \mathcal{N}(0, 1)$ of length $n = 100$ using the following commands in R:

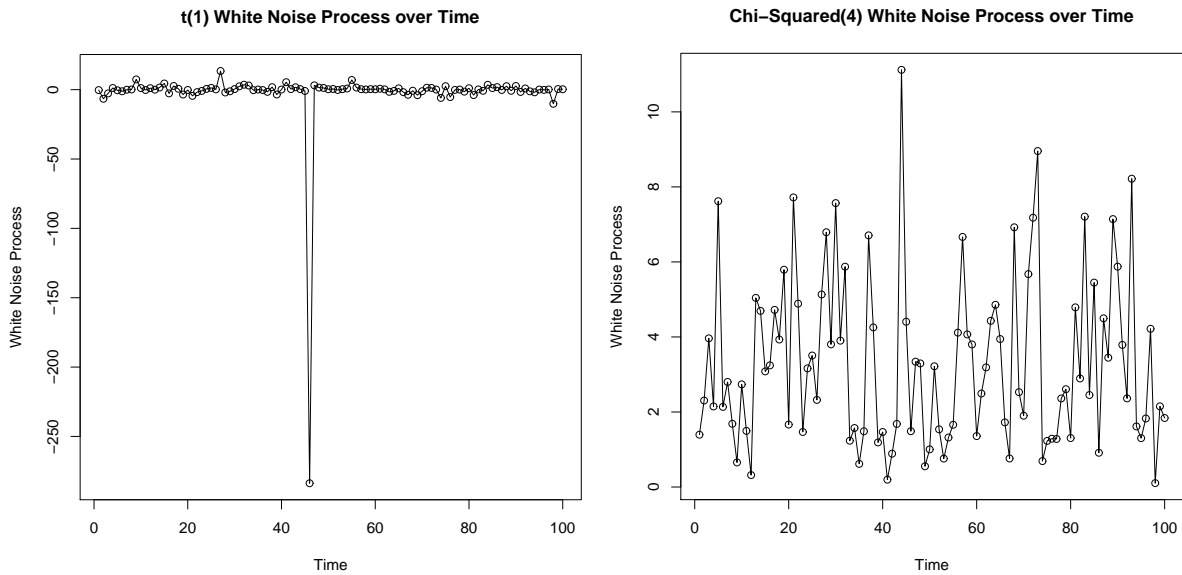
```
> wn.n01 = rnorm(100,0,1)
> plot(wn.n01,ylab="White noise process",xlab="Time",type="o")
```



(b) Repeat part (a) under the assumption that

- $e_t \sim \text{iid } t(1)$
- $e_t \sim \text{iid } \chi^2(4)$

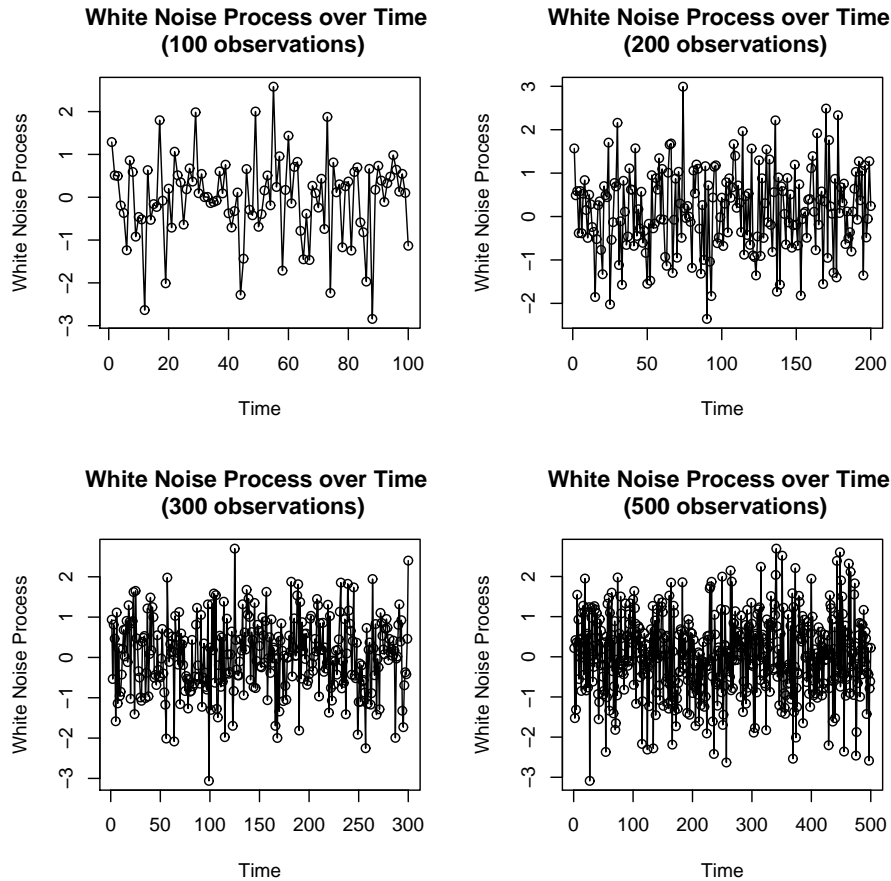
To do this, just replace the first line of the code above with `wn.t1 = rt(100,1)` and `wn.chisq4 = rchisq(100,4)`, respectively. Comment on the differences among the 3 simulated white noise processes.



The normal and t processes have a similar, almost symmetric shape, with the exception that the t process has an extremely small observation. This is expected because the t distribution with 1 degree of freedom is a mound-shaped distribution, similar to the normal distribution, yet with a higher probability of obtaining extreme values. Students with a background in Mathematical Statistics may recognize a t -distribution with 1 degree of freedom as being a Cauchy distribution. The Chi-Squared process is a little different. The most important of which is the fact that it's not symmetric. It also is incapable of obtaining negative values and has a higher number of large values than the other processes.

(c) Repeat part (a) using $n = 200$, $n = 300$, and $n = 500$. With your plot from part (a), take your 4 standard normal white noise processes and put them in a 2×2 matrix of plots using the `par(mfrow=c(2,2))` command in R. Label each plot in the matrix according to the sample size used, e.g.,

```
plot(wn.n01.100,ylab="WN",xlab="Time",main="Sample.size=100",type="o").
```



4. Do the following problems in Chapter 2 from Cryer and Chan: 2.1, 2.2, and 2.4.

2.1 Suppose $E(X) = 2$, $\text{Var}(X) = 9$, $E(Y) = 0$, $\text{Var}(Y) = 4$, and $\text{Corr}(X, Y) = 0.25$. Find:

(a) $\text{Var}(X + Y)$.

$$\begin{aligned}
 \text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y) \\
 &= \text{Var}(X) + \text{Var}(Y) + 2\text{Corr}(X, Y)\sqrt{\text{Var}(X)\text{Var}(Y)} \\
 &= 9 + 4 + 2(.25)\sqrt{9(4)} \\
 &= 9 + 4 + 2(.25)(3)(2) \\
 &= 16
 \end{aligned}$$

(b) $\text{Cov}(X, X + Y)$.

$$\begin{aligned}
 \text{Cov}(X, X + Y) &= \text{Cov}(X, X) + \text{Cov}(X, Y) \\
 &= \text{Var}(X) + \text{Corr}(X, Y)\sqrt{\text{Var}(X)\text{Var}(Y)} \\
 &= 9 + .25\sqrt{9(4)} \\
 &= 9 + .25(3)(2) \\
 &= 10.5
 \end{aligned}$$

(c) $\text{Corr}(X + Y, X - Y)$.

$$\begin{aligned}
 \text{Corr}(X + Y, X - Y) &= \frac{\text{Cov}(X + Y, X - Y)}{\sqrt{\text{Var}(X + Y)\text{Var}(X - Y)}} \\
 &= \frac{\text{Cov}(X + Y, X) + \text{Cov}(X + Y, -Y)}{\sqrt{\text{Var}(X + Y) [\text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)]}} \\
 &= \frac{\text{Cov}(X, X) + \text{Cov}(Y, X) + \text{Cov}(X, -Y) + \text{Cov}(Y, -Y)}{\sqrt{\text{Var}(X + Y) [\text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)]}} \\
 &= \frac{\text{Var}(X) + \text{Cov}(X, Y) - \text{Cov}(X, Y) - \text{Var}(Y)}{\sqrt{\text{Var}(X + Y) [\text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)]}} \\
 &= \frac{\text{Var}(X) - \text{Var}(Y)}{\sqrt{\text{Var}(X + Y) [\text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)]}} \\
 &= \frac{9 - 4}{\sqrt{16 [9 + 4 - 2 \times .25\sqrt{9 \times 4}]}} \\
 &= \frac{5}{\sqrt{16 [9 + 4 - 2 \times .25 \times 3 \times 2]}} \\
 &= \frac{5}{\sqrt{16 [10]}} \\
 &= \frac{5}{\sqrt{160}} \approx .395
 \end{aligned}$$

2.2 If X and Y are dependent, but $\text{Var}(X) = \text{Var}(Y)$, find $\text{Cov}(X + Y, X - Y)$.

In Problem **2.1(c)**, we showed that $\text{Cov}(X + Y, X - Y) = \text{Var}(X) - \text{Var}(Y)$. Therefore, if $\text{Var}(X) = \text{Var}(Y)$, then $\text{Cov}(X + Y, X - Y) = 0$.

2.4 Let $\{e_t\}$ be a zero mean white noise process. Suppose that the observed process is $Y_t = e_t + \theta e_{t-1}$, where θ is either 3 or $\frac{1}{3}$.

(a) Find the autocorrelation function for $\{Y_t\}$ both when $\theta = 3$ and when $\theta = \frac{1}{3}$.

First, let's find the autocovariance function for $\{Y_t\}$.

$$\begin{aligned}
 \gamma_{t,s} &= \text{Cov}(Y_t, Y_s) \\
 &= \text{Cov}(e_t + \theta e_{t-1}, e_s + \theta e_{s-1}) \\
 &= \text{Cov}(e_t, e_s) + \text{Cov}(e_t, \theta e_{s-1}) + \text{Cov}(\theta e_{t-1}, e_s) + \text{Cov}(\theta e_{t-1}, \theta e_{s-1}) \\
 &= \text{Cov}(e_t, e_s) + \theta \text{Cov}(e_t, e_{s-1}) + \theta \text{Cov}(e_{t-1}, e_s) + \theta^2 \text{Cov}(e_{t-1}, e_{s-1})
 \end{aligned}$$

Now, let's break this derivation into three parts. Let $k = t - s$, where $t \geq s$.

Part 1: $k = 0$, i.e. $s = t$.

$$\begin{aligned}\gamma_0 &= \text{Cov}(e_t, e_t) + \theta \text{Cov}(e_t, e_{t-1}) + \theta \text{Cov}(e_{t-1}, e_t) + \theta^2 \text{Cov}(e_{t-1}, e_{t-1}) \\ &= \text{Var}(e_t) + 2\theta \text{Cov}(e_t, e_{t-1}) + \theta^2 \text{Var}(e_{t-1}) \\ &= \sigma^2 + 2\theta \times 0 + \theta^2 \sigma^2 \\ &= \sigma^2(1 + \theta^2)\end{aligned}$$

Part 2: $k = 1$, i.e. $s = t - 1$.

$$\begin{aligned}\gamma_1 &= \text{Cov}(e_t, e_{t-1}) + \theta \text{Cov}(e_t, e_{t-2}) + \theta \text{Cov}(e_{t-1}, e_{t-1}) + \theta^2 \text{Cov}(e_{t-1}, e_{t-2}) \\ &= \text{Cov}(e_t, e_{t-1}) + \theta \text{Cov}(e_t, e_{t-2}) + \theta \text{Var}(e_{t-1}, e_{t-1}) + \theta^2 \text{Cov}(e_{t-1}, e_{t-2}) \\ &= 0 + \theta \times 0 + \theta \sigma^2 + \theta^2 \times 0 \\ &= \sigma^2 \theta\end{aligned}$$

Part 3: $k \geq 2$.

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

So, the autocovariance function for $\{Y_t\}$ is

$$\gamma_k = \begin{cases} \sigma^2(1 + \theta^2), & k = 0 \\ \sigma^2 \theta, & k = 1 \\ 0, & k \geq 2 \end{cases}$$

Now, let's find the autocorrelation function for $\{Y_t\}$.

$$\begin{aligned}
\rho_k &= \text{Corr}(Y_t, Y_s) \\
&= \frac{\text{Cov}(Y_t, Y_s)}{\sqrt{\text{Var}(Y_t)\text{Var}(Y_s)}} \\
&= \frac{\gamma_k}{\sqrt{\gamma_0\gamma_0}} \\
&= \frac{\gamma_k}{\gamma_0} \\
&= \begin{cases} 1, & k = 0 \\ \frac{\sigma^2\theta}{\sigma^2(1+\theta^2)}, & k = 1 \\ 0, & k \geq 2 \end{cases} \\
&= \begin{cases} 1, & k = 0 \\ \frac{\theta}{1+\theta^2}, & k = 1 \\ 0, & k \geq 2 \end{cases}
\end{aligned}$$

It's obvious that this function does not depend on t . Combine this with the fact that $E(Y_t) = E(e_t + \theta e_{t-1}) = E(e_t) + \theta E(e_{t-1}) = 0 + \theta \times 0 = 0$ and we have that this series is stationary.

If we assume $\theta = 3$, then

$$\rho_k = \begin{cases} 1, & k = 0 \\ \frac{3}{10}, & k = 1 \\ 0, & k \geq 2 \end{cases}$$

If we assume $\theta = \frac{1}{3}$, we also have

$$\rho_k = \begin{cases} 1, & k = 0 \\ \frac{3}{10}, & k = 1 \\ 0, & k \geq 2 \end{cases}$$

Therefore, the autocorrelation function for $\{Y_t\}$ is the same regardless of which choice we make for θ .

(b) You should have discovered that the time series is stationary regardless of the value of θ and that the autocorrelation functions are the same for $\theta = 3$ and $\theta = \frac{1}{3}$. For simplicity, suppose that the process mean is known to be zero and the variance of Y_t is known to be 1. You observe the series $\{Y_t\}$ for $t = 1, 2, \dots, n$ and suppose that you can produce good estimates of the autocorrelations ρ_k . Do you think that you could determine which value of θ is correct (3 or $\frac{1}{3}$) based on the estimate of ρ_k ? Why or why not?

If we restrict ourselves to only using the estimates of the autocorrelations, then it would be impossible to determine whether or not the sample came from the population with $\theta = 3$ or $\theta = \frac{1}{3}$. This is known as an "identifiability" issue. However, if we expand our reach, we could still accurately estimate θ . If we observe values from the series and can obtain good estimates of the autocorrelations, then it isn't much of a leap to assume that we can also obtain good estimates of the autocovariances. Therefore, we could use the sample autocovariances to determine which θ was used.