

1. Suppose that  $\{e_t\}$  is a zero mean white noise process with  $\text{var}(e_t) = \sigma_e^2$ , and consider the AR(1) model with  $\phi = 0.8$ , that is,

$$Y_t = 0.8Y_{t-1} + e_t.$$

Suppose that we have observed a series of length  $n = 200$  from this AR(1) process.

(a) Write out the approximate (large-sample) sampling distributions of  $r_1$ ,  $r_2$ ,  $r_5$ , and  $r_{10}$ .

From the notes, we know that for large samples,

$$r_k \sim AN \left( \phi^k, \frac{1}{n} \left[ \frac{\{1 + \phi^2\}\{1 - \phi^{2k}\}}{1 - \phi^2} - 2k\phi^{2k} \right] \right).$$

So, we have the following:

$$\begin{aligned} r_1 &\sim AN \left( \phi^1, \frac{1}{n} \left[ \frac{\{1 + \phi^2\}\{1 - \phi^2\}}{1 - \phi^2} - 2\phi^2 \right] \right) \\ r_2 &\sim AN \left( \phi^2, \frac{1}{n} \left[ \frac{\{1 + \phi^2\}\{1 - \phi^4\}}{1 - \phi^2} - 4\phi^4 \right] \right) \\ r_5 &\sim AN \left( \phi^5, \frac{1}{n} \left[ \frac{\{1 + \phi^2\}\{1 - \phi^{10}\}}{1 - \phi^2} - 10\phi^{10} \right] \right) \\ r_{10} &\sim AN \left( \phi^{10}, \frac{1}{n} \left[ \frac{\{1 + \phi^2\}\{1 - \phi^{20}\}}{1 - \phi^2} - 20\phi^{20} \right] \right) \end{aligned}$$

Next, using the following code, we generate

(b) Use Monte Carlo simulation, like that in Example 6.2 (notes), to simulate the sampling distributions of  $r_1$ ,  $r_2$ ,  $r_5$ , and  $r_{10}$  when  $n = 200$ . Assume that  $e_t \sim \text{iid } \mathcal{N}(0, 1)$ . Do these sampling distributions agree with what the large-sample theory says should happen?

For this simulation, we chose  $\phi = .8$  and  $M = 2000$  to keep continuity with Example 6.2. Your choices for  $\phi$  and  $M$  may differ, which will cause your simulations to differ slightly as well. First, let's find the asymptotic distribution for  $r_1, r_2, r_5$ , and  $r_{10}$ . The values were obtained using a calculator.

$$r_1 \sim AN(.8, .00180)$$

$$r_2 \sim AN(.64, .00526)$$

$$r_5 \sim AN(.328, .01496)$$

$$r_{10} \sim AN(.107, .02136)$$

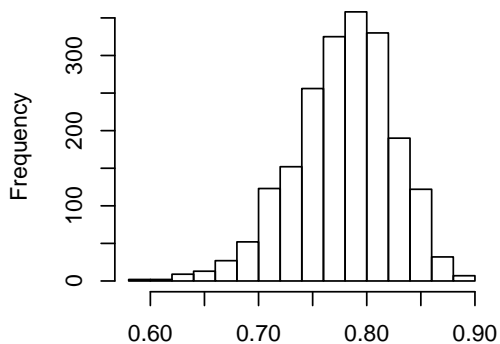
Using the following code, we simulated each of these values  $M = 2000$  times.

```
> M <- 2000
> n <- 200
> phi <- .8
> r1 <- c()
> r2 <- c()
> r5 <- c()
> r10 <- c()
> sim <- c()

> for(i in 1:M){
+ sim <- arima.sim(list(ar = phi), n = n)
+
+ r1[i] <- acf(sim, plot = F)[[1]][1]
+ r2[i] <- acf(sim, plot = F)[[1]][2]
+ r5[i] <- acf(sim, plot = F)[[1]][5]
+ r10[i] <- acf(sim, plot = F)[[1]][10]
+ }
```

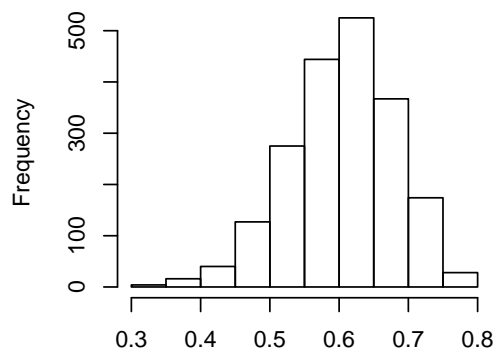
The results from the simulation are as follows:

**Histogram of Lag 1  
Sample Autocorrelation**



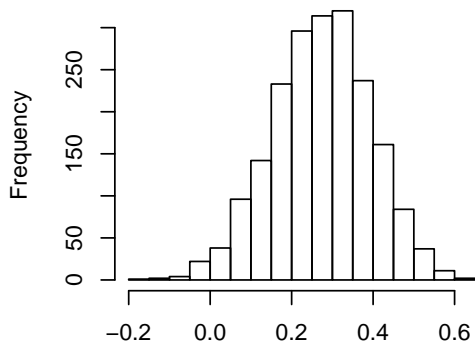
Lag 1 Sample Autocorrelation

**Histogram of Lag 2  
Sample Autocorrelation**



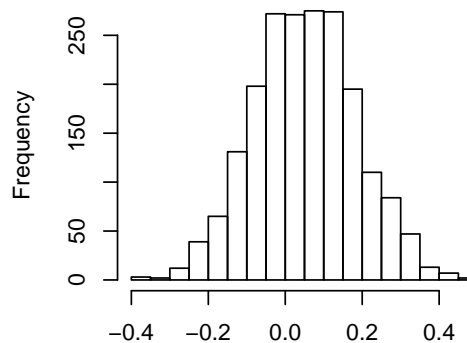
Lag 2 Sample Autocorrelation

**Histogram of Lag 5  
Sample Autocorrelation**



Lag 5 Sample Autocorrelation

**Histogram of Lag 10  
Sample Autocorrelation**



Lag 10 Sample Autocorrelation

```
> mean(r1)
[1] 0.778398
> var(r1)
[1] 0.002102658
> shapiro.test(r1)
```

Shapiro-Wilk normality test

```
data: r1
W = 0.9852, p-value = 1.506e-13
```

```
> mean(r2)
[1] 0.6042862
> var(r2)
[1] 0.005838953
> shapiro.test(r2)
```

## Shapiro-Wilk normality test

```
data: r2
W = 0.9904, p-value = 3.225e-10
```

```
> mean(r5)
[1] 0.2733858
> var(r5)
[1] 0.01427647
> shapiro.test(r5)
```

## Shapiro-Wilk normality test

```
data: r5
W = 0.9981, p-value = 0.01963
```

```
> mean(r10)
[1] 0.05235349
> var(r10)
[1] 0.01774757
> shapiro.test(r10)
```

## Shapiro-Wilk normality test

```
data: r10
W = 0.9986, p-value = 0.1133
```

Before we begin the analysis, it should be noted that while all of these claims can be rigorously tested in a large number of ways, for the sake of this problem, we need only look at the values and see if they are "good enough." First, let's look at  $r_1$ .

	Truth	Observation
Mean	.8	.778
Variance	.00180	.00210
Normality	Yes	No

As you can see, the mean and variance are pretty close to what they should be. However, normality doesn't seem to hold. There seems to be a small amount of skewness on the histogram. But, this is rather intuitive because we are attempting to model a bounded variable with an unbounded distribution.

Now, let's look at  $r_2$ .

	Truth	Observation
Mean	.64	.604
Variance	.00526	.00584
Normality	Yes	No

Again, we see that the observed values are close to their theorized values. However, normality is still an issue for  $r_2$ . This is to be expected because the theoretical mean of .64 is still relatively close to its upper bound of 1.

Next, let's examine  $r_5$ .

	Truth	Observation
Mean	.328	.273
Variance	.01496	.01428
Normality	Yes	Mildly Non-normal

The observed means seem to deviating from their true values. However, this can be explained because the variances are increasing, thereby increasing the standard errors of the sample means, and the means are getting close to zero. Therefore, while we see the observed mean being half of its asymptotic value, this is simply due to random chance and is not uncommon. Also, normality is becoming more reasonable because the theoretical means are deviating from the (-1, 1) boundaries.

Finally, let's look at  $r_{10}$ .

	Truth	Observation
Mean	.107	.0524
Variance	.02136	.01775
Normality	Yes	Yes

The explanation for  $r_{10}$  follows similar logic to that for  $r_5$ , with the exception that normality is now quite reasonable. On a different note, the Shapiro-Wilk Test, along with many other hypothesis tests, is affected by sample size. With a sample size of  $M = 2000$ , it is very likely that many of the sample values will significantly deviate from their theoretical means. Therefore, the p-value should be taken very lightly. One could make the argument that it would be much more appropriate to look at the histograms and make the decision for yourself.

2. In Chapter 6, we described testing hypothesis testing procedures to select MA and AR models using sample autocorrelations and sample partial autocorrelations, respectively.

Use these testing procedures in each of the following situations.

(a) From a time series of  $n = 150$  observations, we compute the following sample autocorrelation coefficients:  $r_1 = -0.37$ ,  $r_2 = 0.28$ ,  $r_3 = 0.31$ ,  $r_4 = -0.13$ ,  $r_5 = 0.04$ , and  $r_6 = -0.15$ . The remaining sample autocorrelations appear to be negligible. Find the MA( $q$ ) process most consistent with this information, that is, determine the value of  $q$ .

In order to find the most appropriate value of  $q$ , we need to conduct a series of tests to find the first value of  $q$  that leaves us with a "reasonable" model. Refer to Example 6.1 in the notes for a more rigorous version of this test. We will control our  $\alpha$  at the .05 level. Therefore, our rejection region for each test will be  $\{z^* : |z^*| > z_{.025} = 1.96\}$ .

Test 1:  $H_0 : q = 0$  vs.  $H_a : q \neq 0$ .

$$\begin{aligned}
 Z^* &= \frac{r_{q+1}}{\sqrt{\frac{1}{n} \left(1 + 2 \sum_{j=1}^q r_j^2\right)}} \\
 &= \frac{r_1}{\sqrt{\frac{1}{n} \left(1 + 2 \sum_{j=1}^0 r_j^2\right)}} \\
 &= \frac{-0.37}{\sqrt{\frac{1}{150} (1 + 2[0])}} \\
 &= \frac{-0.37}{\sqrt{\frac{1}{150}(1)}} \\
 &= \frac{-0.37}{\sqrt{\frac{1}{150}}} \\
 &= -4.532
 \end{aligned}$$

Since our test statistic falls in the rejection region, we have significant evidence that our null hypothesis is not true. So, we conclude that  $q \neq 0$ . Now, let's move on to  $q = 1$ .

Test 2:  $H_0 : q = 1$  vs.  $H_a : q \neq 1$ .

$$\begin{aligned}
Z^* &= \frac{r_{q+1}}{\sqrt{\frac{1}{n} \left(1 + 2 \sum_{j=1}^q r_j^2\right)}} \\
&= \frac{r_2}{\sqrt{\frac{1}{n} \left(1 + 2 \sum_{j=1}^1 r_j^2\right)}} \\
&= \frac{.28}{\sqrt{\frac{1}{150} (1 + 2r_1^2)}} \\
&= \frac{.28}{\sqrt{\frac{1}{150} (1 + 2[-.37^2])}} \\
&= \frac{.28}{\sqrt{\frac{1}{150} (1 + 2[.137])}} \\
&= \frac{.28}{\sqrt{\frac{1}{150} (1 + .274)}} \\
&= \frac{.28}{\sqrt{\frac{1}{150}(1.274)}} \\
&= 3.038
\end{aligned}$$

Since our test statistic falls in the rejection region, we have significant evidence that our null hypothesis is not true. So, we conclude that  $q \neq 1$ . Now, let's move on to  $q = 2$ .

Test 3:  $H_0 : q = 2$  vs.  $H_a : q \neq 2$ .

$$\begin{aligned}
Z^* &= \frac{r_{q+1}}{\sqrt{\frac{1}{n} \left( 1 + 2 \sum_{j=1}^q r_j^2 \right)}} \\
&= \frac{r_3}{\sqrt{\frac{1}{n} \left( 1 + 2 \sum_{j=1}^2 r_j^2 \right)}} \\
&= \frac{.31}{\sqrt{\frac{1}{150} (1 + 2[-.37^2 + .28^2])}} \\
&= \frac{.31}{\sqrt{\frac{1}{150} (1 + 2[.137 + .078])}} \\
&= \frac{.31}{\sqrt{\frac{1}{150} (1 + 2[.215])}} \\
&= \frac{.31}{\sqrt{\frac{1}{150} (1 + .430)}} \\
&= \frac{.31}{\sqrt{\frac{1}{150} (1.430)}} \\
&= 3.147
\end{aligned}$$

Since our test statistic falls in the rejection region, we have significant evidence that our null hypothesis is not true. So, we conclude that  $q \neq 2$ . Now, let's move on to  $q = 3$ .

Test 4:  $H_0 : q = 3$  vs.  $H_a : q \neq 3$ .

$$\begin{aligned}
Z^* &= \frac{r_{q+1}}{\sqrt{\frac{1}{n} \left(1 + 2 \sum_{j=1}^q r_j^2\right)}} \\
&= \frac{r_4}{\sqrt{\frac{1}{n} \left(1 + 2 \sum_{j=1}^3 r_j^2\right)}} \\
&= \frac{-0.13}{\sqrt{\frac{1}{150} \left(1 + 2[-.37^2 + .28^2 + .31^2]\right)}} \\
&= \frac{-0.13}{\sqrt{\frac{1}{150} \left(1 + 2[.137 + .078 + .096]\right)}} \\
&= \frac{-0.13}{\sqrt{\frac{1}{150} \left(1 + 2[.311]\right)}} \\
&= \frac{-0.13}{\sqrt{\frac{1}{150} \left(1 + .622\right)}} \\
&= \frac{-0.13}{\sqrt{\frac{1}{150} (1.622)}} \\
&= -1.250
\end{aligned}$$

Since this value does not fall in our rejection region, we do not have significant evidence to say that our null hypothesis is false. So, we conclude that  $q = 3$  is a reasonable choice for this series.

(b) From a time series of  $n = 150$  observations, we compute the following sample partial autocorrelation coefficients:  $\hat{\phi}_{11} = 0.24$ ,  $\hat{\phi}_{22} = -0.81$ ,  $\hat{\phi}_{33} = 0.31$ ,  $\hat{\phi}_{44} = -0.09$ ,  $\hat{\phi}_{55} = 0.04$ , and  $\hat{\phi}_{66} = -0.02$ . The remaining sample partial autocorrelations appear to be negligible. Find the  $\text{AR}(p)$  process most consistent with this information, that is, determine the value of  $p$ .

In order to find the most appropriate value of  $p$ , we need to conduct a series of tests to find the first value of  $p$  that leaves us with a "reasonable" model. Refer to Example 6.1 in the notes for a more rigorous version of this test in the  $\text{MA}(\mathbf{q})$  case. We will control our  $\alpha$  at the .05 level. Therefore, our rejection region for each test will be  $\{z^* : |z^*| > z_{.025} = 1.96\}$ .

Test 1:  $H_0 : p = 0$  vs.  $H_a : p \neq 0$ .

$$\begin{aligned}
 Z^* &= \frac{\hat{\phi}_{p+1,p+1}}{\sqrt{\frac{1}{n}}} \\
 &= \frac{\hat{\phi}_{1,1}}{\sqrt{\frac{1}{150}}} \\
 &= \frac{.24}{\sqrt{\frac{1}{150}}} \\
 &= 2.940
 \end{aligned}$$

Since our test statistic falls in the rejection region, we have significant evidence that our null hypothesis is not true. So, we conclude that  $p \neq 0$ . Now, let's move on to  $p = 1$ .

Test 2:  $H_0 : p = 1$  vs.  $H_a : p \neq 1$ .

$$\begin{aligned}
 Z^* &= \frac{\hat{\phi}_{p+1,p+1}}{\sqrt{\frac{1}{n}}} \\
 &= \frac{\hat{\phi}_{2,2}}{\sqrt{\frac{1}{150}}} \\
 &= \frac{-.81}{\sqrt{\frac{1}{150}}} \\
 &= -9.920
 \end{aligned}$$

Since our test statistic falls in the rejection region, we have significant evidence that our null hypothesis is not true. So, we conclude that  $p \neq 1$ . Now, let's move on to  $p = 2$ .

Test 3:  $H_0 : p = 2$  vs.  $H_a : p \neq 2$ .

$$\begin{aligned}
 Z^* &= \frac{\hat{\phi}_{p+1,p+1}}{\sqrt{\frac{1}{n}}} \\
 &= \frac{\hat{\phi}_{3,3}}{\sqrt{\frac{1}{150}}} \\
 &= \frac{.31}{\sqrt{\frac{1}{150}}} \\
 &= 3.797
 \end{aligned}$$

Since our test statistic falls in the rejection region, we have significant evidence that our null hypothesis is not true. So, we conclude that  $p \neq 2$ . Now, let's move on to  $p = 3$ .

Test 4:  $H_0 : p = 3$  vs.  $H_a : p \neq 3$ .

$$\begin{aligned} Z^* &= \frac{\hat{\phi}_{p+1,p+1}}{\sqrt{\frac{1}{n}}} \\ &= \frac{\hat{\phi}_{4,4}}{\sqrt{\frac{1}{150}}} \\ &= \frac{-.09}{\sqrt{\frac{1}{150}}} \\ &= -1.102 \end{aligned}$$

Since this value does not fall in our rejection region, we do not have significant evidence to say that our null hypothesis is false. So, we conclude that  $p = 3$  is a reasonable choice for this series.

3. Generate three time series data sets, each of length  $n = 200$ , including (i) an AR(1) with  $\phi = -0.6$ , (ii) an MA(1) with  $\theta = 0.8$ , and (iii) an ARMA(1,1) with  $\phi = -0.6$  and  $\theta = 0.8$ . For each one,

- (a) plot the observed time series.
- (b) plot the sample ACF, the sample PACF, and the sample EACF.
- (c) use the `armasubsets` function in R to identify the best model in terms of the BIC.

Do the plots in part (b) agree with what you know to be true? Remember, you know the correct models! That is, you are assessing here whether the sample identification functions agree with the truth. Does the BIC identify the correct model as the “best” model in each case? If not, where is the correct model ranked, if at all?

### (i) AR(1) process

The code used to generate the output is as follows:

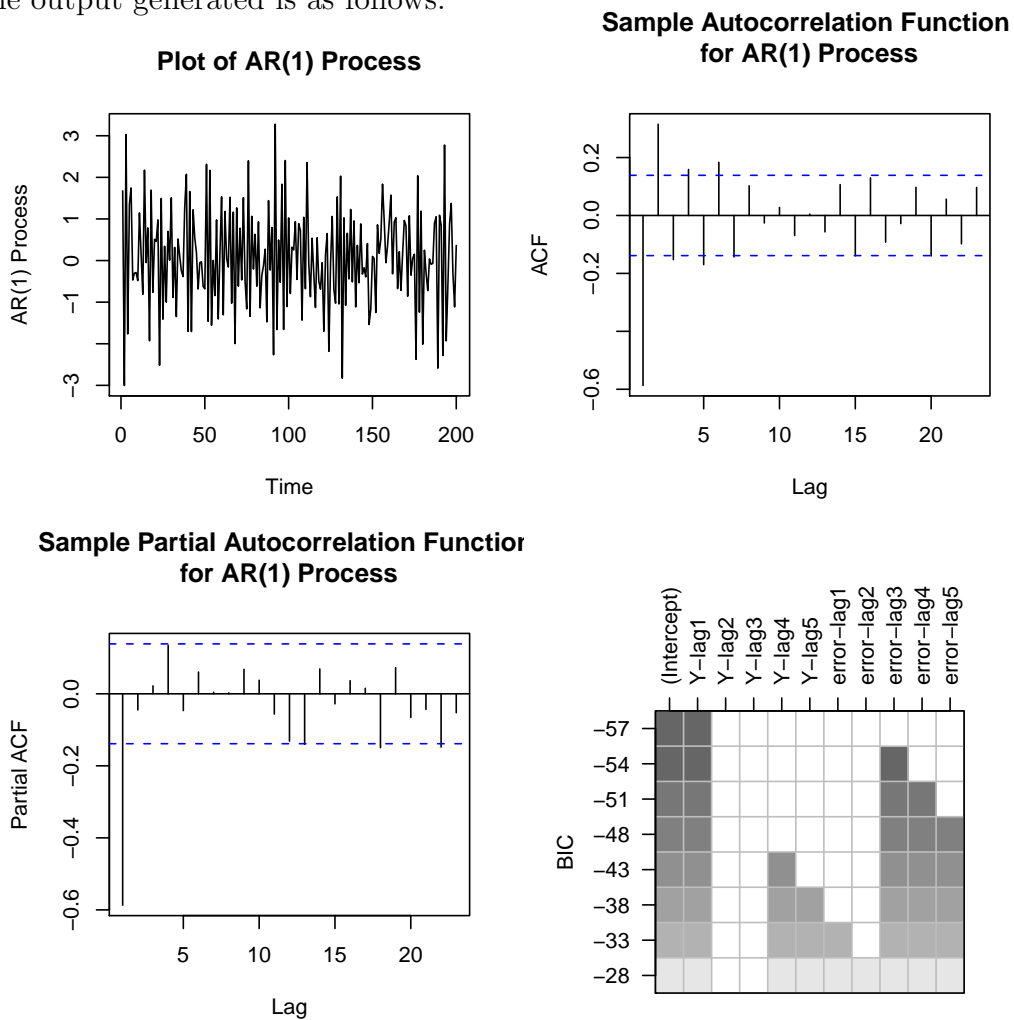
```
> ar1 <- arima.sim(list(ar = -.6), n = 200)
> sub <- armasubsets(ar1, nar = 5, nma = 5)
>
```

```

> par(mfrow = c(2,2))
>
> plot.ts(ar1, main = "Plot of AR(1) Process", xlab = "Time",
+ ylab = "AR(1) Process")
> acf(ar1, main = "Sample Autocorrelation Function
+ for AR(1) Process")
> pacf(ar1, main = "Sample Partial Autocorrelation
+ Function for AR(1) Process")
> plot.armsubsets(sub)
>
> eacf(ar1)

```

The output generated is as follows:



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

Looking at the sample ACF, we can see a faint exponential decay trend, which implies that an autoregressive element exists in the model. The sample PACF strongly implies that this is an **AR(1) process**. The sample EACF and the BIC agree with this conclusion. Therefore, using the model identification tools above, we can properly conclude that the data belongs to an **AR(1) process**.

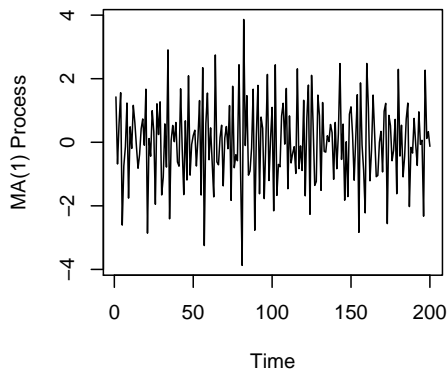
## (ii) MA(1) process

The code used to generate the output is as follows:

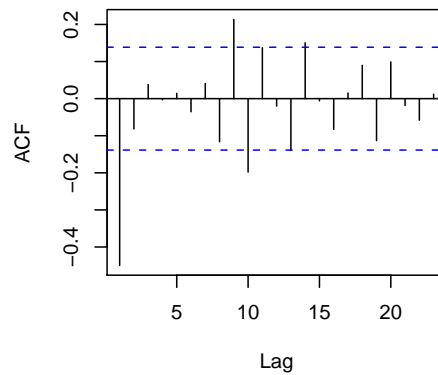
```
> ma1 <- arima.sim(list(ma = -.8), n = 200)
> sub <- armasubsets(ma1, nar = 5, nma = 5)
>
> par(mfrow = c(2,2))
>
> plot.ts(ma1, main = "Plot of MA(1) Process", xlab = "Time",
+ ylab = "MA(1) Process")
> acf(ma1, main = "Sample Autocorrelation Function
+ for MA(1) Process")
> pacf(ma1, main = "Sample Partial Autocorrelation
+ Function for MA(1) Process")
> plot.armsubsets(sub)
>
> eacf(ma1)
```

The output generated is as follows:

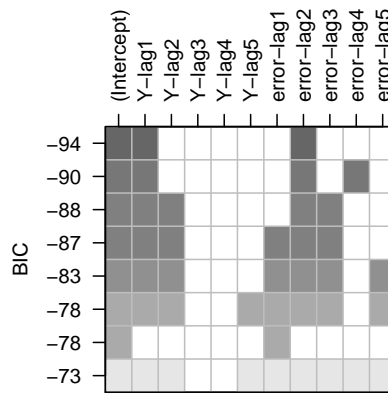
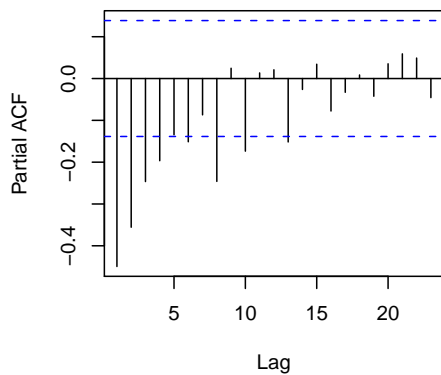
Plot of MA(1) Process



Sample Autocorrelation Function for MA(1) Process



Sample Partial Autocorrelation Function for MA(1) Process



AR/MA

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

Looking at the sample ACF, we can see a strong spike at lag  $k = 1$  followed by what seems to be white noise. This implies that this data may belong to an **MA(1) process**. The sample PACF has a rather strong decaying trend, which further supports our **MA(1)** conjecture. The sample EACF agrees with this conclusion. However, the **MA(1) process** is seventh on the BIC chart. Since we know that the true model is an **MA(1) process**, we are seeing the effects of random chance and imperfect diagnostic tools. Therefore, using the model identification tools above, we can properly conclude that the data belongs to an **MA(1) process**.

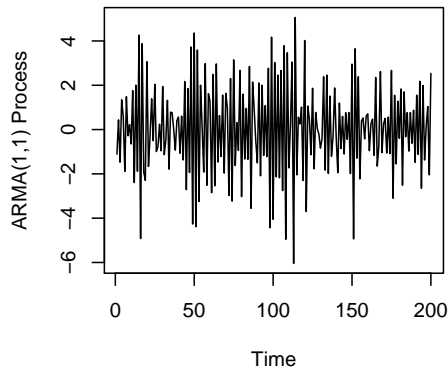
## (iii) ARMA(1,1) process

The code used to generate the output is as follows:

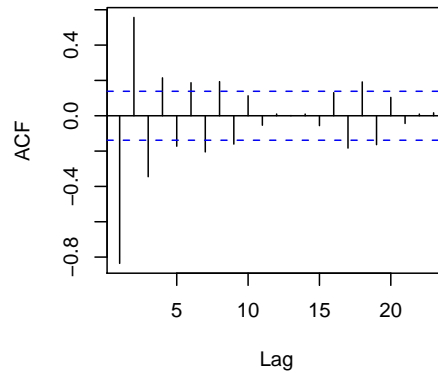
```
> arma11 <- arima.sim(list(ar = -.6, ma = -.8), n = 200)
> sub <- armasubsets(arma11, nar = 5, nma = 5)
>
> par(mfrow = c(2,2))
>
> plot.ts(arma11, main = "Plot of ARMA(1,1) Process", xlab = "Time",
+ ylab = "ARMA(1,1) Process")
> acf(arma11, main = "Sample Autocorrelation Function
+ for ARMA(1) Process")
> pacf(arma11, main = "Sample Partial Autocorrelation
+ Function for ARMA(1,1) Process")
> plot.armasubsets(sub)
>
> eacf(arma11)
```

The output generated is as follows:

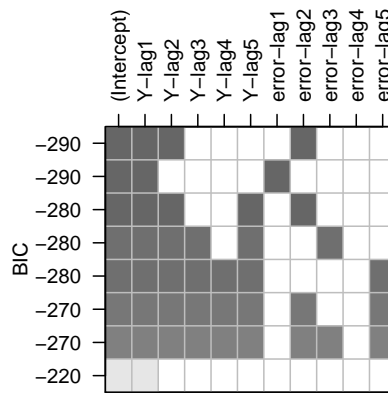
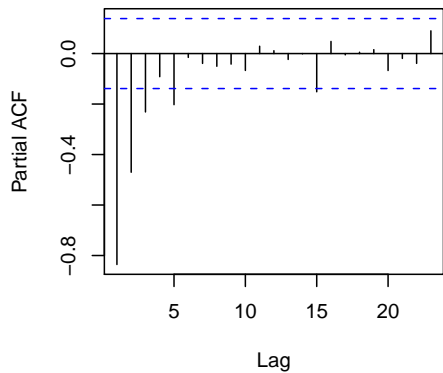
Plot of ARMA(1,1) Process



Sample Autocorrelation Function for ARMA(1) Process



Sample Partial Autocorrelation Function for ARMA(1,1) Process



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

Looking at the sample ACF, we can see a relatively strong decaying trend. This implies that this data may have an autoregressive component. The sample PACF has a rather strong decaying trend as well, which implies that the data may have a moving average component. The sample EACF implies that an **ARMA(1,1) process** may be an appropriate model for this data. The **ARMA(1,1)** and **ARMA(2,1) processes** are tied for first on the BIC chart. However, we prescribe to the school of parsimony. Therefore, using the model identification tools above, we can properly conclude that the data belongs to an **ARMA(1,1) process**.

4. I have put two new data sets on the course web site:

- **ibm**: daily closing IBM stock prices (dates not given)
- **internet**: number of users logged on to an Internet server each minute.

In addition, consider the data set

- **robot**: final horizontal position of an industrial robot put through a series of planned exercises

which is in the **TSA** package.

Using the methods from Chapter 6, identify a small set of candidate  $ARIMA(p, d, q)$  models for each data set. There may be a single model that emerges as a “clear favorite.” For guidance, use the summary described in Section 6.7 (notes) and follow it exactly. For each data set, write up detailed notes that describe how you decided on the model(s) you did. Your summary should convince me that your model(s) is (are) worthy of further consideration.

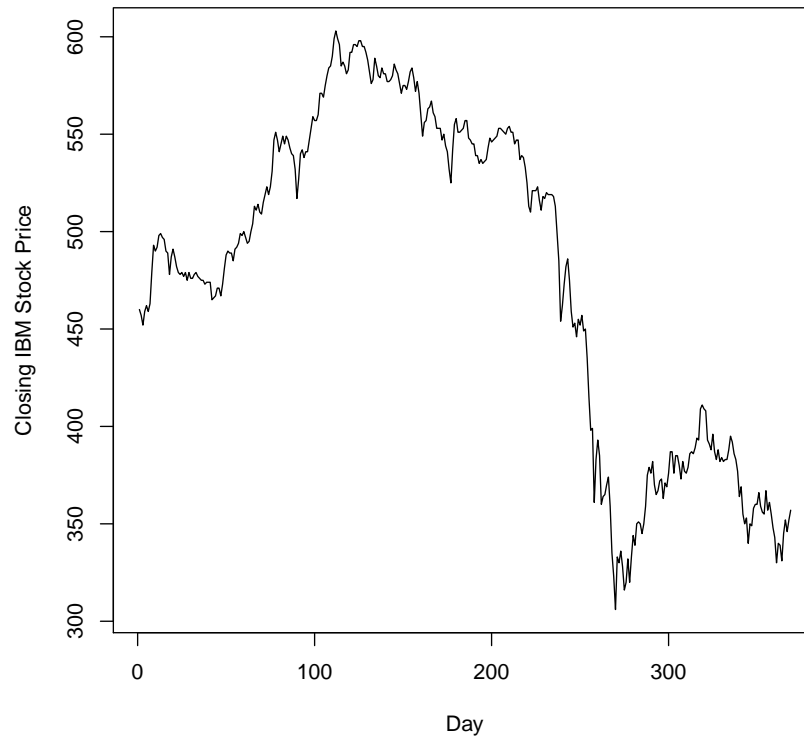
**Note:** Problem 4 is important because your class project will involve you finding data of your own, specifying a model (or models), fitting the model(s), and diagnosing model fit. This problem will help you in the model specification phase of your project.

In this section, any non-trivial code used to generate output will be listed.

Data: **ibm**

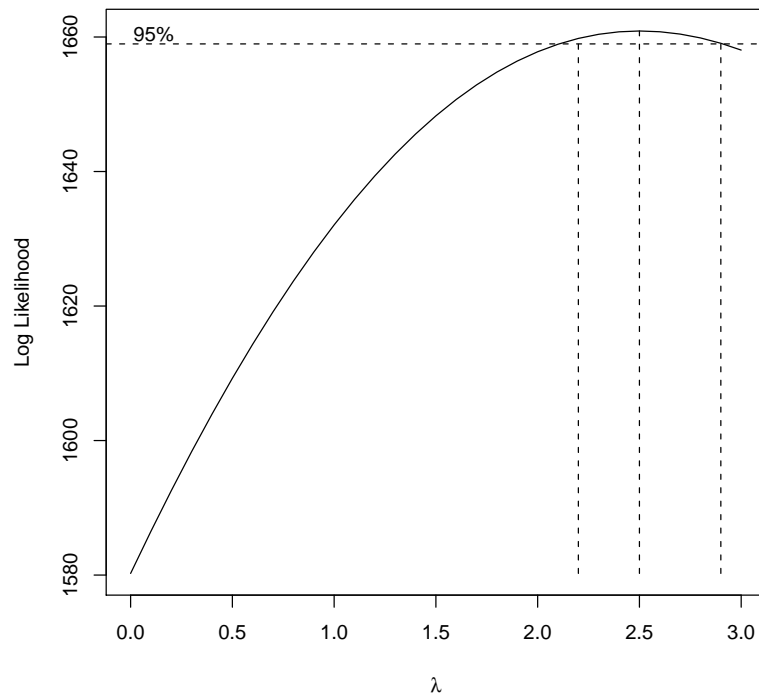
Before we attempt to model this data, we should see what it looks like.

Plot of Daily Closing IBM Stock Prices



As we can see, this series is definitely not stationary. Also, there doesn't seem to be any obvious transformation that would yield a stationary series. So, let's find an appropriate Box-Cox transformation (the appropriate  $\lambda$  values can be by trial-and-error).

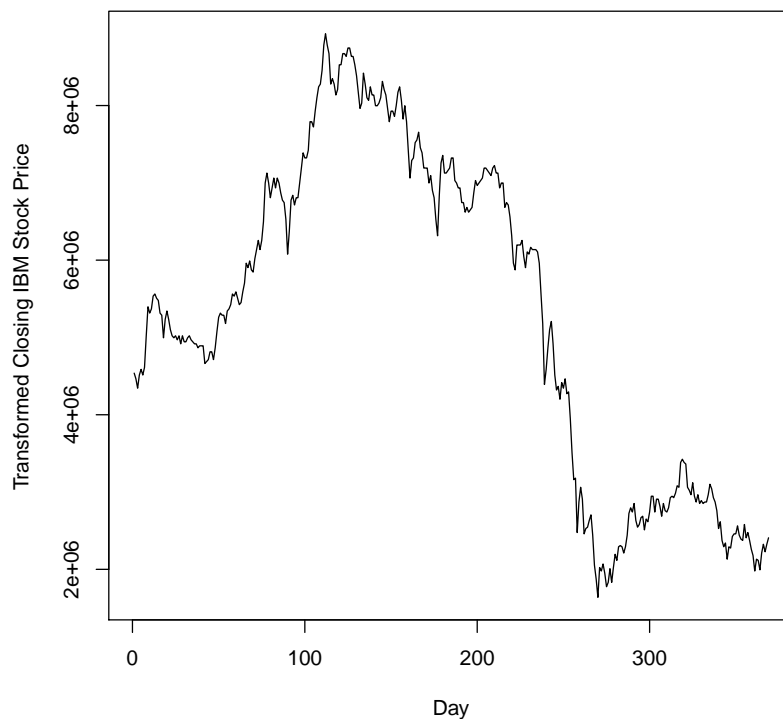
```
> BoxCox.ar(ibm, lambda = seq(0, 3, .1))
```



The Box-Cox technique implies that an appropriate power transformation has order 2.5. So, let's apply this transformation and see what the new series looks like.

```
> ibm2.5 <- ibm^(2.5)
```

Plot of Transformed Daily Closing IBM Stock Prices



This plot looks very similar to the original plot. It is not uncommon for the Box-Cox technique to yield a useless transformation. So, for simplicity's sake, let's use the original data. Now, let's use the Augmented Dickey-Fuller Unit Root (ADF) Test to see if differencing is warranted.

```
> ar(diff(ibm))
```

```
Call:
```

```
ar(x = diff(ibm))
```

```
Coefficients:
```

```
      1
0.0856
```

```
Order selected 1  sigma^2 estimated as  52.44
```

```
>
```

```
> ADF.test(ibm, selectlags = list(1), itsd = c(1,0,0))
```

```
----- - -----
Augmented Dickey & Fuller test
----- - -----
```

```
Null hypothesis: Unit root.
```

```
Alternative hypothesis: Stationarity.
```

-----

ADF statistic:

	Estimate	Std. Error	t value	Pr(> t )
adf.reg	-0.002	0.005	-0.343	0.1

Lag orders: 1

Number of available observations: 367

Warning message:

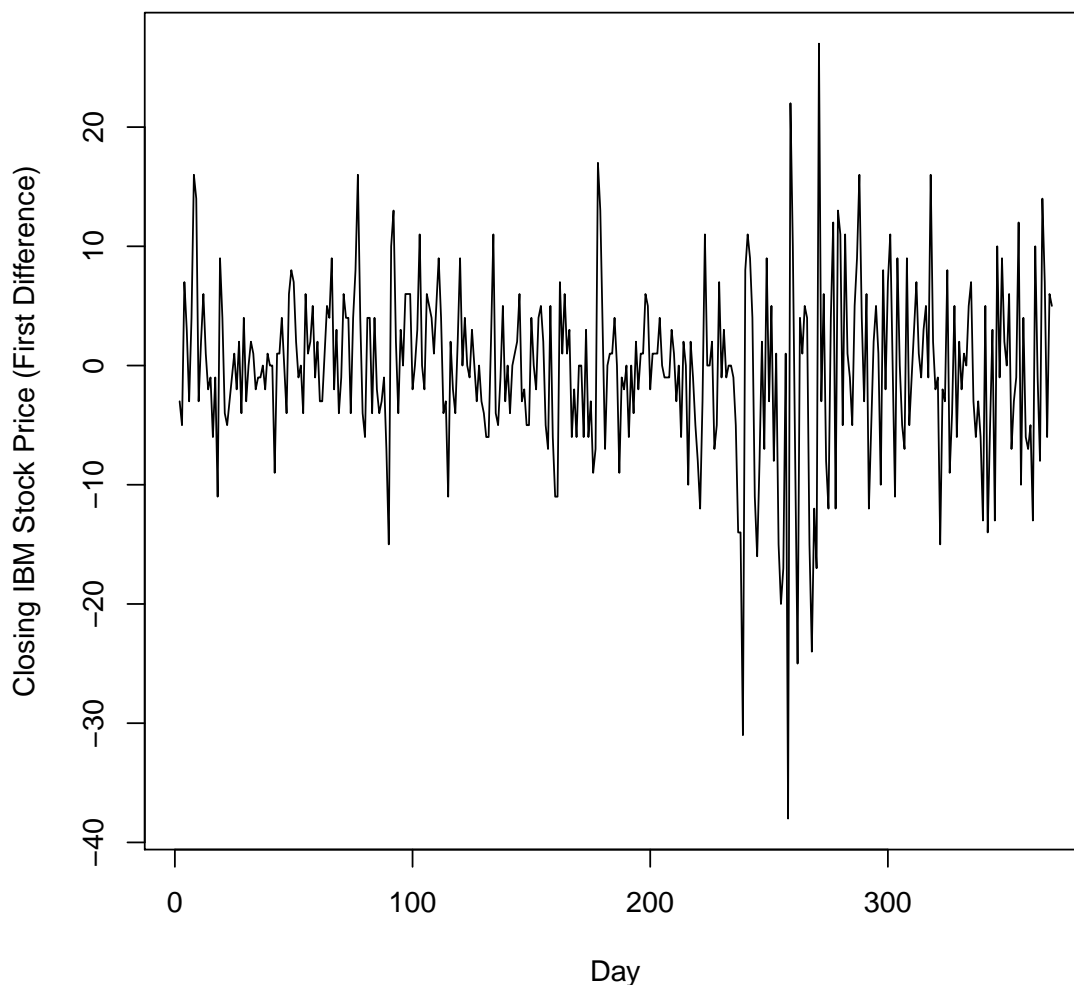
In interpolpval(code = code, stat = adfreg[, 3], N = N) :

p-value is greater than printed p-value

With a  $P$ -value greater than .1, we do not have sufficient evidence at the  $\alpha = .05$  level to say that this series is stationary. So, let's take the first difference and see what the new series looks like.

```
> ibm.diff <- diff(ibm)
```

### Plot of Daily Closing IBM Stock Prices (First Difference)

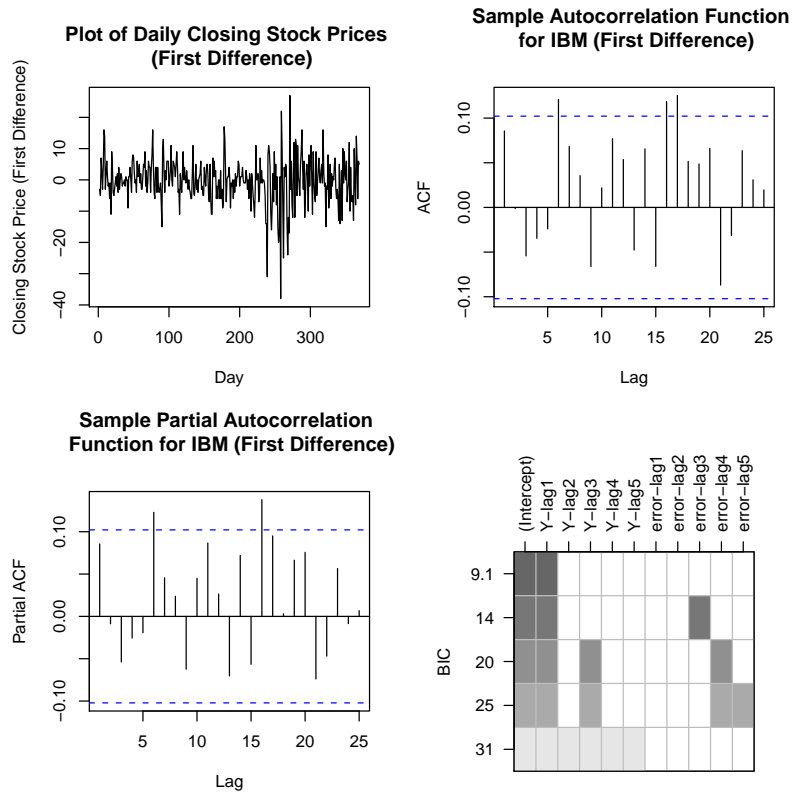




```

>
> par(mfrow = c(2,2))
>
> plot.ts(ibm.diff, main = "Plot of Daily Closing IBM Stock Prices
+ (First Difference)", xlab = "Day", ylab =
+ "Closing IBM Stock Price (First Difference)")
> acf(ibm.diff, main = "Sample Autocorrelation Function
+ for IBM (First Difference)")
> pacf(ibm, main = "Sample Partial Autocorrelation
+ Function for IBM (First Difference)")
> plot.armsubsets(sub)

```



```

> eacf(ibm.diff)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 o o o o o x o o o o o o o
1 o o o o o x o o o o o o o
2 x x o o o x o o o o o o o
3 x x o o o o o o o o o o o
4 x o x o o o o o o o o o o
5 x x x x x o o o o o o o o
6 x o x x o x o o o o o o o
7 x o x x x x o o o o o o o

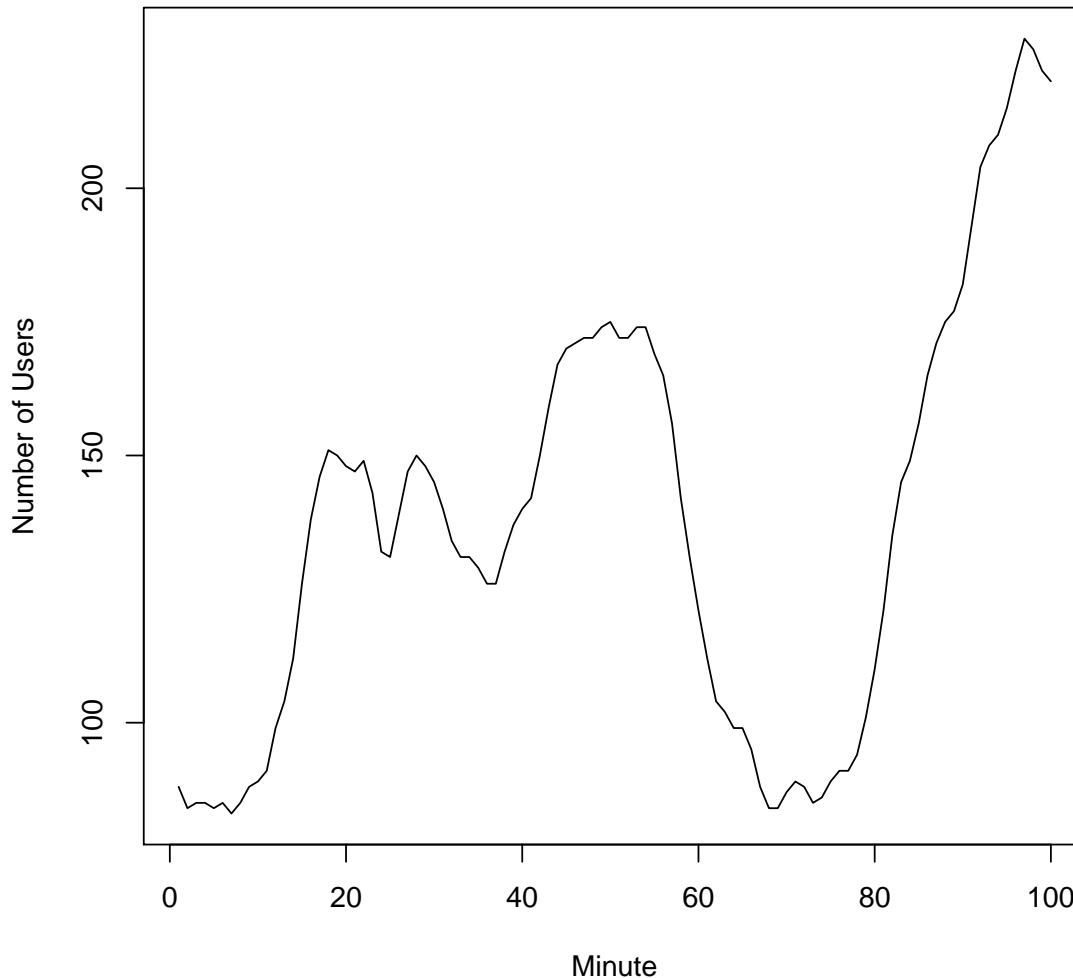
```

The sample ACF and PACF both appear to be reflective of a white noise process. This conclusion is supported by the sample EACF. However, the white noise model is completely absent from the BIC plot. The BIC plot suggests that an **AR(1) process** appropriately models the differenced data. But, we would be very wary of an **AR(1)** model with  $\phi < .1$ . Therefore, we feel most comfortable modeling these first differences as white noise, i.e. modeling the original process as a random walk.

Data: **internet**

Before we attempt to model this data, we should see what it looks like.

### Plot of Number of Internet Users Logged On Each Minute



As we can see, this series is definitely not stationary. Also, there doesn't seem to be any obvious transformation that would yield a stationary series. So, let's find an appropriate Box-Cox transformation (the appropriate  $\lambda$  values can be found by trial-and-error).

```
> BoxCox.ar(internet, lambda = seq(0, 3, .1))
```

The BoxCox.ar() function in R returns an error for this series. It's origin is unknown at this time. So, let's use the Augmented Dickey-Fuller Unit Root (ADF) Test to see if differencing is warranted.

```
> ar(diff(internet))
```

Call:

```
ar(x = diff(internet))
```

Coefficients:

```
      1      2      3
1.1060 -0.5957  0.3029
```

Order selected 3 sigma^2 estimated as 10.32

```
>
```

```
> ADF.test(internet, selectlags = as.list(1:3), itsd = c(1,0,0))
```

```
-----
Augmented Dickey & Fuller test
-----
```

Null hypothesis: Unit root.

Alternative hypothesis: Stationarity.

```
----
```

ADF statistic:

	Estimate	Std. Error	t value	Pr(> t )
adf.reg	-0.011	0.009	-1.233	0.1

Lag orders: 1 2

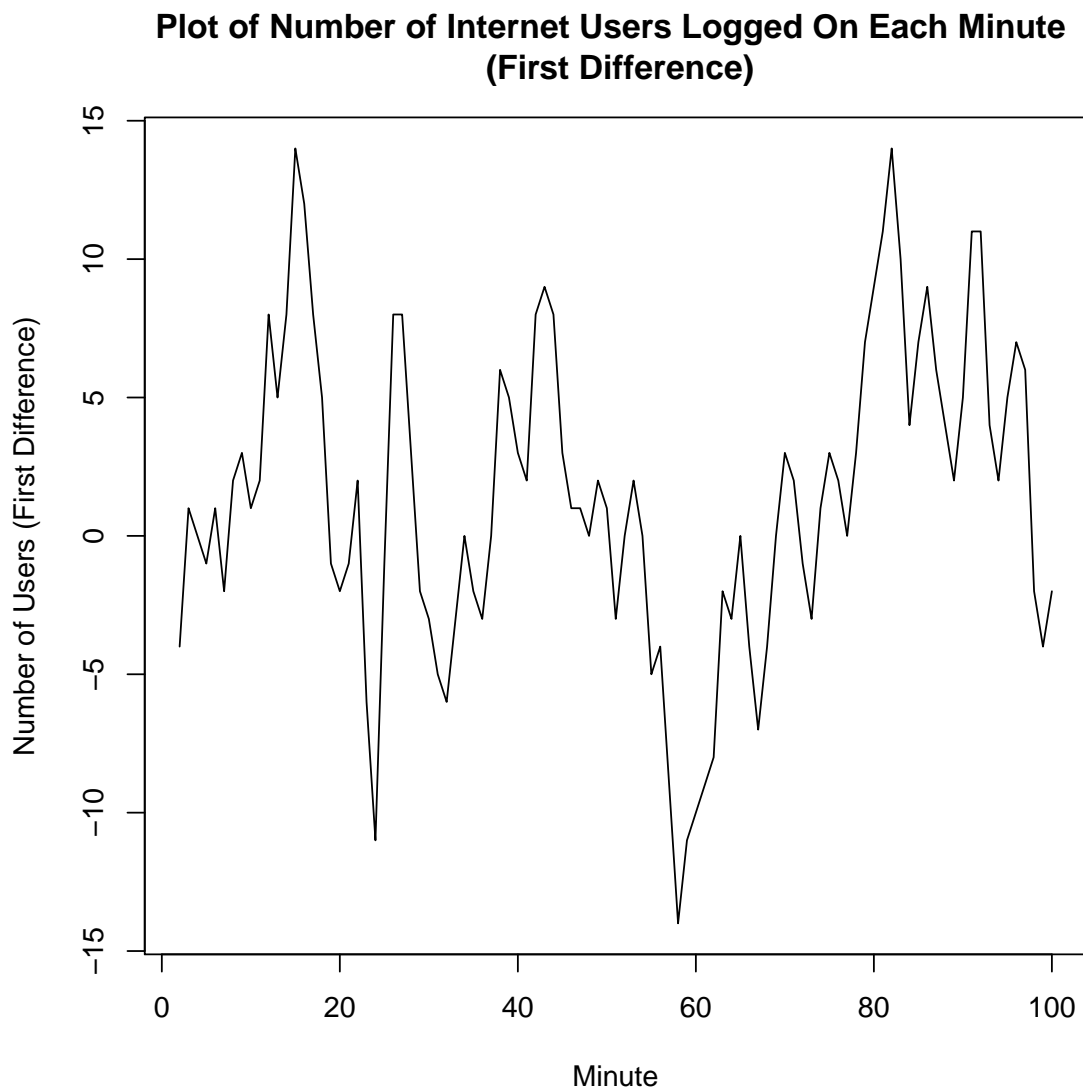
Number of available observations: 97

Warning message:

```
In interpval(code = code, stat = adfreg[, 3], N = N) :
  p-value is greater than printed p-value
```

With a  $P$ -value greater than .1, we do not have sufficient evidence at the  $\alpha = .05$  level to say that this series is stationary. So, let's take the first difference and see what the new series looks like.

```
> internet.diff <- diff(internet)
```



This series still seems to have some structure, which could imply non-stationarity. However, in this case, this is most likely due to the structure of the **ARMA(p,q) process**. However, out of curiosity, let's run the ADF test and see what it tells us.

```
> ar(diff(internet.diff))
```

Call:

```
ar(x = diff(internet.diff))
```

Coefficients:

1	2
0.2489	-0.4341

```
Order selected 2 sigma^2 estimated as 10.56
```

```

>
> ADF.test(internet.diff, selectlags = as.list(1:2), itsd = c(1,0,0))
-----
Augmented Dickey & Fuller test
-----

Null hypothesis: Unit root.
Alternative hypothesis: Stationarity.

-----

ADF statistic:

      Estimate Std. Error t value Pr(>|t|)
adf.reg  -0.175    0.064  -2.722   0.077

Lag orders: 1 2
Number of available observations: 96

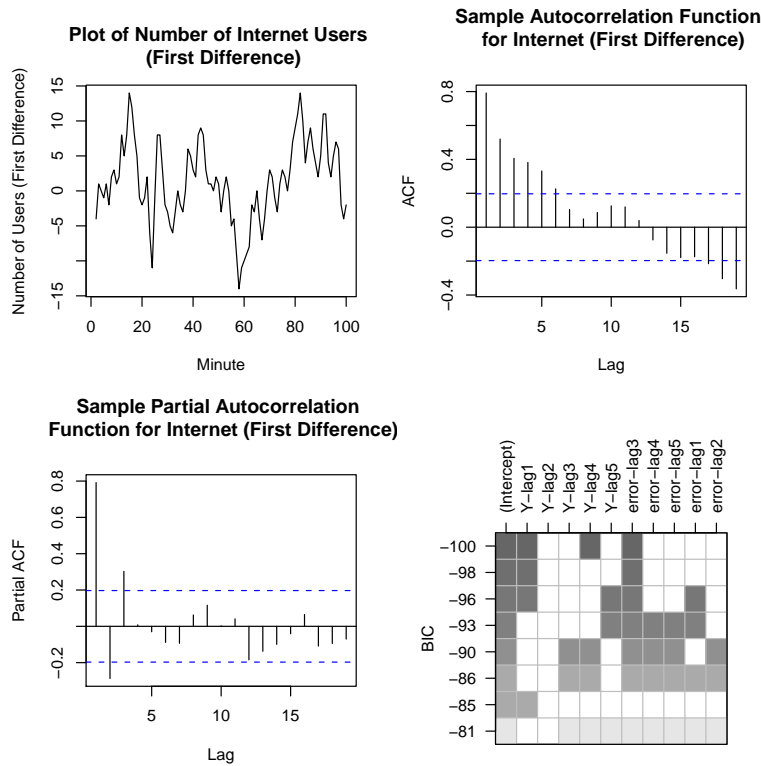
```

With a  $P$ -value less than .077, we have mild evidence at the  $\alpha = .05$  level to conclude that the series is stationary. However, our general rule is to require strong evidence of non-stationarity before we difference the data. So, we will not difference this data again. Now, let's move on to picking an **ARMA(p, q)** model for the differenced series. First, let's look at the sample ACF, sample PACF, sample EACF, and BIC plots.

```

> sub <- armasubsets(internet.diff, nar = 5, nma = 5)
>
> par(mfrow = c(2,2))
>
> plot.ts(internet.diff, main = "Plot of Number of Internet Users
+ (First Difference)", xlab = "Minute", ylab =
+ "Number of Users (First Difference)")
> acf(internet.diff, main = "Sample Autocorrelation Function
+ for Internet (First Difference)")
> pacf(internet.diff, main = "Sample Partial Autocorrelation
+ Function for Internet (First Difference)")
> plot.armasubsets(sub)

```

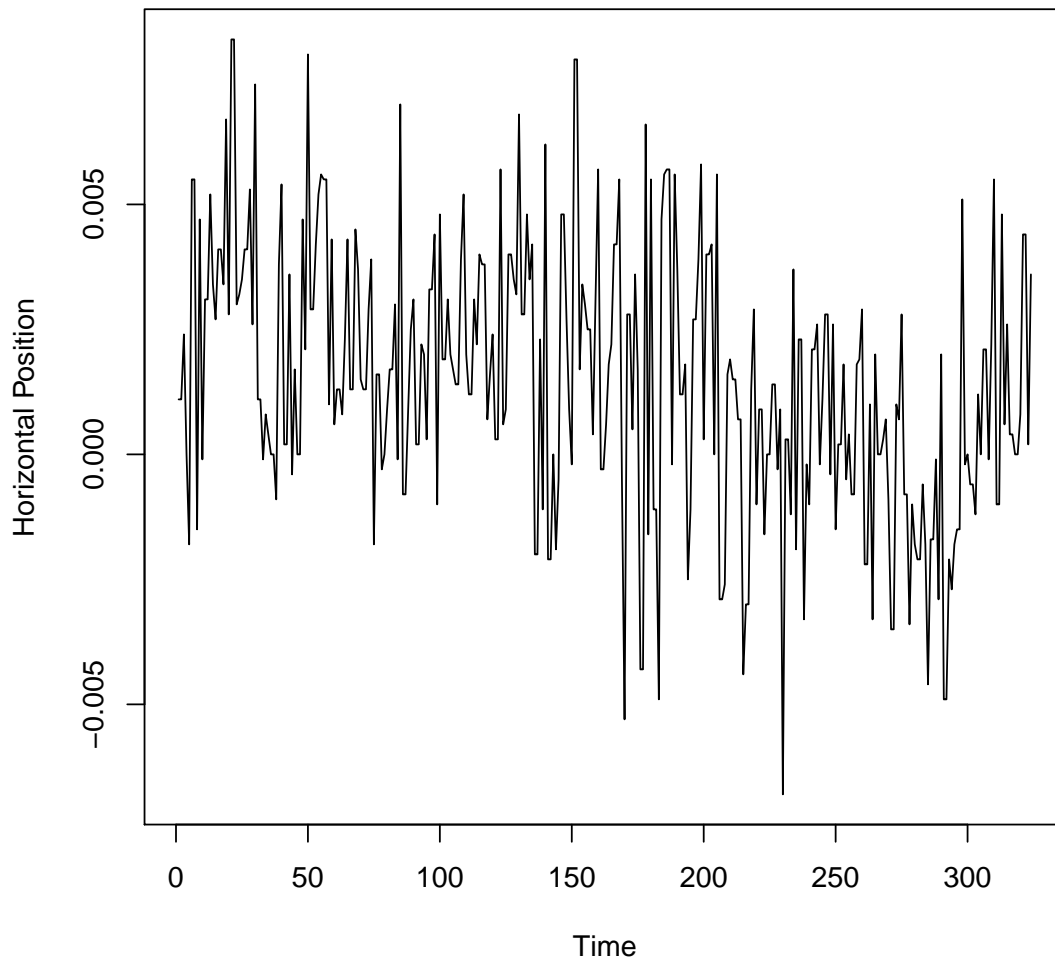


```
> eacf(internet.diff)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x o o o o o o o
1 x x o o o o o o o o o o o
2 x x x o o o o o o o o o o
3 o o o o o o o o o o o o o
4 x o o x o o o o o o o o o
5 x o o x o o o o o o o o o
6 x o o x o o o o o o o o o
7 x x x x o o o o o o o o o
```

The sample ACF seems to have some sort of decaying trend to it. This implies that there may be an autoregressive component to this model. The sample PACF has a strong spike at lags  $k = 1, 2,$  and  $3,$  and insignificant values after. This implies that this series could have up to 3 autoregressive components. The sample EACF suggests that good models for the first differences are **ARMA(1,2)** and **AR(3)** processes. The BIC plot does not have any of these models on it. So, we would say that the best model for this data is an **ARI(3,1)** process, with an **ARIMA(1,1,2)** process placing in a distant second.

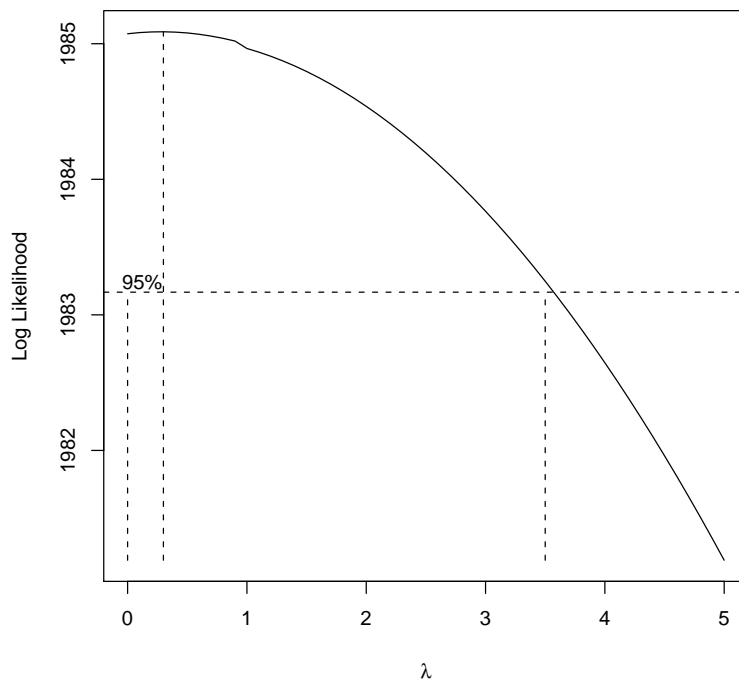
Data: robot

Before we attempt to model this data, we should see what it looks like.

**Plot of Horizontal Position of Robot**

This series seems to exhibit a mild trend in its mean function. However, we are not sure whether this is enough to call the series non-stationary. So, let's find an appropriate Box-Cox transformation (the appropriate  $\lambda$  values can be by trial-and-error).

```
> BoxCox.ar(robot + .1, lambda = seq(0, 5, .1))
```



Since 1 is in the 95% confidence interval, we do not have sufficient evidence that the identity transformation is inappropriate. Now, let's use the ADF test to check for non-stationarity.

```
> ADF.test(robot, itsd = c(1,0,0))
```

```
-----
Augmented Dickey & Fuller test
-----
```

```
Null hypothesis: Unit root.
Alternative hypothesis: Stationarity.
```

```
----
```

```
ADF statistic:
```

	Estimate	Std. Error	t value	Pr(> t )
adf.reg	-0.296	0.084	-3.523	0.01

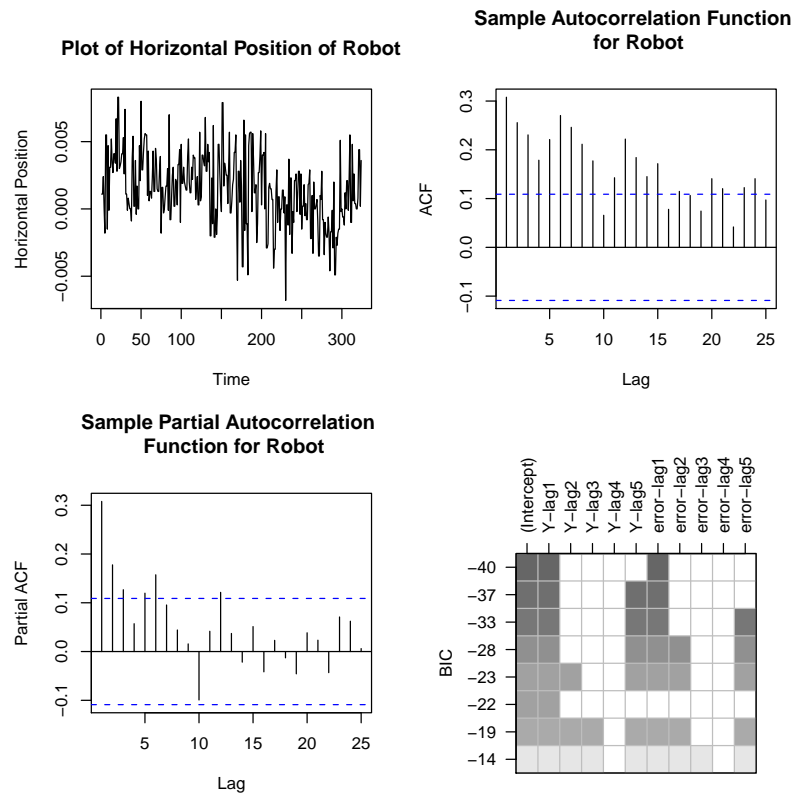
```
Lag orders: 1 2 3 4 5 6 10 11
Number of available observations: 312
```

```
Warning message:
```

```
In interpval(code = code, stat = adfreg[, 3], N = N) :
p-value is smaller than printed p-value
```

For some reason, the `ADF.test()` function in R would not accept values for the `selectlags` parameter. So, we used the default value. With a  $P$ -value less than .01, we have sufficient evidence at the  $\alpha = .05$  level to say that this series is stationary. Now, let's move on to picking an **ARMA(p, q)** model for the differenced series. First, let's look at the sample ACF, sample PACF, sample EACF, and BIC plots.

```
> sub <- armasubsets(robot, nar = 5, nma = 5)
>
> par(mfrow = c(2,2))
>
> plot.ts(robot, main = "Plot of Horizontal Position of Robot",
+ xlab = "Time", ylab = "Horizontal Position")
> acf(robot, main = "Sample Autocorrelation Function
+ for Robot")
> pacf(robot, main = "Sample Partial Autocorrelation
+ Function for Robot")
> plot.armasubsets(sub)
```



```

> eacf(robot)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x x x x x x x x x o x x x x
1 x o o o o o o o o o o o o o
2 x x o o o o o o o o o o o o
3 x x o o o o o o o o o o o o
4 x x x x o o o o o o o o x o
5 x x x o o o o o o o o o x o
6 x o o o o x o o o o o o o o
7 x o o x o x x o o o o o o o

```

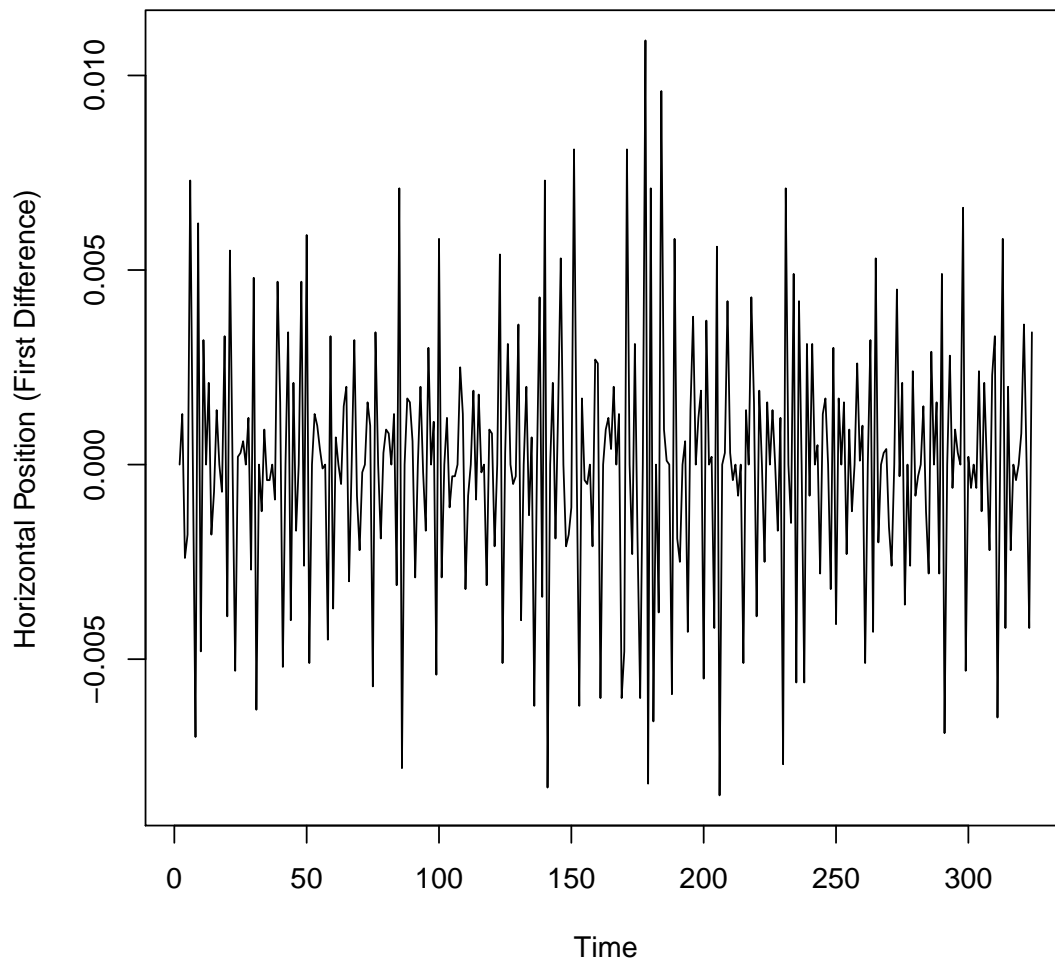
There is a significant decaying trend in both the sample ACF and the sample PACF. This indicates that there could be an autoregressive component and a moving average component. The sample EACF and the BIC plot both agree that this series is best modeled by an **ARIMA(1,0,1) process**. However, we aren't convinced that this is the only route to go down. There seems to be what could be some kind of trend in the mean function. So, let's take the first differences and see what we get.

```

> robot.diff <- diff(robot)

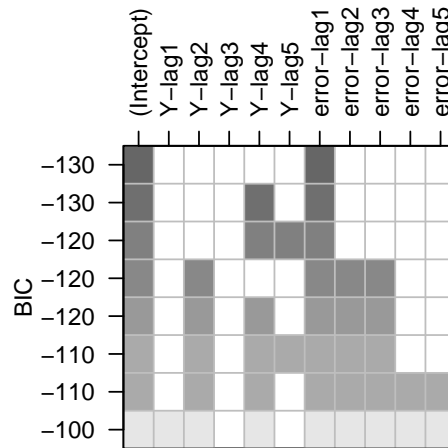
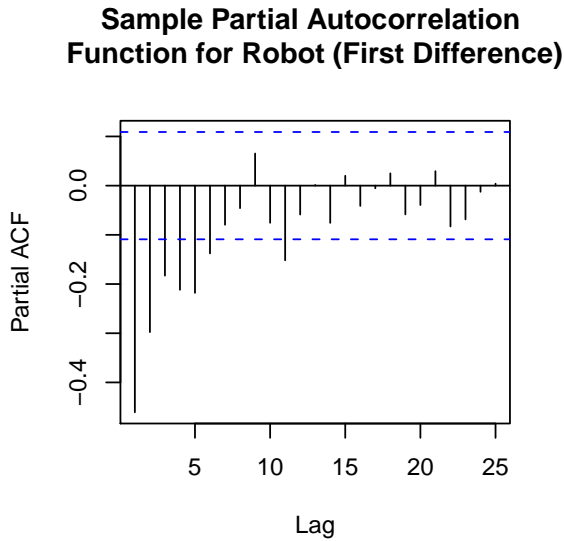
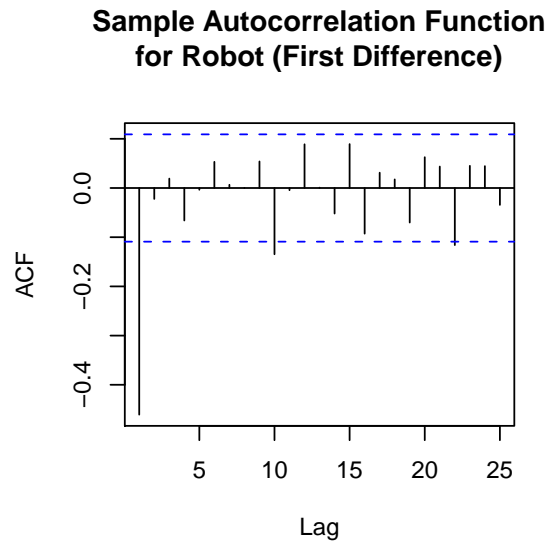
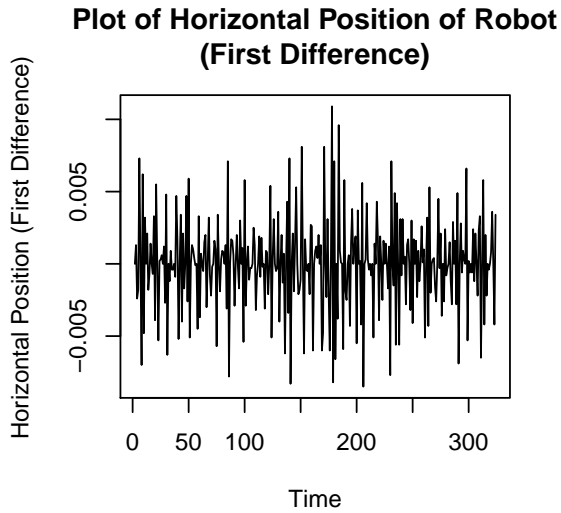
```

**Plot of Horizontal Position of Robot  
(First Difference)**



The differenced data looks to substantially more stationary. In fact, we think it would be unreasonable, and possibly misleading, to conduct the ADF test on this data. So, let's look at the ACF, PACF, EACF, and BIC plot to find a set of candidate models for the first differences.

```
> par(mfrow = c(2,2))
>
> plot.ts(robot.diff, main = "Plot of Horizontal Position of Robot
+ (First Difference)", xlab = "Time", ylab =
+ "Horizontal Position (First Difference)")
> acf(robot.diff, main = "Sample Autocorrelation Function
+ for Robot (First Difference)")
> pacf(robot.diff, main = "Sample Partial Autocorrelation
+ Function for Robot (First Difference)")
> plot.armsubsets(sub)
```



```
> eacf(robot.diff)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o o o o o o o o x o o o o
1 x x o o o o o o o x o o o o
2 x x x o o o o o o o o o o o
3 x x x x o o o o o o o o o o
4 x x x o o o o o o o o o o o
5 x o o o o x o o o o o o o o
6 x x o x o x x o o o o o o o
7 x o o o o x x o o o o o o o
```

Every single one of these plots tells us the exact same thing. The **MA(1) process** has arisen as a clear candidate to model the first differences. In fact, we think that this model is substantially better than our original **ARIMA(1,0,1)** decision.