

In HW 4, you were asked to identify a small set of candidate $ARIMA(p, d, q)$ models (perhaps some of you even made a “final choice”) for each of the following data sets:

- **ibm**: daily closing IBM stock prices (dates not given)
- **internet**: number of users logged on to an Internet server each minute.
- **robot**: final horizontal position of an industrial robot put through a series of planned exercises.

Remembering your candidate models for each data set, fit and diagnose your model selections. That is, use the methods from Chapter 7 to fit your chosen models (for uniformity, you could just use maximum likelihood for each model fit). Then, diagnose your fitted model(s) by doing a thorough analysis of the residuals and implementing the overfitting technique (Chapter 8).

For each data set and for each model you entertained in HW 4, what do you think now? Would you like to suggest another model for further investigation? Or, are you satisfied with your HW 4 model selections?

- If your original model choices in HW 4 are “reasonable,” convince me that they are.
- If your original model choices are deemed “not reasonable,” use the information from your diagnoses to specify another model. Then, evaluate the merit of this new model using the methods from Chapter 7 and Chapter 8.

Your goal is to come up with **one final model** for each data set—the “best” one. Convince me that your final model does a good job at explaining the variability in the data, but also adhere to the Principle of Parsimony. There are no “right” answers here, but there are certainly bad answers (stay away from these).

Important: Remember your final model for each data set.

Data: **ibm**

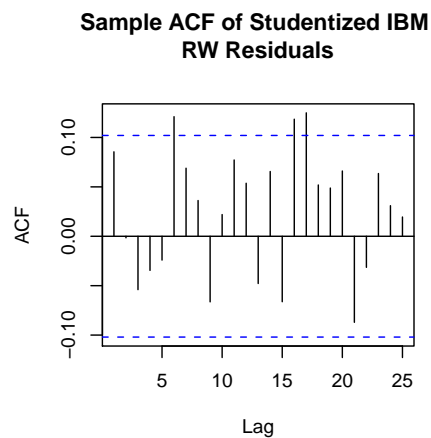
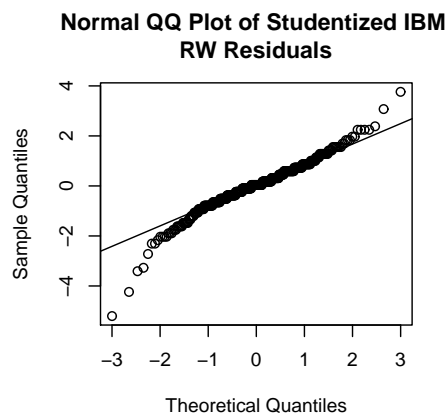
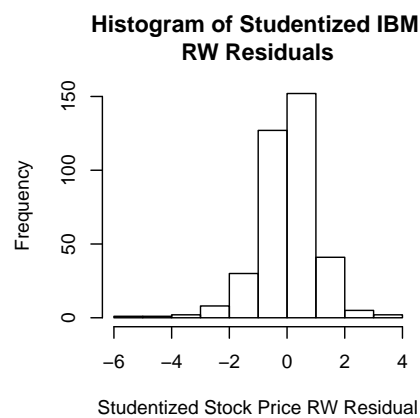
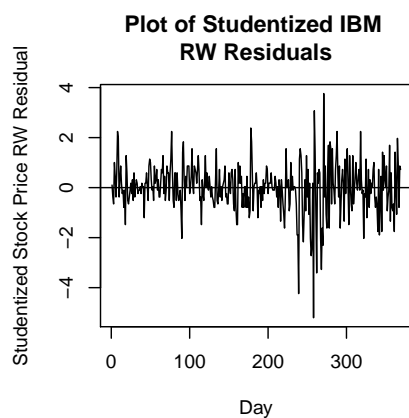
We chose to model this series as a random walk. Because of this choice, we have only one parameter to estimate, the white noise variance. So, let’s check the residuals to see if they resemble white noise

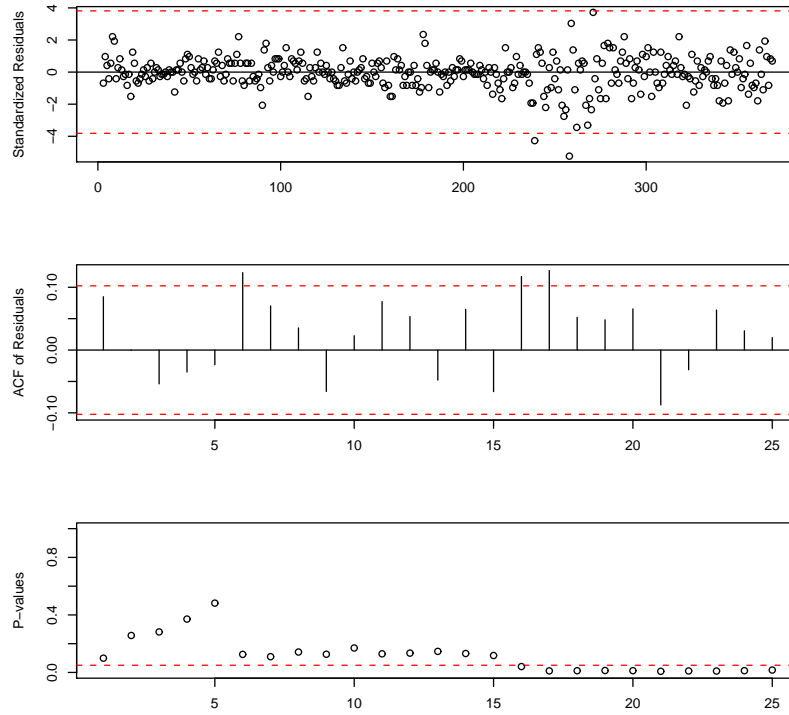
```
> ibm.rw <- arima(ibm, order = c(0,1,0), method = "ML")
> ibm.rw
> ibm.rw.res <- ibm.rw$residuals
> ibm.rw.stud.res <- (ibm.rw.res - mean(ibm.rw.res)) / sd(ibm.rw.res)
>
> par(mfrow = c(2,2))
```

```

>
> plot.ts(ibm.rw.stud.res, main = "Plot of Studentized IBM
+ RW Residuals", xlab = "Day", ylab =
+ "Studentized Stock Price RW Residual")
> abline(h = 0)
> hist(ibm.rw.stud.res, main = "Histogram of Studentized IBM
+ RW Residuals", xlab = "Studentized Stock Price RW Residual")
> qqnorm(ibm.rw.stud.res, main = "Normal QQ Plot of Studentized IBM
+ RW Residuals")
> qqline(ibm.rw.stud.res)
> acf(ibm.rw.stud.res, main = "Sample ACF of Studentized IBM
+ RW Residuals")
>
> par(mfrow = c(1,1))
>
> tsdiag(ibm.rw)

```





```
> shapiro.test(ibm.rw.stud.res)
```

Shapiro-Wilk normality test

```
data: ibm.rw.stud.res  
W = 0.9582, p-value = 9.527e-09
```

```
> runs(ibm.rw.stud.res)
```

```
$pvalue  
[1] 0.777
```

```
$observed.runs  
[1] 181
```

```
$expected.runs  
[1] 184.1978
```

```
$n1  
[1] 169
```

```
$n2  
[1] 200
```

```
$k
[1] 0
```

It is obvious by the histogram, QQ plot, and the Shapiro-Wilk Test for Normality that these standardized residuals are not normally distributed. Therefore, the random walk process is not a very good model for this data. Now, let's try the **ARI(1,1) process** that we mentioned in Homework 4. We will use Maximum Likelihood Estimation to estimate the autoregressive coefficient ϕ .

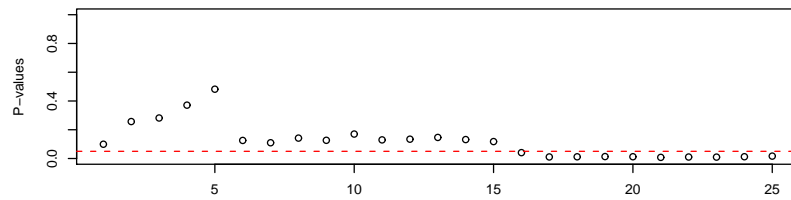
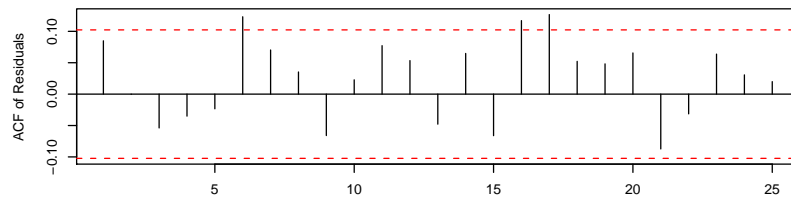
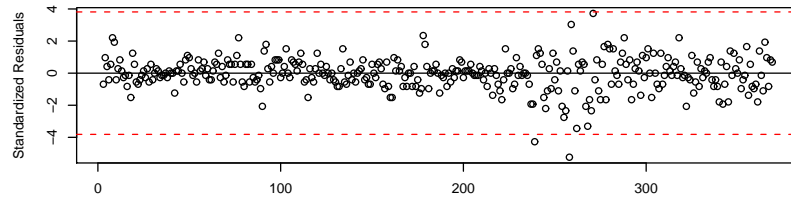
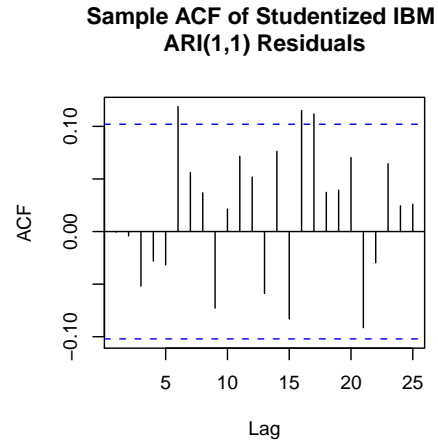
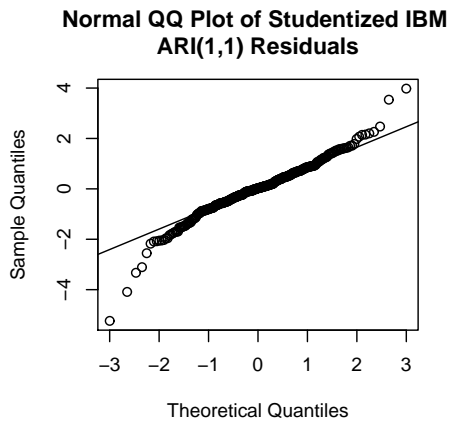
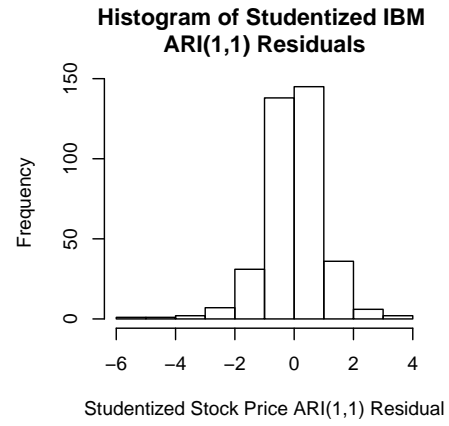
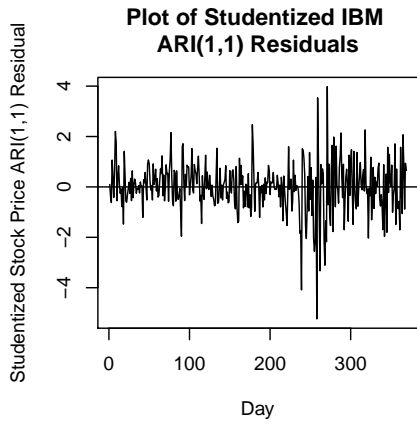
```
> ibm.ari1.1 <- arima(ibm, order = c(1,1,0), method = "ML")
> ibm.ari1.1
```

```
Call:
arima(x = ibm, order = c(1, 1, 0), method = "ML")
```

```
Coefficients:
          ar1
          0.0869
s.e.      0.0519
```

```
sigma^2 estimated as 52.22:  log likelihood = -1249.97,  aic = 2501.94
```

```
>
> ibm.ari1.1.res <- ibm.ari1.1$residuals
> ibm.ari1.1.stud.res <- (ibm.ari1.1.res - mean(ibm.ari1.1.res)) / sd(ibm.ari1.1.res)
>
> par(mfrow = c(2,2))
>
> plot.ts(ibm.ari1.1.stud.res, main = "Plot of Studentized IBM
+ ARI(1,1) Residuals", xlab = "Day", ylab =
+ "Studentized Stock Price ARI(1,1) Residual")
> abline(h = 0)
> hist(ibm.ari1.1.stud.res, main = "Histogram of Studentized IBM
+ ARI(1,1) Residuals", xlab = "Studentized Stock Price ARI(1,1) Residual")
> qqnorm(ibm.ari1.1.stud.res, main = "Normal QQ Plot of Studentized IBM
+ ARI(1,1) Residuals")
> qqline(ibm.ari1.1.stud.res)
> acf(ibm.ari1.1.stud.res, main = "Sample ACF of Studentized IBM
+ ARI(1,1) Residuals")
>
> par(mfrow = c(1,1))
>
> tsdiag(ibm.ari1.1)
```



```
> shapiro.test(ibm.ari1.1.stud.res)
```

Shapiro-Wilk normality test

```
data:  ibm.ari1.1.stud.res
W = 0.9612, p-value = 2.626e-08
```

```
> runs(ibm.ari1.1.stud.res)
$pvalue
[1] 0.746
```

```
$observed.runs
[1] 189
```

```
$expected.runs
[1] 185.3902
```

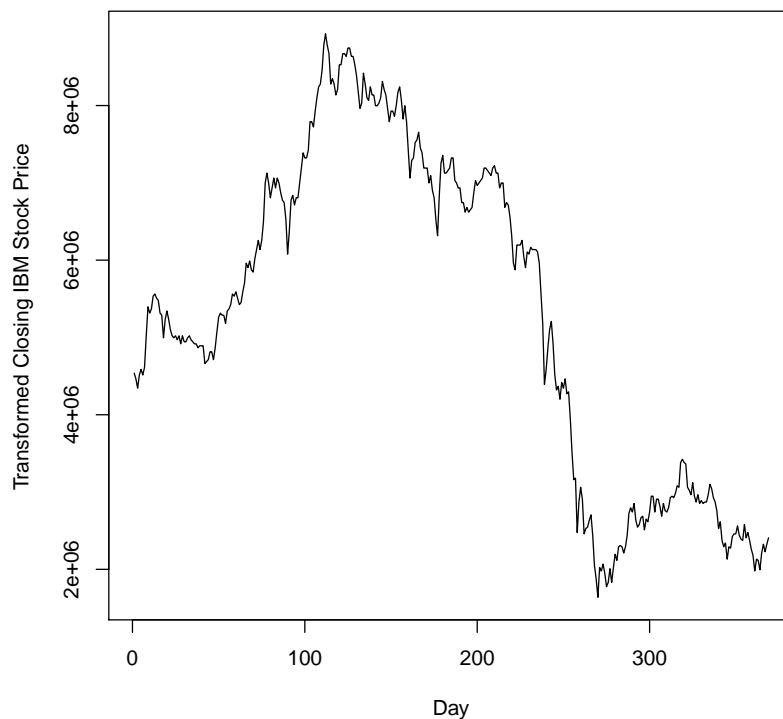
```
$n1
[1] 180
```

```
$n2
[1] 189
```

```
$k
[1] 0
```

These residuals look very similar to the ones obtained from the random walk process. They are definitely not zero mean normal white noise. Also, the ϕ parameter is not significantly different from zero. So, it seems that we can't find a suitable model in our candidate set. Now, let's use the $\lambda = 2.5$ power transformation that we threw away on Homework 4.

Plot of Transformed Daily Closing IBM Stock Prices



It seems very likely to us that this series is not stationary. Let's run the ADF test to be sure though.

```
> ADF.test(ibm2.5, 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	-0.004	0.004	-0.939	0.1

```
Lag orders: 1 3 14 16 17 20
```

```
Number of available observations: 348
```

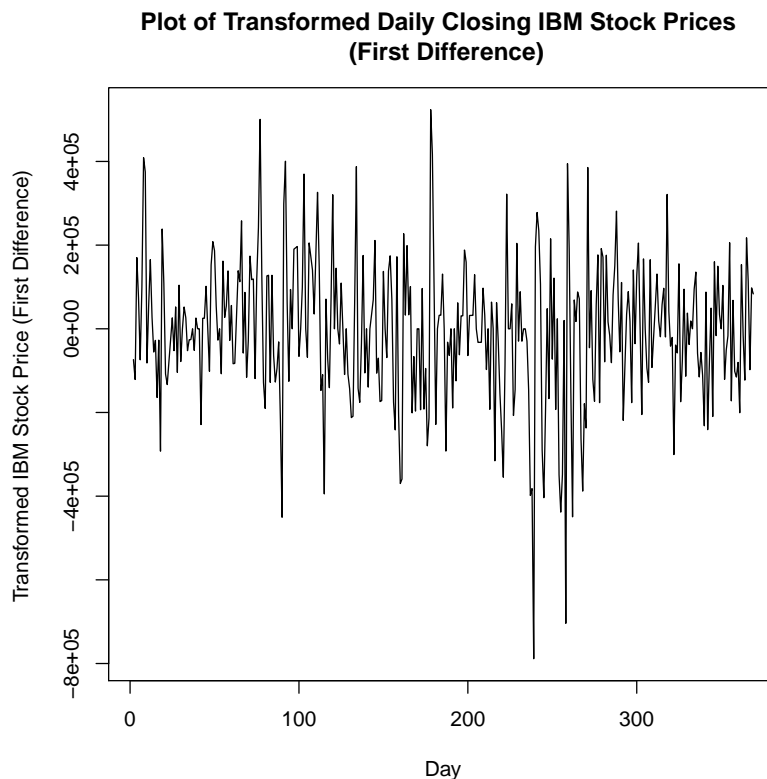
```
Warning messages:
```

```
1: In summary(lmref) : bytecode version mismatch; using eval
```

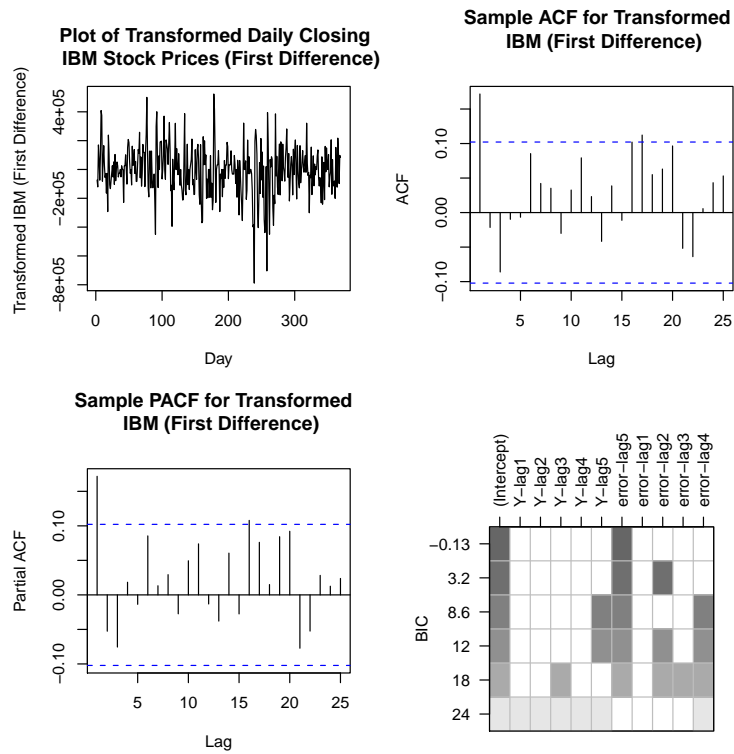
```
2: In interpolpval(code = code, stat = adfreg[, 3], N = N) :
p-value is greater than printed p-value
```

With a p-value greater than .1, we do not have sufficient evidence to reject the null hypothesis. Therefore, we can conclude that the series could be non-stationary. So, let's take the first difference.

```
> ibm2.5.diff <- diff(ibm2.5)
```



This series looks much more stationary. We feel it would not be appropriate to run the ADF test on this data because there is no obvious trend in the mean or variance over time. So, let's look at the ACF, PACF, EACF, and BIC plots.



```
> eacf(ibm2.5.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 o o o o o
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 o x o o o o o o o o o o o
4 x o 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 x x x o o o o o o o o o o
7 x x o x o x x o o o o o o o
```

The sample ACF and PACF look very clean. They each show a spike at lag 1 and insignificant values afterward. This can happen simultaneously because the values of r_1 and $\hat{\phi}_{11}$ are small. Looking at the BIC plot, we see that the top model for the transformed, differenced data is an **MA(1) process**. Looking at the EACF, we see that two reasonable models for this data are the **MA(1)** and **ARMA(1,1) processes**, which both agree with what the ACF and PACF tell us. Therefore, let's move forward using the **IMA(1,1) process** to model the transformed data and keep the **ARIMA(1,1,1) process** in second just in case there is something wrong with our **IMA(1,1) model**.

```
> ibm2.5.ima1.1 <- arima(ibm2.5, order = c(0,1,1), method = "ML")
> ibm2.5.ima1.1
```

```
Call:
arima(x = ibm2.5, order = c(0, 1, 1), method = "ML")

Coefficients:
      ma1
      0.1771
s.e.    0.0500

sigma^2 estimated as 2.840e+10:  log likelihood = -4950.97,  aic = 9903.95

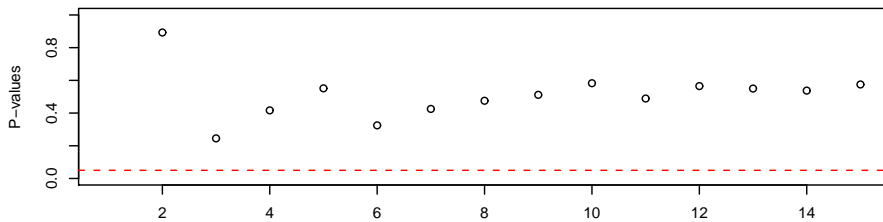
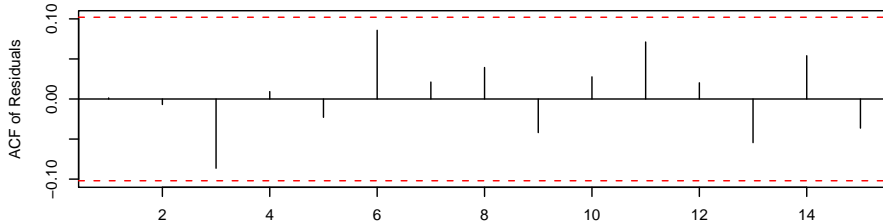
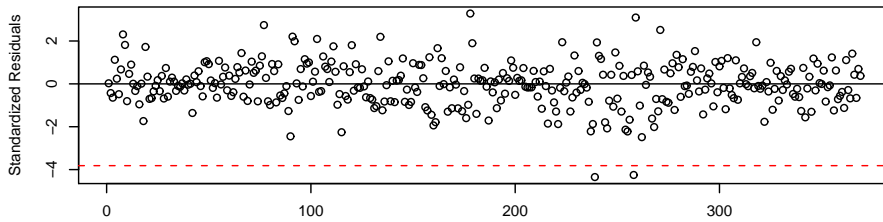
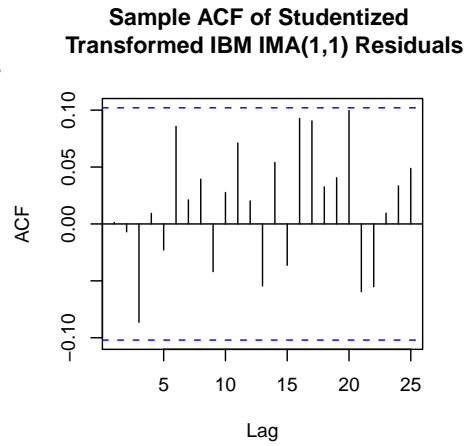
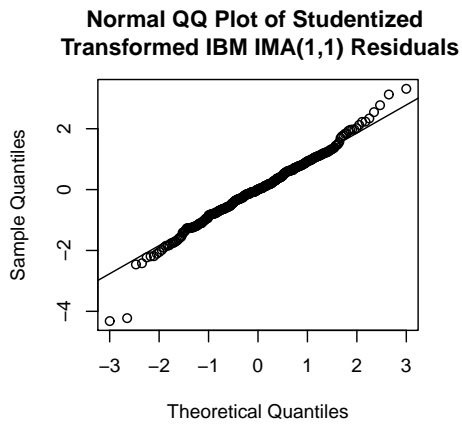
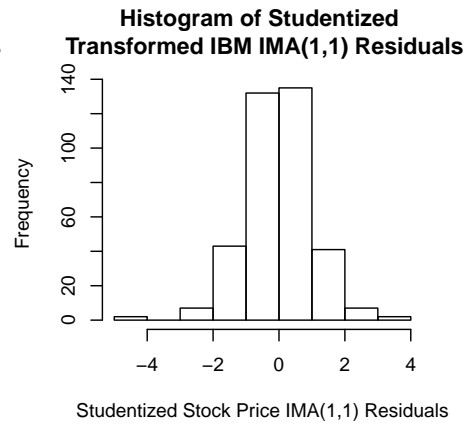
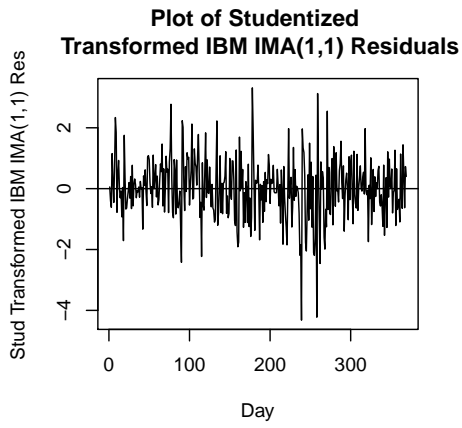
> ibm2.5.ima1.1.res <- ibm2.5.ima1.1$residuals
> ibm2.5.ima1.1.stud.res <- (ibm2.5.ima1.1.res - mean(ibm2.5.ima1.1.res)) /
+ sd(ibm2.5.ima1.1.res)

> par(mfrow = c(2,2))

> plot.ts(ibm2.5.ima1.1.stud.res, main = "Plot of Studentized
+ Transformed IBM IMA(1,1) Residuals", xlab = "Day", ylab =
+ "Studentized Transformed IBM IMA(1,1) Residuals")
> abline(h = 0)
> hist(ibm2.5.ima1.1.stud.res, main = "Histogram of Studentized
+ Transformed IBM IMA(1,1) Residuals", xlab =
+ "Studentized Stock Price IMA(1,1) Residuals")
> qqnorm(ibm2.5.ima1.1.stud.res, main = "Normal QQ Plot of Studentized
+ Transformed IBM IMA(1,1) Residuals")
> qqline(ibm2.5.ima1.1.stud.res)
> acf(ibm2.5.ima1.1.stud.res, main = "Sample ACF of Studentized
+ Transformed IBM IMA(1,1) Residuals")

> par(mfrow = c(1,1))

> tsdiag(ibm2.5.ima1.1, gof = 15, omit.initial = F)
```



```
> shapiro.test(ibm2.5.ima1.1.stud.res)
```

```
Shapiro-Wilk normality test
```

```
data:  ibm2.5.ima1.1.stud.res  
W = 0.9834, p-value = 0.0003011
```

```
> runs(ibm2.5.ima1.1.stud.res)
```

```
$pvalue  
[1] 0.175
```

```
$observed.runs
```

```
[1] 199
```

```
$expected.runs
```

```
[1] 185.4986
```

```
$n1
```

```
[1] 184
```

```
$n2
```

```
[1] 185
```

```
$k
```

```
[1] 0
```

We need to also overfit the models to check for sufficient parameterization.

```
> ibm2.5.arima1.1.1 <- arima(ibm2.5, order = c(1,1,1), method = "ML")  
> ibm2.5.arima1.1.1
```

```
Call:
```

```
arima(x = ibm2.5, order = c(1, 1, 1), method = "ML")
```

```
Coefficients:
```

```
          ar1      ma1  
          0.0300  0.1493  
s.e.      0.2103  0.2048
```

```
sigma^2 estimated as 2.839e+10:  log likelihood = -4950.96,  aic = 9905.93
```

```
> ibm2.5.ima1.2 <- arima(ibm2.5, order = c(0,1,2), method = "ML")  
> ibm2.5.ima1.2
```

```
Call:
```

```
arima(x = ibm2.5, order = c(0, 1, 2), method = "ML")
```

Coefficients:

	ma1	ma2
	0.1807	0.0104
s.e.	0.0537	0.0533

sigma² estimated as 2.839e+10: log likelihood = -4950.95, aic = 9905.91

The plot shows that the residuals seem to have zero mean and constant variance. The histogram shows a nice bell curve shape with a few outlying observations past -4. The normal Q-Q plot agrees with this, showing a very linear trend between the sample and theoretical quantiles. The Shapiro-Wilk Test for Normality shows a very low p-value of .0003. This would seem to imply that the data is not normally distributed. However, we know that the Shapiro-Wilk Test is greatly influenced by extreme observations. So, this p-value is most likely the result of the previously mentioned observations past -4. These conclusions are backed up by the Box-Ljung plot. The Sample ACF and Runs Test both tell us that independence is reasonable. So, we feel very confident saying that these residuals resemble a white noise process. Also, overfitting the model yielded additional parameters that were not statistically different from zero. Therefore, our final model for the IBM data is an **IMA(1,1) process** with a $\lambda = 2.5$ power transformation.

Data: **internet**

In Homework 4, we chose to model this data with an **ARI(3,1) process**. So, let's run diagnostics and see if this model is a good choice.

```
> internet.ari3.1 <- arima(internet, order = c(3,1,0), method = "ML")
> internet.ari3.1
```

Call:

```
arima(x = internet, order = c(3, 1, 0), method = "ML")
```

Coefficients:

	ar1	ar2	ar3
	1.1513	-0.6612	0.3407
s.e.	0.0950	0.1353	0.0941

sigma² estimated as 9.363: log likelihood = -252, aic = 509.99

>

```
> internet.ari3.1.res <- internet.ari3.1$residuals
> internet.ari3.1.stud.res <- (internet.ari3.1.res -
```

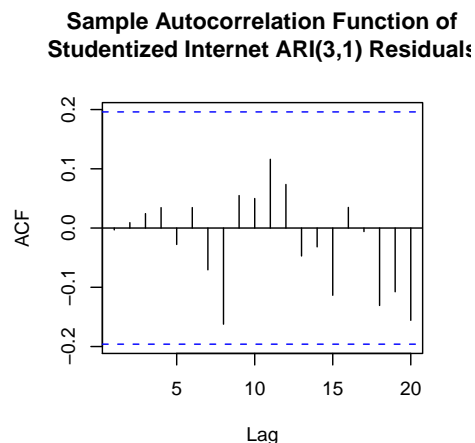
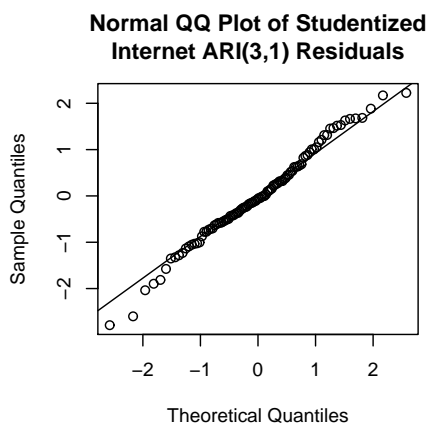
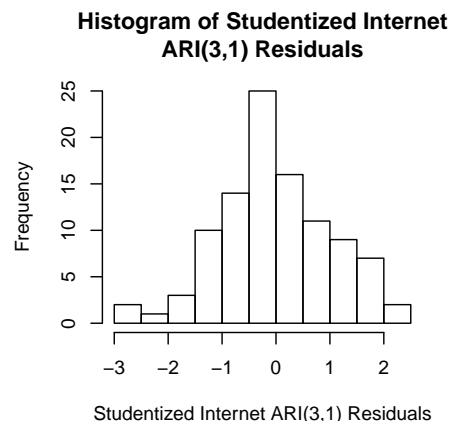
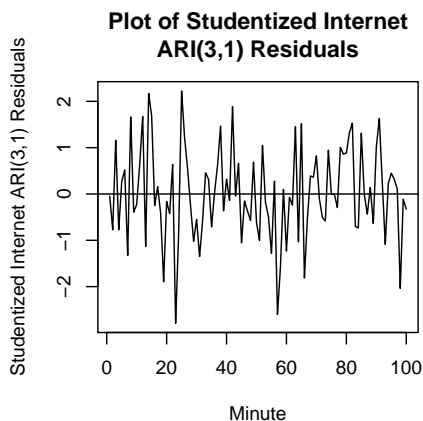
```

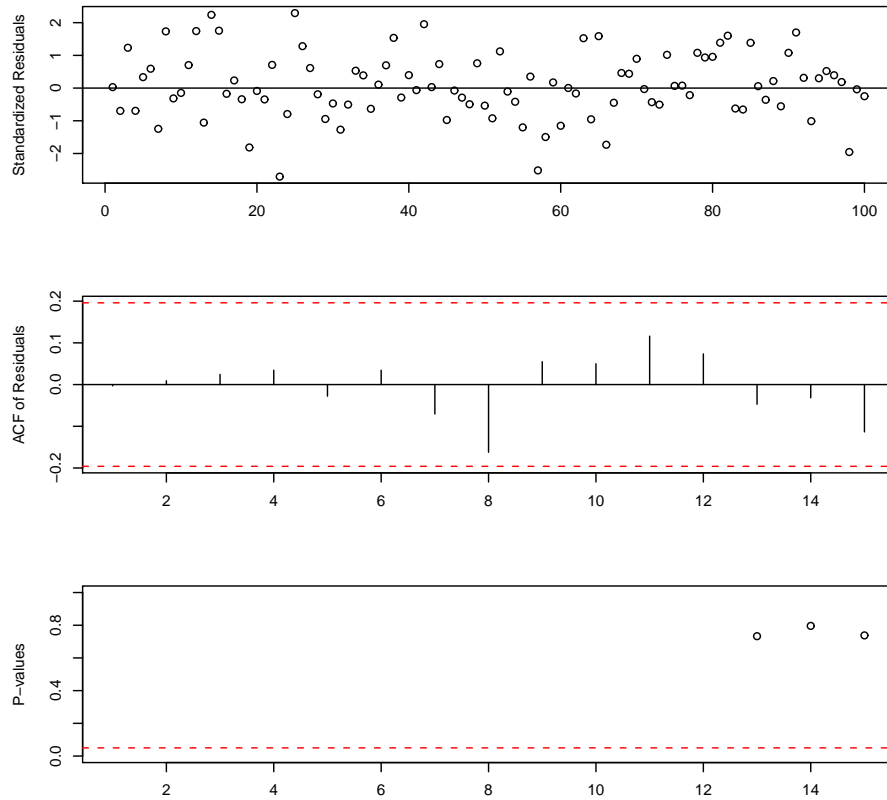
+ mean(internet.ari3.1.res)) / sd(internet.ari3.1.res)
>
> par(mfrow = c(2,2))
>
> plot.ts(internet.ari3.1.stud.res, main = "Plot of Studentized Internet
+ ARI(3,1) Residuals", xlab = "Minute", ylab =
+ "Studentized Internet ARI(3,1) Residuals")
> abline(h = 0)
> hist(internet.ari3.1.stud.res, main = "Histogram of Studentized Internet
+ ARI(3,1) Residuals", xlab = "Studentized Internet ARI(3,1) Residuals")
> qqnorm(internet.ari3.1.stud.res, main = "Normal QQ Plot of Studentized
+ Internet ARI(3,1) Residuals")
> qqline(internet.ari3.1.stud.res)
> acf(internet.ari3.1.stud.res, main = "Sample Autocorrelation Function of
+ Studentized Internet ARI(3,1) Residuals")

> par(mfrow = c(1,1))

> tsdiag(internet.ari3.1, gof = 15, omit.initial = F)

```





```
> shapiro.test(internet.ari3.1.stud.res)
```

Shapiro-Wilk normality test

```
data: internet.ari3.1.stud.res  
W = 0.9891, p-value = 0.5951
```

```
> runs(internet.ari3.1.stud.res)
```

```
$pvalue  
[1] 0.685
```

```
$observed.runs  
[1] 53
```

```
$expected.runs  
[1] 50.5
```

```
$n1  
[1] 55
```

```
$n2
```

```
[1] 45
```

```
$k
```

```
[1] 0
```

We need to also overfit the models to check for sufficient parameterization.

```
> internet.ari4.1 <- arima(internet, order = c(4,1,0), method = "ML")
> internet.ari4.1
```

```
Call:
```

```
arima(x = internet, order = c(4, 1, 0), method = "ML")
```

```
Coefficients:
```

	ar1	ar2	ar3	ar4
	1.1603	-0.6784	0.3710	-0.0260
s.e.	0.1013	0.1512	0.1521	0.1025

```
sigma^2 estimated as 9.357: log likelihood = -251.96, aic = 511.93
```

```
> internet.ari3.1.1 <- arima(internet, order = c(3,1,1), method = "ML")
> internet.ari3.1.1
```

```
Call:
```

```
arima(x = internet, order = c(3, 1, 1), method = "ML")
```

```
Coefficients:
```

	ar1	ar2	ar3	ma1
	1.0925	-0.5995	0.3232	0.0667
s.e.	0.2694	0.2999	0.1242	0.2835

```
sigma^2 estimated as 9.358: log likelihood = -251.97, aic = 511.94
```

The plot shows no significant deviation from zero mean or constant variance. The histogram, Normal Q-Q Plot, and Shapiro-Wilk Test show that normality seems quite reasonable for the residuals. The ACF and Runs Test imply that independence is also reasonable. The Box-Ljung test shows no significant problems either. Overfitting the model yields statistically insignificant parameters. Therefore, we feel confident saying that the **ARI(3,1) process** models these data well.

Data: **robot**

In Homework 4, we chose the **IMA(1,1) process** to model this data. So, let's run the diagnostics and see how it fares.

```
> robot.ima1.1 <- arima(robot, order = c(0,1,1), method = "ML")
> robot.ima1.1

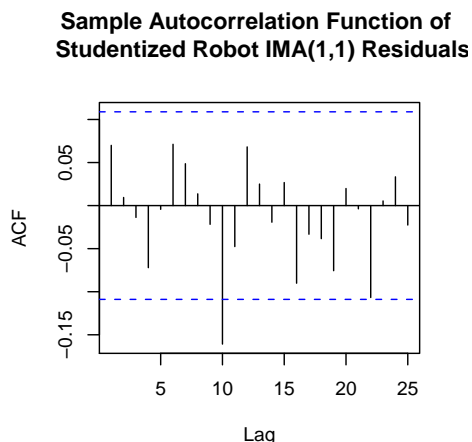
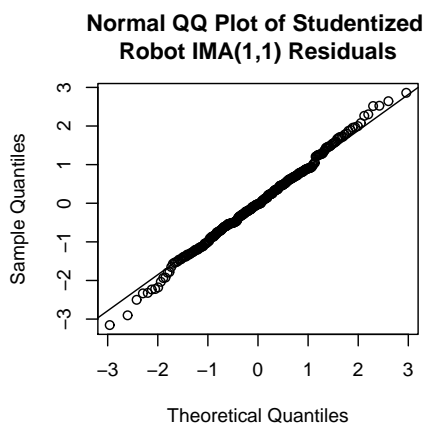
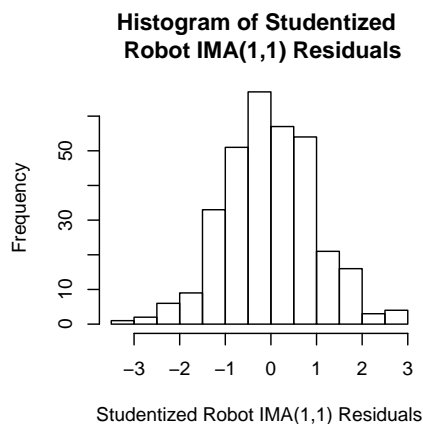
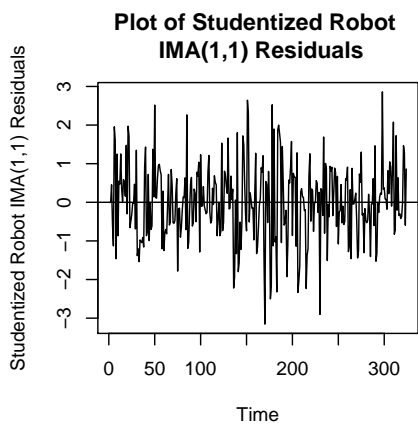
Call:
arima(x = robot, order = c(0, 1, 1), method = "ML")

Coefficients:
          ma1
      -0.8713
s.e.      0.0389

sigma^2 estimated as 6.069e-06:  log likelihood = 1480.95,  aic = -2959.9
>
> robot.ima1.1.res <- robot.ima1.1$residuals
> robot.ima1.1.stud.res <- (robot.ima1.1.res - mean(robot.ima1.1.res))
+ / sd(robot.ima1.1.res)
>
> par(mfrow = c(2,2))
>
> plot.ts(robot.ima1.1.stud.res, main = "Plot of Studentized Robot
+ IMA(1,1) Residuals", xlab = "Time", ylab =
+ "Studentized Robot IMA(1,1) Residuals")
> abline(h = 0)
> hist(robot.ima1.1.stud.res, main = "Histogram of Studentized
+ Robot IMA(1,1) Residuals", xlab = "Studentized Robot IMA(1,1) Residuals")
> qqnorm(robot.ima1.1.stud.res, main = "Normal QQ Plot of Studentized
+ Robot IMA(1,1) Residuals")
> qqline(robot.ima1.1.stud.res)
> acf(robot.ima1.1.stud.res, main = "Sample Autocorrelation Function of
+ Studentized Robot IMA(1,1) Residuals")

> par(mfrow = c(1,1))

> tsdiag(robot.ima1.1, gof = 15, omit.initial = F)
```



```
> shapiro.test(robot.ima1.1.stud.res)
```

Shapiro-Wilk normality test

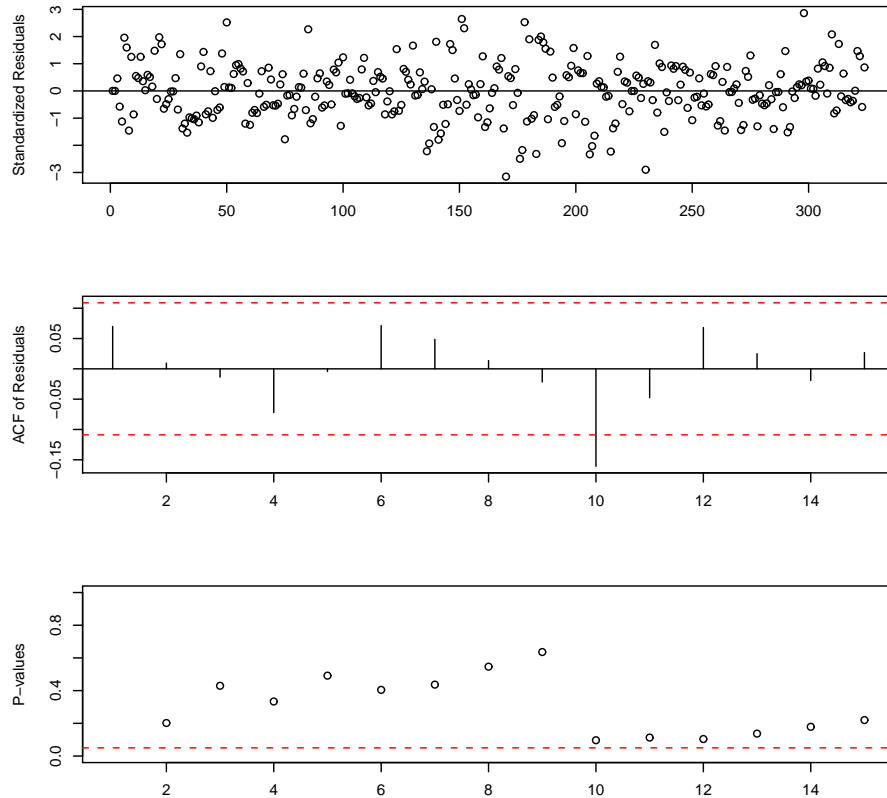
```
data: robot.ima1.1.stud.res
W = 0.9969, p-value = 0.791
```

```
> runs(robot.ima1.1.stud.res)
$pvalue
[1] 0.0132
```

```
$observed.runs
[1] 140
```

```
$expected.runs
[1] 162.6975
```

```
$n1
[1] 169
```



```
$n2
[1] 155
```

```
$k
[1] 0
```

We need to also overfit the models to check for sufficient parameterization.

```
> robot.arma1.1.1 <- arima(robot, order = c(1,1,1), method = "ML")
> robot.arma1.1.1
```

```
Call:
arima(x = robot, order = c(1, 1, 1), method = "ML")
```

```
Coefficients:
      ar1      ma1
  0.1208 -0.9215
s.e.  0.0715  0.0428
```

```
sigma^2 estimated as 6.012e-06:  log likelihood = 1482.35,  aic = -2960.7
```

```
> robot.ima1.2 <- arima(internet, order = c(0,1,2), method = "ML")
> robot.ima1.2
```

Call:

```
arima(x = internet, order = c(0, 1, 2), method = "ML")
```

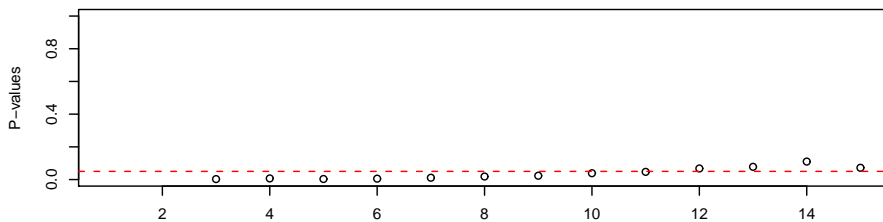
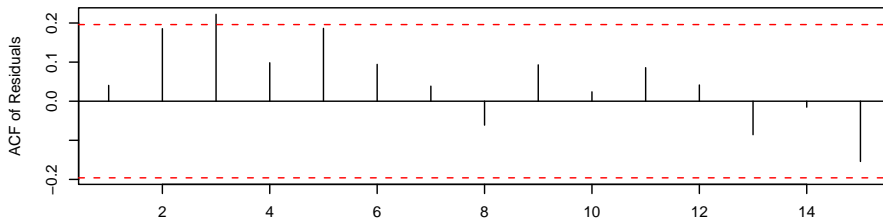
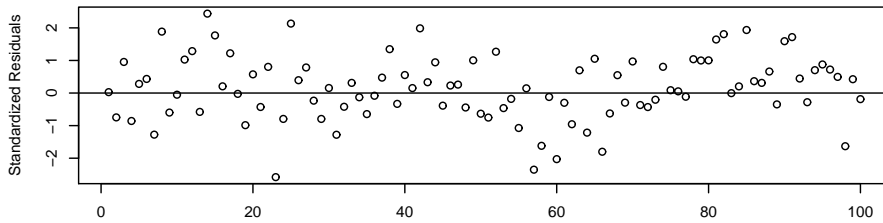
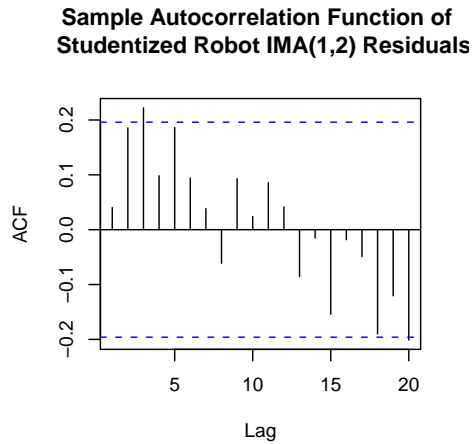
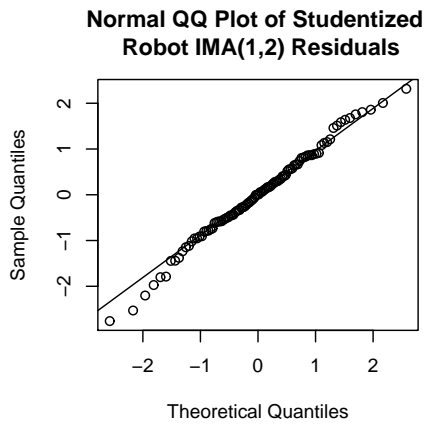
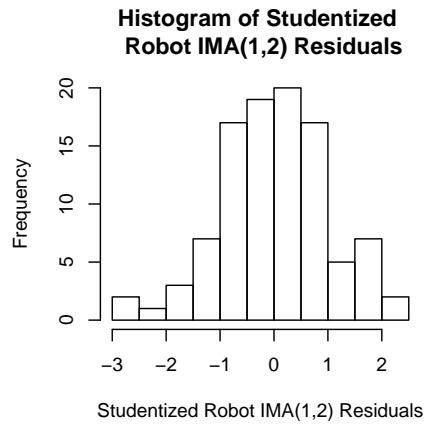
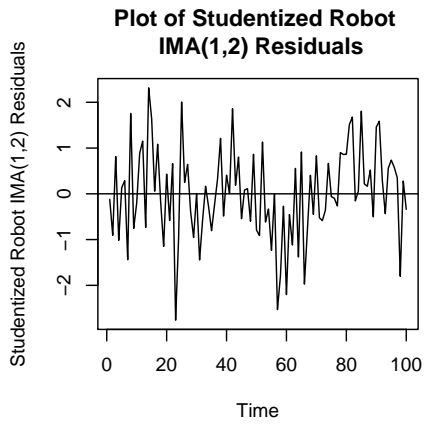
Coefficients:

```
          ma1      ma2
          1.1978  0.5780
s.e.      0.0862  0.0912
```

```
sigma^2 estimated as 10.34:  log likelihood = -256.94,  aic = 517.87
```

Despite the fact that the plots and tests look nice, our overfitting has yielded an interesting result. The second moving average parameter is quite significant. So, let's rerun our diagnostics with the **IMA(1,2) process**.

```
> robot.ima1.2.res <- robot.ima1.2$residuals
> robot.ima1.2.stud.res <- (robot.ima1.2.res - mean(robot.ima1.2.res)) /
+ sd(robot.ima1.2.res)
>
> par(mfrow = c(2,2))
>
> plot.ts(robot.ima1.2.stud.res, main = "Plot of Studentized Robot
+ IMA(1,2) Residuals", xlab = "Time", ylab =
+ "Studentized Robot IMA(1,2) Residuals")
> abline(h = 0)
> hist(robot.ima1.2.stud.res, main = "Histogram of Studentized
+ Robot IMA(1,2) Residuals", xlab =
+ "Studentized Robot IMA(1,2) Residuals")
> qqnorm(robot.ima1.2.stud.res, main = "Normal QQ Plot of Studentized
+ Robot IMA(1,2) Residuals")
> qqline(robot.ima1.2.stud.res)
> acf(robot.ima1.2.stud.res, main = "Sample Autocorrelation Function of
+ Studentized Robot IMA(1,2) Residuals")
>
> par(mfrow = c(1,1))
>
> tsdiag(robot.ima1.2, gof = 15, omit.initial = F)
```



```
> shapiro.test(robot.ima1.2.stud.res)
```

Shapiro-Wilk normality test

```
data: robot.ima1.2.stud.res
W = 0.9916, p-value = 0.7895
```

```
> runs(robot.ima1.2.stud.res)
```

```
$pvalue
[1] 0.924
```

```
$observed.runs
[1] 51
```

```
$expected.runs
[1] 50.98
```

```
$n1
[1] 49
```

```
$n2
[1] 51
```

```
$k
[1] 0
```

We need to also overfit the models to check for sufficient parameterization.

```
> robot.arima1.1.2 <- arima(robot, order = c(1,1,2), method = "ML")
> robot.arima1.1.2
```

Call:

```
arima(x = robot, order = c(1, 1, 2), method = "ML")
```

Coefficients:

	ar1	ma1	ma2
	0.8333	-1.6751	0.6831
s.e.	0.1146	0.1392	0.1335

```
sigma^2 estimated as 5.956e-06: log likelihood = 1483.69, aic = -2961.38
```

```
> robot.ima1.3 <- arima(internet, order = c(0,1,3), method = "ML")
```

```
> robot.ima1.3
```

```
Call:
```

```
arima(x = internet, order = c(0, 1, 3), method = "ML")
```

```
Coefficients:
```

	ma1	ma2	ma3
	1.2179	0.6672	0.1376
s.e.	0.1018	0.1220	0.1026

```
sigma^2 estimated as 10.19:  log likelihood = -256.14,  aic = 518.28
```

The plot appears to have approximate zero mean and constant variance. However, there seems to be some structure to it, i.e. it is not white noise. This is back up by the fact that the ACF has significant decreasing structure. So, in spite of the maximum likelihood estimate being significant, we have substantial evidence that we overparameterized this model. In fact, if we had looked at the AIC and estimate white noise variance components from the **IMA(1,1)** and **IMA(1,2) processes**, we would have immediately seen that the **IMA(1,1) process** is a better fit for the data. So, let's move back to the **IMA(1,1) model**.

The plot appears to have zero mean and constant variance. The histogram, Normal Q-Q Plot, and Shapiro-Wilk Test for Normality all agree that normality is reasonable. The Runs Test returned a significant p-value, which could imply a lack of independence. The cause for this could be the large sample size combined with a significant sample autocorrelation at lag $k = 10$. So, we believe this is reasonable to overlook. The Box-Ljung Plot did not present any new issues. Therefore, we feel confident saying that the **IMA(1,1) process** models this data well.