Time Series HW 2

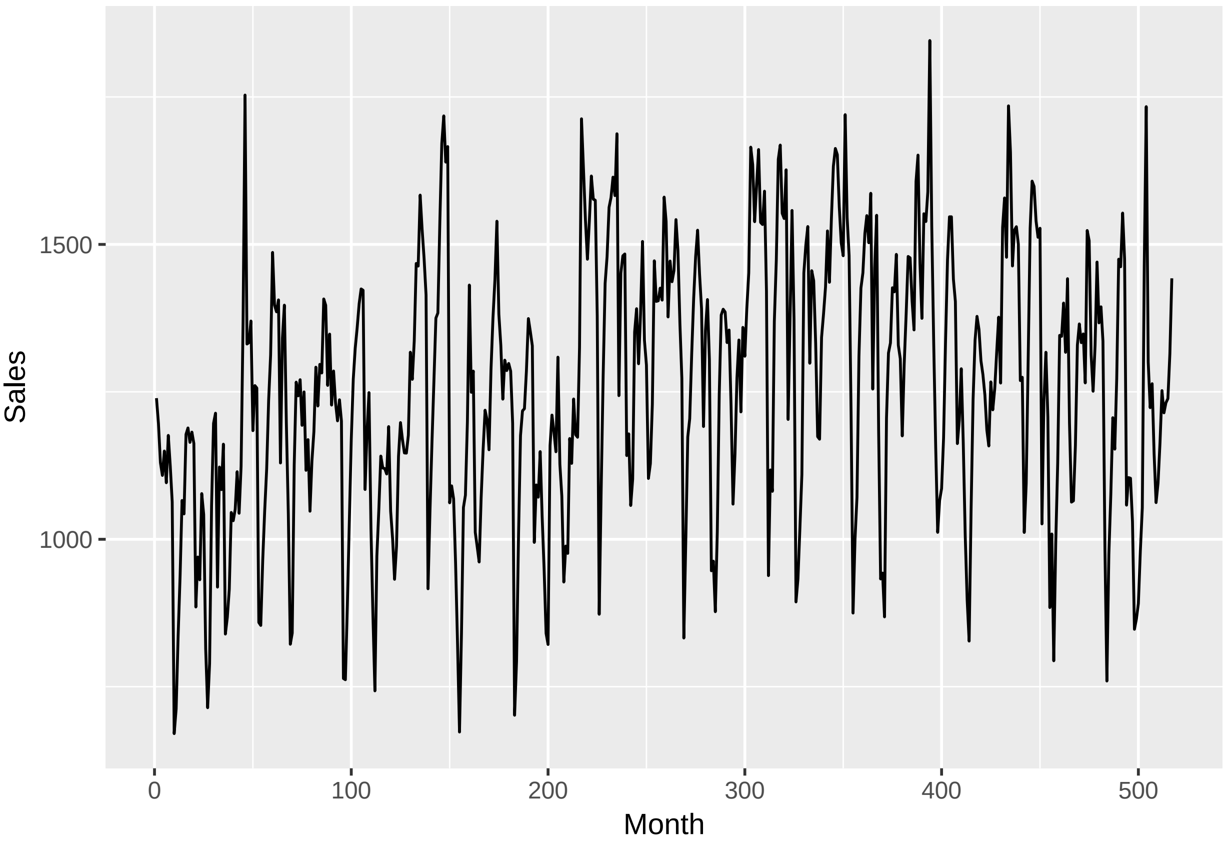
Kylie Taylor

2/14/2019

Question 1

I will be using a data set obtained from the FRED on the number of vehicle sales each month in the United States, starting in Jan 1976 until Jan 2019. The following plot is a plot of the time series. There is an obvious positive trend in the series. Income is clearly increasing with each period.

```r
sales$y <- sales$sales
TS <- ts(sales, start = c(1976, 1), frequency = 1)
ts <- data.frame(TS)
ggplot(ts, aes(date, y)) + geom_line() +
  xlab("Month") + ylab("Sales")
```



The Dickey Fuller test results below reveal that there are indications of a unit root. A unit root implies that a shock in one period had lasting effects on the series. I ran three types of ADF tests with up to 4 lags, 1) no constant and no trend, 2) with a constant and no trend, and 3) with a constant and trend. The p-values from every test reveal that there is a unit root. My choice of test would be the ADF with a lag of 1, since it has the highest p-value across all three types of tests. This means I fail to reject the null hypothesis that there is a unit root, implying that a unit root is present in the series.

I would continue working with the series by taking the first difference of the series. This would hopefully return a stationary series with a constant mean and variance.

```r
adf.test(ts$y, nlag = NULL, output = TRUE)
```

```
## Augmented Dickey-Fuller Test
## alternative: stationary
##
## Type 1: no drift no trend
##      lag    ADF p.value
## [1,]   0 -1.445   0.162
## [2,]   1 -0.861   0.371
## [3,]   2 -0.655   0.445
## [4,]   3 -0.722   0.421
## [5,]   4 -0.629   0.454
## [6,]   5 -0.609   0.461
## Type 2: with drift no trend
##      lag   ADF p.value
## [1,]   0 -9.28    0.01
## [2,]   1 -6.31    0.01
## [3,]   2 -4.97    0.01
## [4,]   3 -5.29    0.01
## [5,]   4 -4.87    0.01
## [6,]   5 -4.79    0.01
## Type 3: with drift and trend
##      lag   ADF p.value
## [1,]   0 -9.95    0.01
## [2,]   1 -6.80    0.01
## [3,]   2 -5.41    0.01
## [4,]   3 -5.81    0.01
## [5,]   4 -5.38    0.01
## [6,]   5 -5.33    0.01
## ----
## Note: in fact, p.value = 0.01 means p.value <= 0.01
```

The following plots are the ACF's and PACF's of the original series and the first differenced series. The plots of the original series and the transformed series reveal that the first differences make the series appear to be stationary. This is validated by the ACF's and PACF's. The ACF of the original series reveals very high correlation between lags of sales and the PACF reveal high correlation as well, especially with the first lag and 12th lag.

The ACF for the transformed series reveals much less persistent shocks, and converges to a correlation of 0 significantly quicker than the original series. The PACF plot has a large spike only at the first lag, which means that all the higher-order auto-correlations are effectively explained by the first lag auto-correlation.

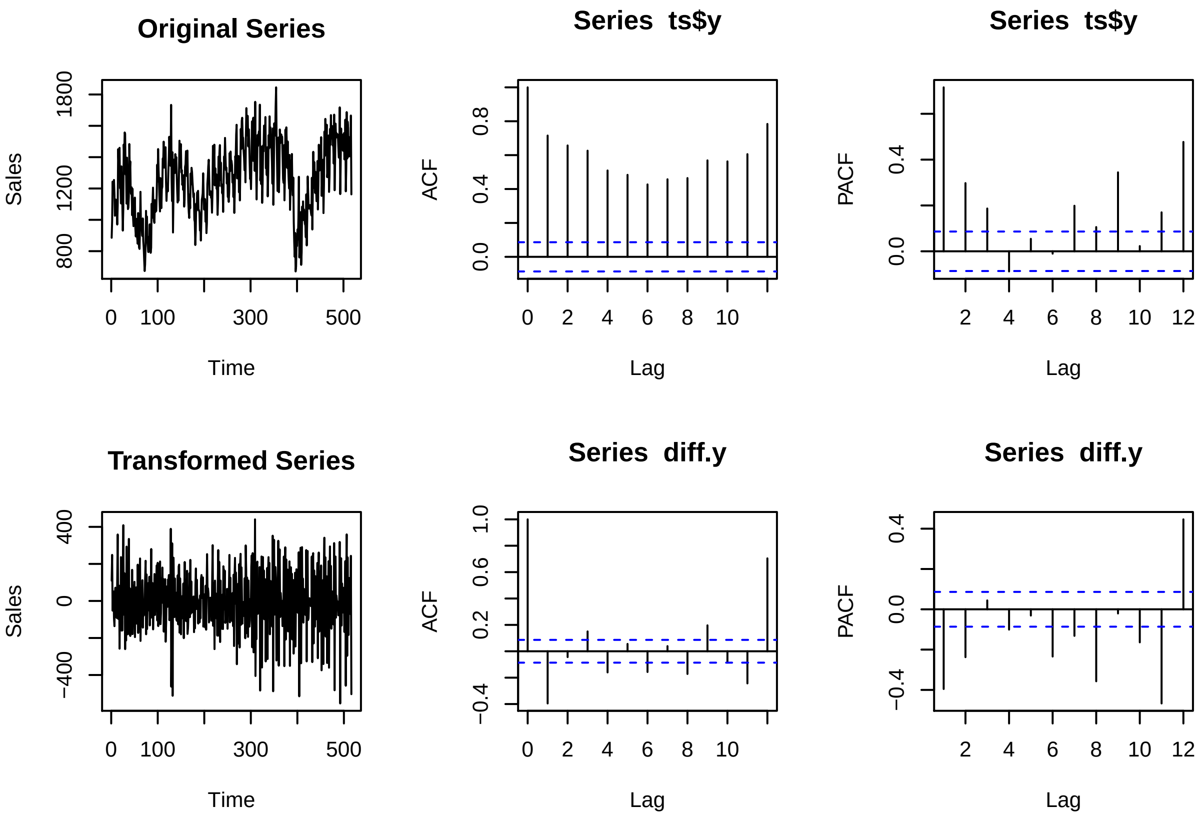
The two models I would choose to start with would be the AR(12) and the ARMA(2,2). I choose this by observing the ACF's and PACF's of the transformed series. The large spike at the first lag of the PACF reveals that I should include an AR term of at least the first order. The first and second lags in the ACF are greater than 0.2 and -0.2, respectively, which leads me to include at least the first order MA term.

```r
set.seed(123456)
diff.y <- diff(ts$y, differences = 1)
par(mfrow=c(2,3))

plot.ts(ts$y, ylab= "Sales", main = "Original Series")
acf(ts$y, lag.max=12, ylab = "ACF")
pacf(ts$y, lag.max=12, ylab="PACF")

plot.ts(diff.y, ylab= "Sales", main = "Transformed Series")
```

```
acf(diff.y, lag.max=12, ylab = "ACF")
pacf(diff.y, lag.max=12, ylab="PACF")
```



The AIC and the BIC of the AR(12) and ARMA(2,2) reveal that the AR(12) is a slightly better model to use since it has marginally smaller AIC and BIC.

```
library(tseries)
```

```
##
## Attaching package: 'tseries'

## The following objects are masked from 'package:aTSA':
##
##     adf.test, kpss.test, pp.test
```

```
ARMA1 <- arima(x=diff.y, order=c(12,0,0), method="ML")
ARMA2 <- arima(x = diff.y, order = c(2,0,2), method="ML")
pander::pander(ARMA1)
```

Call: arima(x = diff.y, order = c(12, 0, 0), method = "ML")

Table 1: Coefficients (continued below)

	ar1	ar2	ar3	ar4	ar5	ar6	ar7
	-0.4698	-0.3584	-0.2562	-0.276	-0.2485	-0.3289	-0.3288
s.e.	0.03949	0.04386	0.04553	0.04606	0.04521	0.04378	0.04378

	ar8	ar9	ar10	ar11	ar12	intercept
	-0.3408	-0.2004	-0.236	-0.1995	0.4447	0.7575
s.e.	0.04487	0.04585	0.04524	0.04388	0.03958	1.176

sigma^2 estimated as 10160: log likelihood = -3116.76, aic = 6261.52

```
pander::pander(ARMA2)
```

Call: arima(x = diff.y, order = c(2, 0, 2), method = "ML")

Table 3: Coefficients

	ar1	ar2	ma1	ma2	intercept
	-0.5011	0.1997	-0.02172	-0.6903	0.6473
s.e.	0.08638	0.06139	0.07142	0.06242	1.443

sigma^2 estimated as 21511: log likelihood = -3306.46, aic = 6624.92

```
AIC1 = AIC(ARMA1)
BIC1 = AIC(ARMA1, k = log(length(ts$y)))
AIC2 = AIC(ARMA2)
BIC2 = AIC(ARMA2, k = log(length(ts$y)))
AIC1
```

```
## [1] 6261.52
```

```
BIC1
```

```
## [1] 6320.993
```

```
AIC2
```

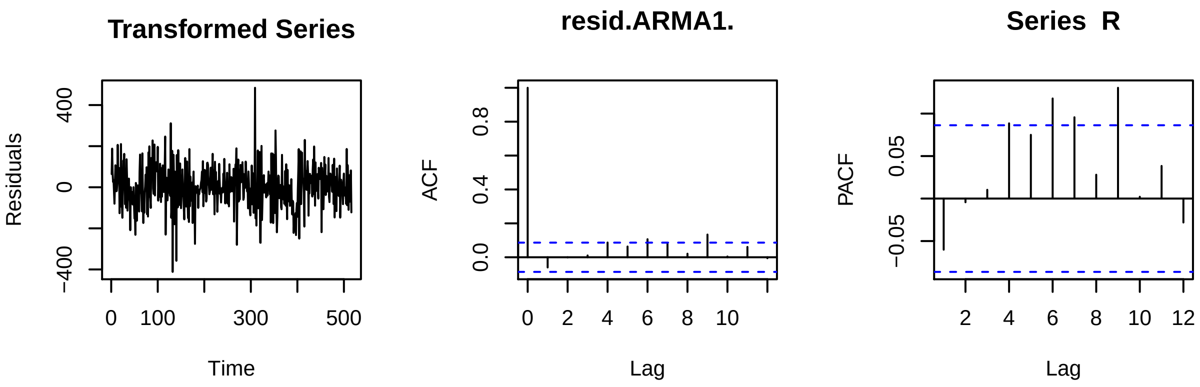
```
## [1] 6624.918
```

```
BIC2
```

```
## [1] 6650.406
```

** do a Newwy-West test here ** I found that the AR(12) residuals are also stationary. The ACF has one large spike at the first lag, revealing that the series of residuals may be MA(1). The PACF is contained withing the 0.1 and -0.1 bounds, which reveals that there is not likely a AR(q) term.

```
R <- data.frame(resid(ARMA1))
par(mfrow = c(2,3))
plot.ts(R, ylab= "Residuals", main = "Transformed Series")
acf(R, lag.max=12, ylab = "ACF")
pacf(R, lag.max=12, ylab="PACF")
```



The ARCH test are testing the following model for various p's:

$a_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \alpha_2 a_{t-2}^2 + ... + \alpha_p a_{t-p}^2 + e_t$

Where we are testing if

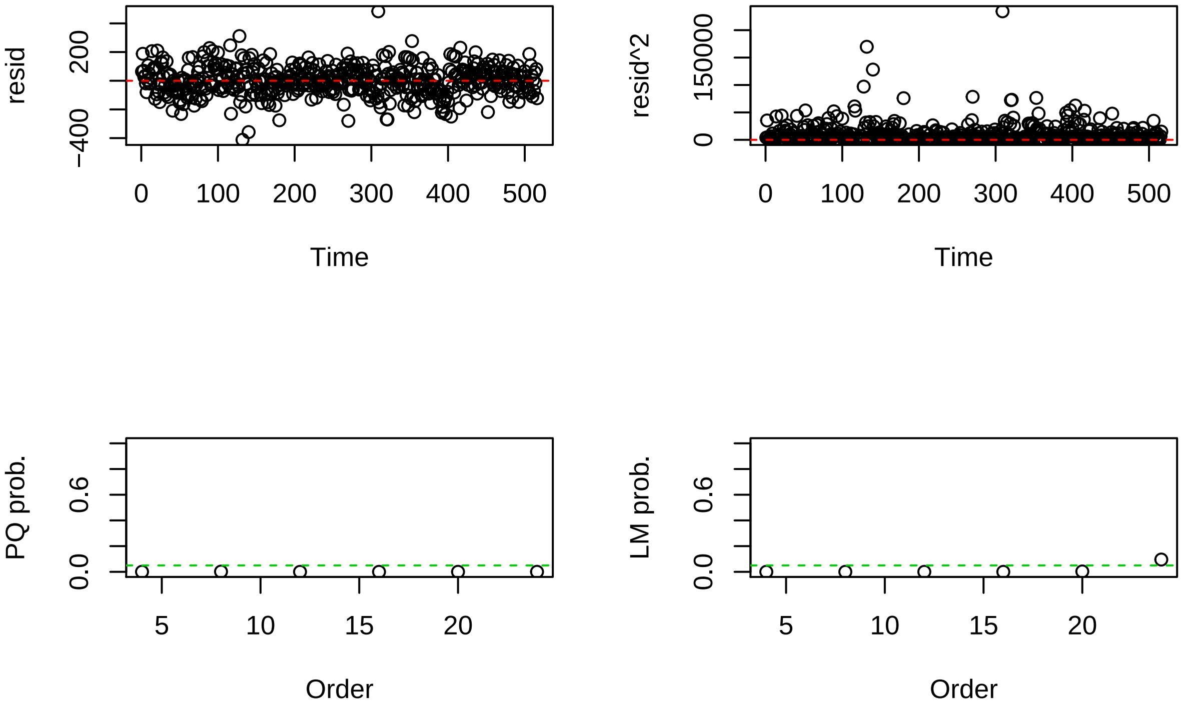
$\alpha_0 = \alpha_1 = \alpha_2 = ... = \alpha_p = 0$

If hypothesis is accepted then we can say that series have no ARCH effects. If it is rejected then one or more coefficients are non zero and we say that there are ARCH effects. +

The p-values of the LM test reveal that we reject the null hypothesis that all the regressors are equal, for all various orders (4, 8, 12, 16, 20, and 24 orders were tested). This leads me to believe that heteroskedasticity is not present. Based on the output of the LM test, I see that all of the ARCH orders do reveal heteroskedasticity of the transformed series.

```r
library(aTSA)
arch.test(ARMA1, output = TRUE)
```

```
## ARCH heteroscedasticity test for residuals
## alternative: heteroscedastic
##
## Portmanteau-Q test:
##      order   PQ  p.value
## [1,]     4 20.6 3.79e-04
## [2,]     8 24.7 1.72e-03
## [3,]    12 49.4 1.81e-06
## [4,]    16 50.3 2.05e-05
## [5,]    20 53.9 5.92e-05
## [6,]    24 57.7 1.36e-04
## Lagrange-Multiplier test:
##      order    LM  p.value
## [1,]     4 283.1 0.00e+00
## [2,]     8 130.0 0.00e+00
## [3,]    12  78.4 3.04e-12
## [4,]    16  51.8 5.99e-06
## [5,]    20  40.2 3.08e-03
## [6,]    24  32.2 9.55e-02
```



The following plots are of various ARMA models that I thought make sense economically and statistically. I included an ARMA(1,1), ARMA(2,2), AR(12), ARMA(12,1), and ARMA(12,2). The two models I chose are the ARMA(12,1) and the ARMA(12,2). The statistical reasoning for picking the ARMA(12,2) model was that it has the lowest AIC's and BIC's out of all the models tested, 6230 and 6298 respectively. My economic reasoning for picking the ARMA(1) model was because sales from a year ago are likely most similar to sales of this month, because there tends to be seasonality in car sales (think Christmas gifts). While the ARMA(12,1) is not the most simple model I could use, I think it still maintains significance because of the magnitude of the data set. The next model I chose was the ARMA(12,2). This model has very comparable AIC and BIC coefficients, at 6631 and 6296 respectively. My main driving reason for picking this as a top two models, was my economic reasoning. I think it is important to include the AR 12 order, because it has shown to be the best predictor of current month's sales, which is intuitive. I also think to include a MA(2) term because errors, or noise, from last 2 months sales is also likely to affect this month's sales. If sales take a hit or jumps up quite a bit for various reasons, the effects from that will likely still be felt in the current month.

```
library(pander)
ARMA1 <- arima(x=diff.y, order=c(1,0,1), method="ML")
pander(ARMA1)
```

Call: arima(x = diff.y, order = c(1, 0, 1), method = "ML")

Table 4: Coefficients

	ar1	ma1	intercept
	0.3097	-0.8495	0.6513
s.e.	0.0537	0.02724	1.426

sigma^2 estimated as 21638: log likelihood = -3307.97, aic = 6623.94

```
AIC(ARMA1)
```

```
## [1] 6623.94
```

```r
AIC(ARMA1, k = log(length(ts$y)))
```

```
## [1] 6640.932
```

```r
ARMA2 <- arima(x = diff.y, order = c(2,0,2), method="ML")
pander(ARMA2)
```

Call: arima(x = diff.y, order = c(2, 0, 2), method = “ML”)

Table 5: Coefficients

	ar1	ar2	ma1	ma2	intercept
	-0.5011	0.1997	-0.02172	-0.6903	0.6473
s.e.	0.08638	0.06139	0.07142	0.06242	1.443

sigma^2 estimated as 21511: log likelihood = -3306.46, aic = 6624.92

```r
AIC(ARMA2)
```

```
## [1] 6624.918
```

```r
AIC(ARMA2, k = log(length(ts$y)))
```

```
## [1] 6650.406
```

```r
ARMA3 <- arima(x = diff.y, order = c(12,0,0), method="ML")
pander(ARMA3)
```

Call: arima(x = diff.y, order = c(12, 0, 0), method = “ML”)

Table 6: Coefficients (continued below)

	ar1	ar2	ar3	ar4	ar5	ar6	ar7
	-0.4698	-0.3584	-0.2562	-0.276	-0.2485	-0.3289	-0.3288
s.e.	0.03949	0.04386	0.04553	0.04606	0.04521	0.04378	0.04378

	ar8	ar9	ar10	ar11	ar12	intercept
	-0.3408	-0.2004	-0.236	-0.1995	0.4447	0.7575
s.e.	0.04487	0.04585	0.04524	0.04388	0.03958	1.176

sigma^2 estimated as 10160: log likelihood = -3116.76, aic = 6261.52

```r
AIC(ARMA3)
```

```
## [1] 6261.52
```

```r
AIC(ARMA3, k = log(length(ts$y)))
```

```
## [1] 6320.993
```

```r
ARMA4 <- arima(x = diff.y, order = c(12,0,1), method="ML")
pander(ARMA4)
```

Call: arima(x = diff.y, order = c(12, 0, 1), method = “ML”)

Table 8: Coefficients (continued below)

	ar1	ar2	ar3	ar4	ar5	ar6
	-0.01034	-0.04355	0.003663	-0.08707	-0.01369	-0.1139
s.e.	0.04906	0.04043	0.03819	0.0346	0.03612	0.03461

Table 9: Table continues below

	ar7	ar8	ar9	ar10	ar11	ar12
	-0.06105	-0.09827	0.0582	-0.06618	0.01042	0.655
s.e.	0.03768	0.03648	0.03776	0.03417	0.03487	0.03354

	ma1	intercept
	-0.6266	0.9043
s.e.	0.05857	2.075

sigma^2 estimated as 9552: log likelihood = -3101.23, aic = 6232.46

```
AIC(ARMA4)
```

```
## [1] 6232.456
```

```
AIC(ARMA4, k = log(length(ts$y)))
```

```
## [1] 6296.176
```

```
ARMA5 <- arima(x = diff.y, order = c(12,0,2), method="ML")
pander(ARMA5)
```

Call: arima(x = diff.y, order = c(12, 0, 2), method = “ML”)

Table 11: Coefficients (continued below)

	ar1	ar2	ar3	ar4	ar5	ar6
	-0.06356	-0.002081	0.02181	-0.06114	-0.01739	-0.1024
s.e.	0.05047	0.04372	0.03785	0.03588	0.03483	0.03402

Table 12: Table continues below

	ar7	ar8	ar9	ar10	ar11	ar12
	-0.0676	-0.08387	0.061	-0.03923	0.01116	0.6676
s.e.	0.03613	0.03601	0.03619	0.03576	0.03371	0.03312

	ma1	ma2	intercept
	-0.5242	-0.1313	0.9376
s.e.	0.06779	0.06097	2.158

sigma^2 estimated as 9470: log likelihood = -3099.06, aic = 6230.11

```
AIC(ARMA5)
```

```
## [1] 6230.113
```

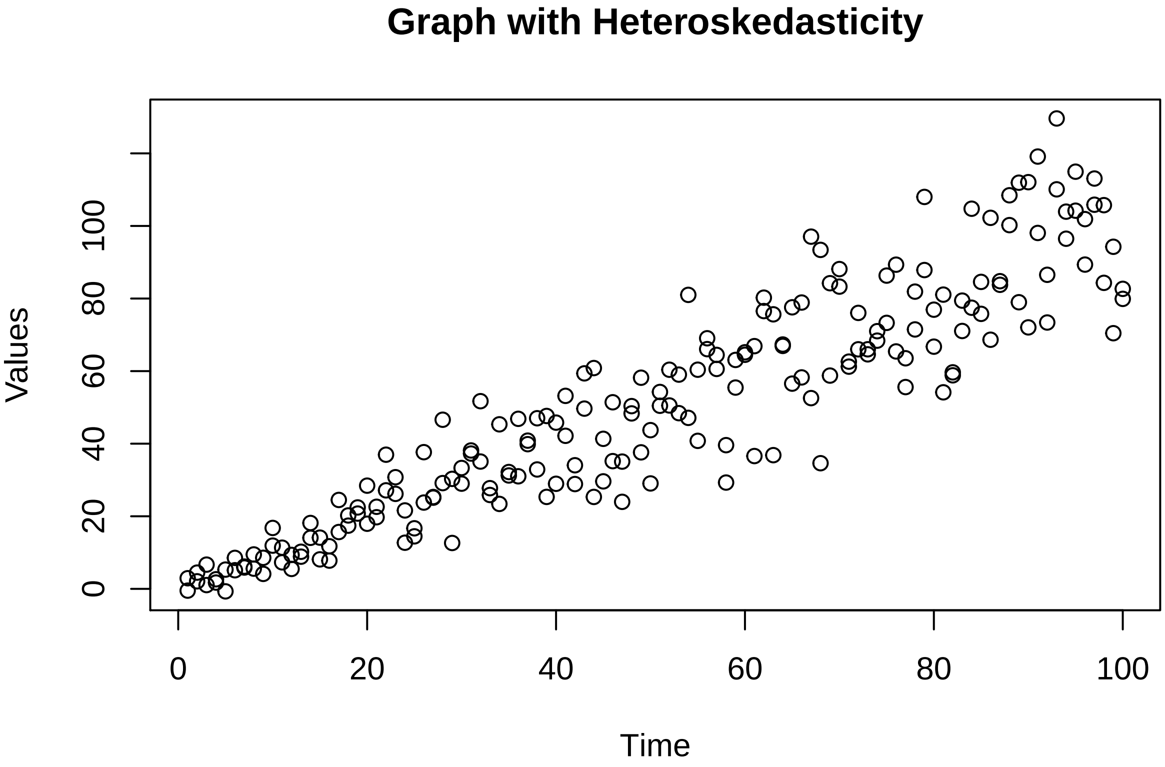
```
AIC(ARMA5, k = log(length(ts$y)))
```

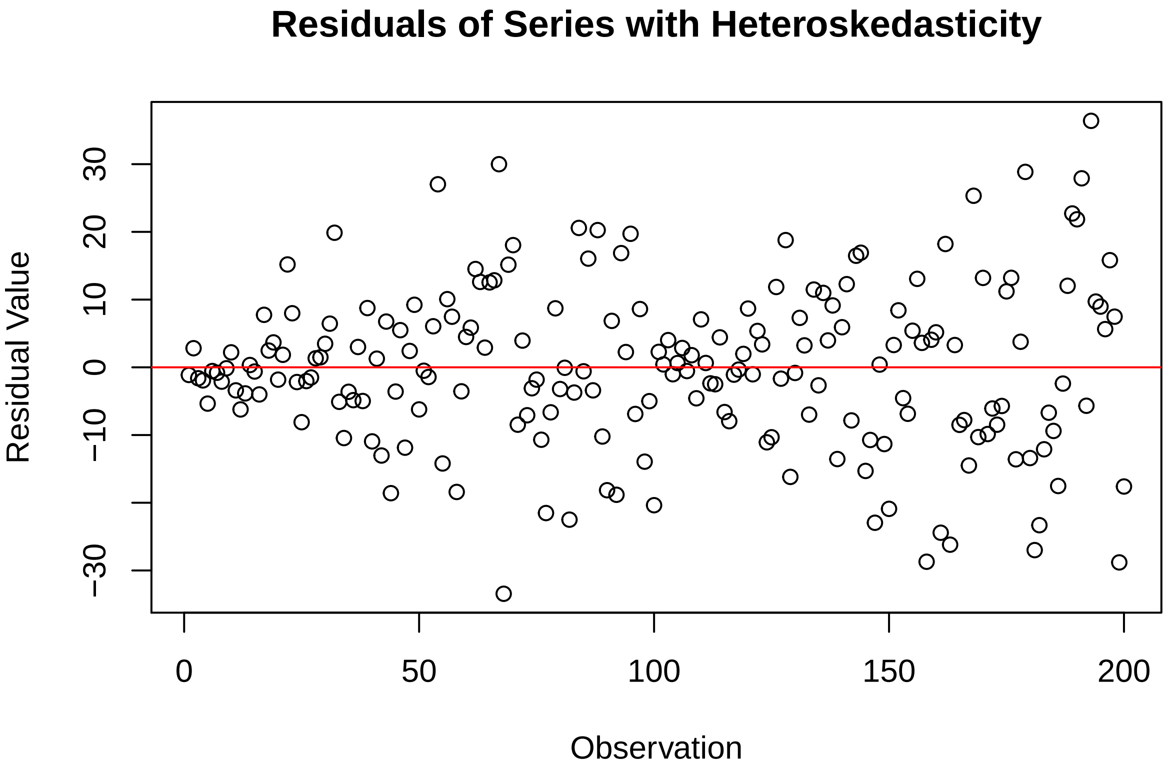
```
## [1] 6298.082
```

2

a)

Heteroskedasticity is when the errors of a series do not have constant variance. This might be important for a specific series an individual is using, because observations with large variances will have a larger effect than other observations, resulting in bias. In the case of the income example above, if a particular quarter has a large variance of income in comparison to other quarters, the model we run will try to fit to that variance, and the findings will be biased or swayed because of this uncommon observation.





Instead of drawing a graph of a series with heteroskedasticity, I figured I would simulate a series with heteroskedasticity and its residuals. A series with heteroskedasticity has a cone shaped appearance, where the plotted points either get further apart of closer together over time. Heteroskedasticity can be confirmed by analyzing the residuals. Homoskedastistic errors will be evenly distributed around zero, whereas heteroskedastistic errors will also have a cone shaped distribution, revealing that variance is not constant over time.

I would first examine the graphs of the series and residuals to see if I can visually identify heteroskedasticity. Next, I would use the Breusch-Pagan test for heteroskedasticity to verify if heteroskedasticity is present in the series. I do this by testing

$u_t^2 = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_p x_p + e_t$

Where we are testing if

$\alpha_0 = \alpha_1 = \alpha_2 = ... = \alpha_p = 0$

If I find that my series has heteroskedasticty or ARCH or GARCH errors, I would re-run my model with heteroskedastic robust standard errors, or preform a weighted least squares instead.

b)

Auto-correlation is when a given variable's observation is correlated to past observation(s) of the same variable. The OLS property that is not valid in the face of auto-correlation is TS5, no serial correlation. TS1 - TS4 and TS6 are valid under auto-correlation. Issues arising by the presence of auto-correlation when there is a lagged dependent variables is that the OLS estimators could be biased and inconsistent. In an AR(1) model, the lag of the dependent variable is correlated with the residuals of the lag, which is also correlated with current residuals. This means that contemporaneous exogeneity does not hold. In the presence of auto-correlation, OLS standard errors overstate statistical significance because there is less independent variation. To correct for this issue, I would first difference the series. We would do inference in the presence of auto-correlation first by checking the model's specifications, making adjustments I feel that are necessary (like adding more lags) and testing for auto-correlation on the correctly specified model. Next I would take the first difference of the dependent variable, in hopes to make the series stationary. Last I would use Newey-West auto-correlation (and heteroskedasticity) adjusted standard errors.

3)

See attachment.