1. Introduction

In the world of finance, the Black Scholes Merton (BSM) model has been the standard option pricing model for the past 50 years. The model's output is central to pricing derivatives and therefore indispensable to portfolio managers everywhere.

Options provide security from price fluctuations to parties that buy them, so the more uncertain future prices are going to be, the more the option value. Within the BSM model volatility is the representation of uncertainty. To be relevant and survive in the market one has to accurately forecast volatility.

General market volatility can be used by economist, policy makers, central banks to gauge market sentiment about the future; they can thereafter alter their decisions based on what the index shows. For instance a central bank might issue a statement about future interest rates if it sees the index trending upward in order to assure investors that it will provide liquidity in case the market needs it hence calming traders down.

There are multiple ways of estimating market volatility but one of the widely accepted means is by using the volatility index (VIX) from the Chicago Board Options Exchange (CBOE). The CBOE is a trusted source of information because the exchange trades billions of US dollars in security value every week. The volatility is tracked continuously during the day but just like stock prices, the close is the most important to traders.

Modeling volatility is usually done over short periods of time e.g. Daily, weekly or monthly mostly because predicting too much into the future has significant uncertainty error that makes the estimation process not worth the endeavor.

2. Data

We used the daily closing VIX from January 2004 to December 2007. We picked this period because it was just before the global financial crisis of 2008. The volatility from July 2008 were significant outliers and would have distorted our estimates. The period we chose also had some interesting features because volatility was not constant during the period, meaning modeling the series was still not an easy task.

3. Methodology

Once we analysed the data we concluded that we needed to have some differencing since our original data is not stationary.

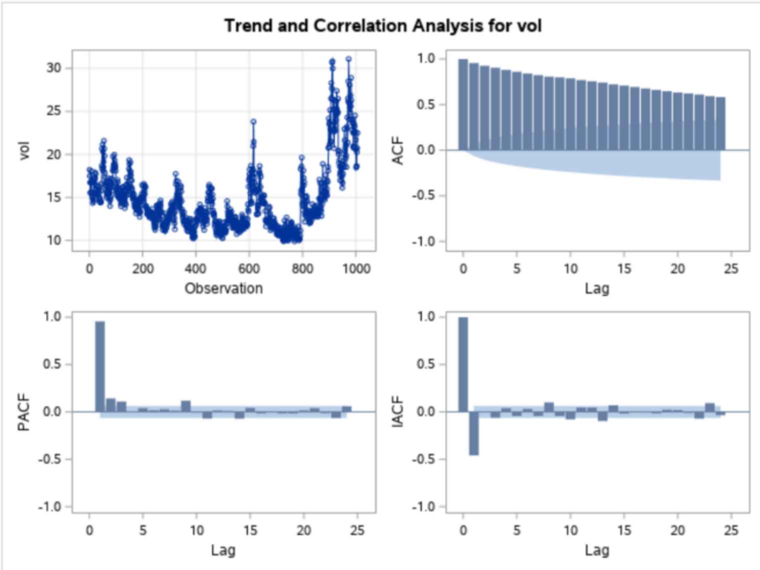
The original dataset

The ARIMA Procedure

Name of Variable = vol	
Mean of Working Series	14.65645
Standard Deviation	3.717646
Number of Observations	1006

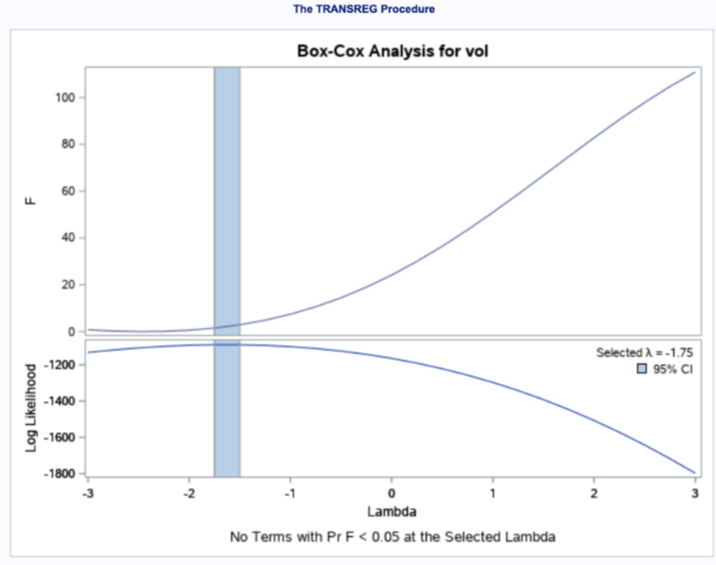
Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4865.42	6	<.0001	0.956	0.926	0.904	0.880	0.860	0.841
12	8682.72	12	<.0001	0.823	0.807	0.799	0.788	0.769	0.756
18	9999.99	18	<.0001	0.742	0.722	0.708	0.693	0.677	0.662
24	9999.99	24	<.0001	0.647	0.633	0.622	0.611	0.593	0.582



The fist 600 observations had some decreasing trend but after the trend seemed to be upward. The ACF graph clearly shows that the that there is some serial correlation for past lags.

For this project we decided to model log volatility because we did not want our final log have a chance of being negative. If we ran a Box Cox transformation procedure to determine what transformation we needed to use, we ended up with 1/Y^1.75. This ended up yielding negative volatility values which do not make sense and would break the BSM model.

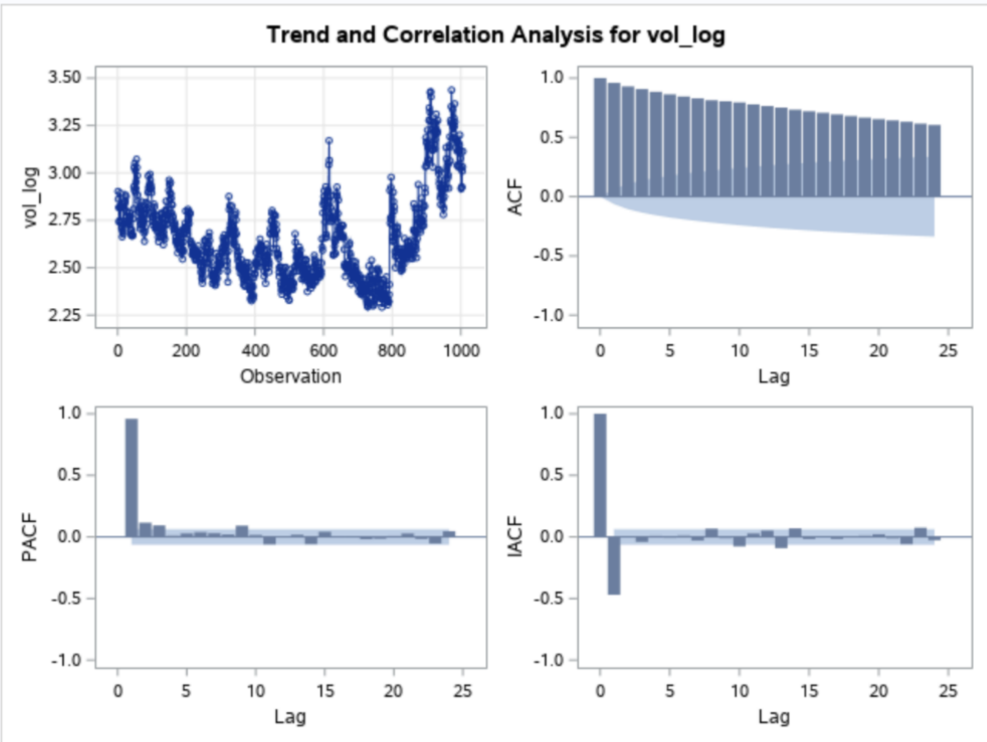


Log transformed data

We performed trend and correlation analysis on the log transformed data.

Name of Variable = vol_log	
Mean of Working Series	2.657567
Standard Deviation	0.225566
Number of Observations	1006

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4901.71	6	<.0001	0.959	0.930	0.907	0.884	0.863	0.845
12	8784.20	12	<.0001	0.829	0.813	0.804	0.794	0.778	0.764
18	9999.99	18	<.0001	0.751	0.733	0.720	0.707	0.694	0.680
24	9999.99	24	<.0001	0.666	0.654	0.643	0.631	0.615	0.604

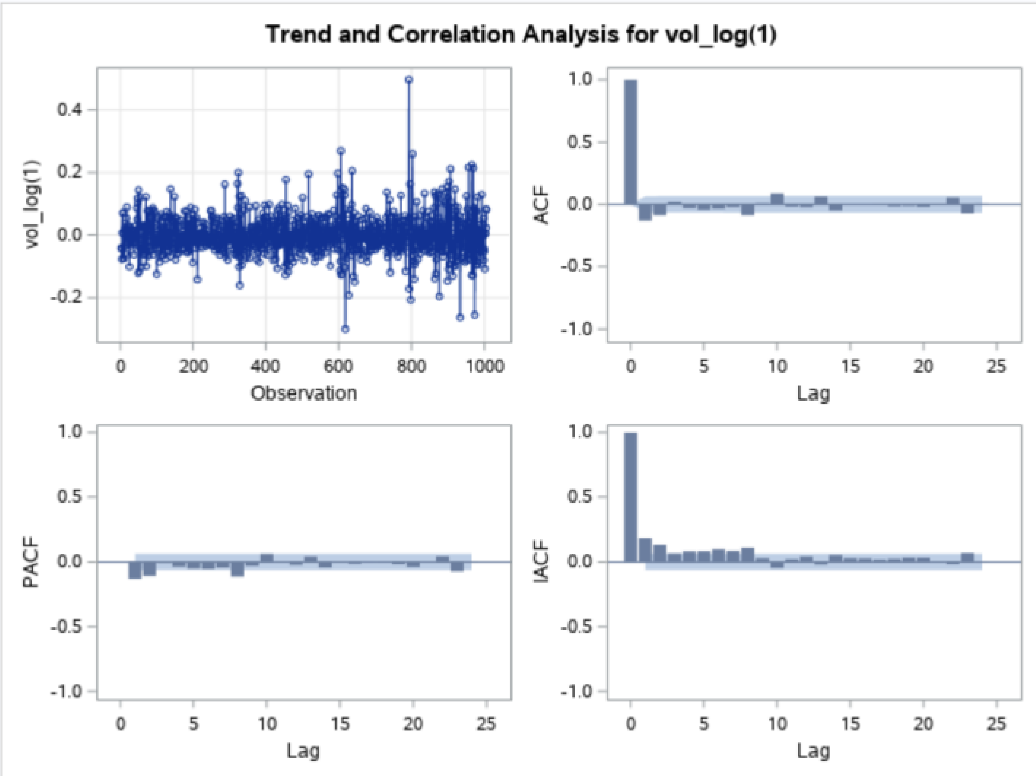


The log transformed data still has very many of the same features of the original data we now have the desirable feature of only producing positive values for volatility.

The ACF plot indicates that we need to consider differencing the data; we therefore differenced at lag one and the output was as below:

Name of Variable = vol_log	
Period(s) of Differencing	1
Mean of Working Series	0.00021
Standard Deviation	0.06226
Number of Observations	1005
Observation(s) eliminated by differencing	1

Autocorrelation Check for White Noise									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	29.35	6	<.0001	-0.131	-0.089	0.020	-0.029	-0.042	-0.034
12	46.63	12	<.0001	-0.021	-0.088	0.010	0.089	-0.017	-0.021
18	53.12	18	<.0001	0.062	-0.048	-0.000	-0.000	-0.005	-0.014
24	62.09	24	<.0001	-0.011	-0.020	0.005	0.054	-0.072	0.008



Looking at the autocorrelation check for white noise the Pr>chisq is quite low and this would mean that the white noise alone does not explain our model. There is ACF and PACF plots point to some factors at lag1, lag2, lag 6 and lag 8. What we can see is that at these lags the bands are outside the Confidence interval limit and this would mean that we could possible need some lags there.We can also see that our model does not show any significant seasonality. Hence we do not consider that in our estimated model.

We tried other factors in between but they never turned out to be good candidates. We can also see that the IACF is dampening after a certain point. This trend indicates that pursuing a MA model of differences might yield a suitable result.

4. Modeling

Fitting an ARMA(p=0,d=1,q=1) model,

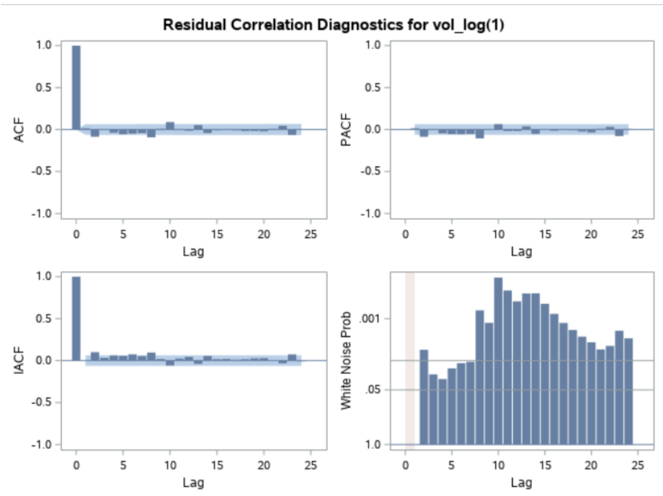
Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.0002028	0.0016299	0.12	0.9010	0
MA1,1	0.16224	0.03119	5.20	<.0001	1

Constant Estimate	0.000203
Variance Estimate	0.003802
Std Error Estimate	0.061658
AIC	-2746.1
SBC	-2736.27
Number of Residuals	1005

From the conditional least squares estimation table, the p value for the MA(1) component is very less stating that it is a significant component.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	14.75	5	0.0115	0.014	-0.086	-0.000	-0.038	-0.056	-0.049
12	33.82	11	0.0004	-0.043	-0.093	0.010	0.089	-0.005	-0.014
18	38.88	17	0.0019	0.053	-0.041	-0.007	-0.003	-0.008	-0.018
24	46.07	23	0.0029	-0.017	-0.021	0.009	0.045	-0.064	0.001
30	55.43	29	0.0022	0.015	-0.022	-0.007	0.028	-0.056	-0.066
36	57.75	35	0.0091	0.006	0.010	-0.005	-0.038	0.017	-0.018
42	60.79	41	0.0239	0.028	0.011	0.029	0.008	-0.010	-0.031
48	62.62	47	0.0633	-0.002	0.034	0.016	-0.010	-0.005	0.015

P-values of residuals are all not larger than 0.05 indicating it is not a white noise.



Also the white noise probability doesn’t seem to be good, so we proceed to fit further models.

Fitting an ARMA(p=0,d=1,q=2) model,

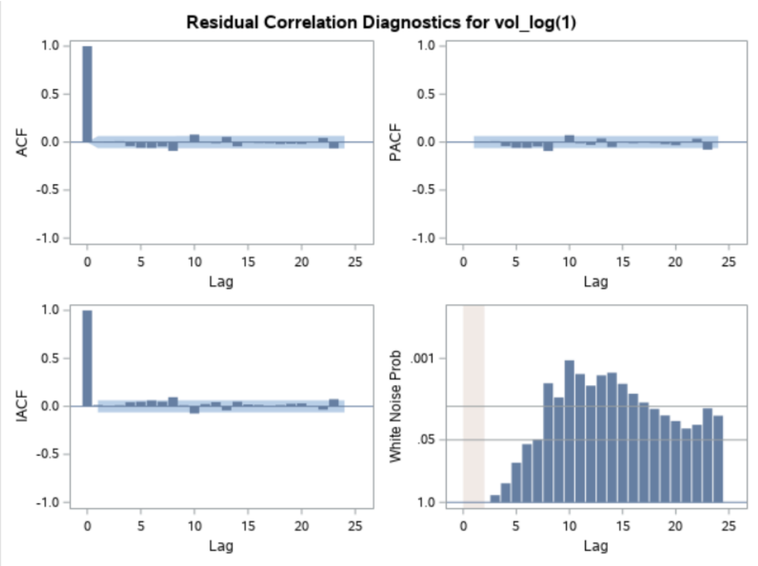
Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.0001987	0.0014518	0.14	0.8911	0
MA1,1	0.14881	0.03147	4.73	<.0001	1
MA1,2	0.10210	0.03148	3.24	0.0012	2

Constant Estimate	0.000199
Variance Estimate	0.003772
Std Error Estimate	0.061415
AIC	-2753.05
SBC	-2738.31
Number of Residuals	1005

From the conditional least squares estimation table, the p value for both MA(1) and MA(2) components are very less stating that they are significant.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	8.99	4	0.0612	0.001	0.007	0.009	-0.041	-0.059	-0.060
12	26.03	10	0.0037	-0.046	-0.091	0.004	0.079	-0.003	-0.011
18	31.60	16	0.0113	0.054	-0.044	-0.003	-0.009	-0.011	-0.020
24	38.61	22	0.0157	-0.017	-0.020	0.001	0.044	-0.065	0.004
30	47.29	28	0.0128	0.009	-0.022	-0.012	0.021	-0.057	-0.064
36	49.88	34	0.0387	0.000	-0.001	-0.003	-0.039	0.021	-0.022
42	52.92	40	0.0829	0.032	0.008	0.031	0.006	-0.007	-0.027
48	54.35	46	0.1863	-0.001	0.030	0.014	-0.006	-0.006	0.013

Not all the P-values in the above table are greater than 0.05 indicating it is not a white noise.



Also, the white noise probability does not seem to be good, so we proceed to fit further models.

Fitting an ARMA(p=0,d=1,q=(1,2,6)) model,

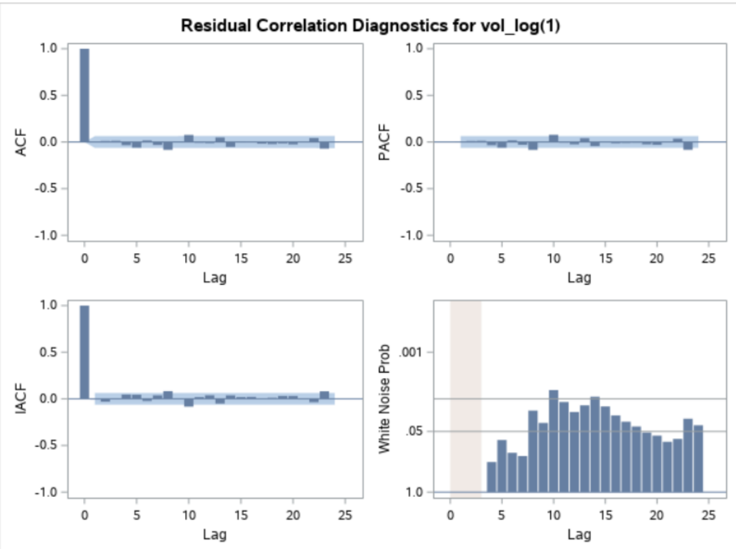
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.0002030	0.0012448	0.16	0.8705	0
MA1,1	0.16068	0.03135	5.13	<.0001	1
MA1,2	0.11478	0.03133	3.66	0.0003	2
MA1,3	0.08111	0.03096	2.62	0.0089	6

Constant Estimate	0.000203
Variance Estimate	0.003752
Std Error Estimate	0.061253
AIC	-2757.35
SBC	-2737.7
Number of Residuals	1005

From the conditional least squares estimation table, the p value for MA(1), MA(2) and MA(6) components are very less stating that they all are significant.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	5.42	3	0.1433	0.003	0.011	0.013	-0.034	-0.060	0.016
12	19.81	9	0.0191	-0.032	-0.084	0.007	0.076	-0.008	-0.012
18	25.89	15	0.0391	0.049	-0.052	-0.005	-0.002	-0.018	-0.022
24	33.94	21	0.0368	-0.014	-0.025	-0.003	0.042	-0.072	-0.004
30	43.75	27	0.0219	0.006	-0.025	-0.014	0.019	-0.061	-0.067
36	46.48	33	0.0599	0.002	-0.004	-0.003	-0.039	0.015	-0.029
42	49.71	39	0.1169	0.032	0.010	0.032	0.001	-0.006	-0.030
48	51.07	45	0.2475	-0.000	0.029	0.016	-0.011	-0.005	0.009

Not all the P-values in the above table are greater than 0.05 indicating it is not a white noise.



Also the white noise probability doesn't seem to be good but better than the previous models, so we proceed to fit further models.

Fitting an ARMA(p=0,d=1,q=(1,2,4,8)) model,

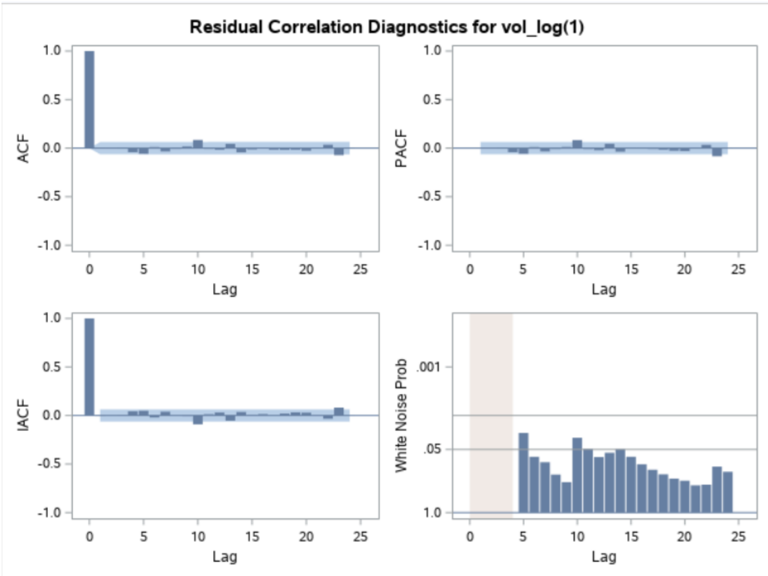
Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.0002183	0.0011265	0.19	0.8464	0
MA1,1	0.15939	0.03126	5.10	<.0001	1
MA1,2	0.09577	0.03137	3.05	0.0023	2
MA1,3	0.07853	0.03114	2.52	0.0118	6
MA1,4	0.08299	0.03127	2.65	0.0081	8

Constant Estimate	0.000218
Variance Estimate	0.003731
Std Error Estimate	0.061086
AIC	-2761.86
SBC	-2737.3
Number of Residuals	1005

From the conditional least squares estimation table, the p value for MA(1), MA(2), MA(6) and MA(8) components are very less stating that they all are significant.

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	5.30	2	0.0708	0.001	-0.001	0.005	-0.042	-0.058	0.011
12	14.41	8	0.0717	-0.034	-0.007	0.017	0.085	-0.007	-0.016
18	19.06	14	0.1628	0.044	-0.043	-0.012	-0.001	-0.015	-0.017
24	26.69	20	0.1442	-0.016	-0.026	-0.003	0.035	-0.072	-0.005
30	35.84	26	0.0946	0.006	-0.030	-0.013	0.017	-0.058	-0.064
36	38.75	32	0.1914	0.000	-0.003	-0.005	-0.044	0.014	-0.026
42	41.96	38	0.3033	0.029	0.003	0.030	0.001	-0.009	-0.035
48	43.17	44	0.5070	-0.001	0.021	0.019	-0.013	-0.003	0.013

All the P-values in the above table are greater than 0.05 indicating it is a white noise process.



Also, the white noise probability seems to be good and better than the previous models. But further we try to fit more models to get better AIC.

Fitting an ARMA(p=0,d=1,q=(1,2,4,8,10)),

An outstanding candidate for a model was an ARMA (p=0,d=1,q=(1, 2, 6, 8)). Other models just did not turn out to be a proper fit.

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MU	0.0001968	0.0012779	0.15	0.8777	0
MA1,1	0.15848	0.03121	5.08	<.0001	1
MA1,2	0.09847	0.03142	3.13	0.0018	2
MA1,3	0.07364	0.03109	2.37	0.0180	6
MA1,4	0.08591	0.03140	2.74	0.0063	8
MA1,5	-0.08098	0.03125	-2.59	0.0097	10

Constant Estimate	0.000197
Variance Estimate	0.003712
Std Error Estimate	0.060922
AIC	-2766.25
SBC	-2736.77
Number of Residuals	1005

* AIC and SBC do not include log determinant.

Correlations of Parameter Estimates						
Parameter	MU	MA1,1	MA1,2	MA1,3	MA1,4	MA1,5
MU	1.000	-0.002	-0.003	-0.002	0.001	0.001
MA1,1	-0.002	1.000	-0.168	0.026	0.003	-0.016
MA1,2	-0.003	-0.168	1.000	0.027	-0.073	-0.082
MA1,3	-0.002	0.026	0.027	1.000	-0.126	0.035
MA1,4	0.001	0.003	-0.073	-0.126	1.000	-0.116
MA1,5	0.001	-0.016	-0.082	0.035	-0.116	1.000

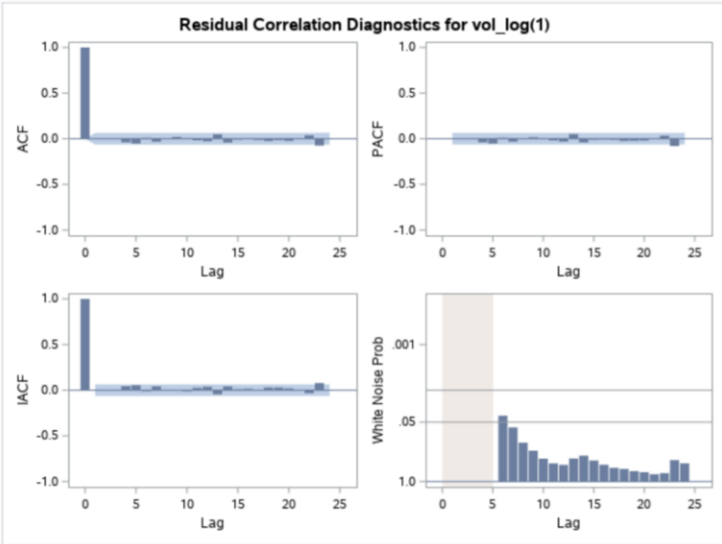
Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	4.39	1	0.0361	0.000	0.004	0.005	-0.038	-0.052	0.010
12	6.97	7	0.4315	-0.032	-0.001	0.019	0.008	-0.017	-0.028
18	11.95	13	0.5317	0.048	-0.041	-0.009	-0.005	-0.012	-0.025
24	19.89	19	0.4010	-0.014	-0.023	-0.002	0.038	-0.075	0.001
30	28.92	25	0.2673	0.005	-0.029	-0.015	0.018	-0.059	-0.061
36	31.61	31	0.4357	0.001	-0.004	0.001	-0.044	0.013	-0.021
42	35.11	37	0.5581	0.031	0.000	0.037	0.008	-0.006	-0.030
48	36.81	43	0.7356	-0.002	0.027	0.020	-0.014	-0.009	0.015

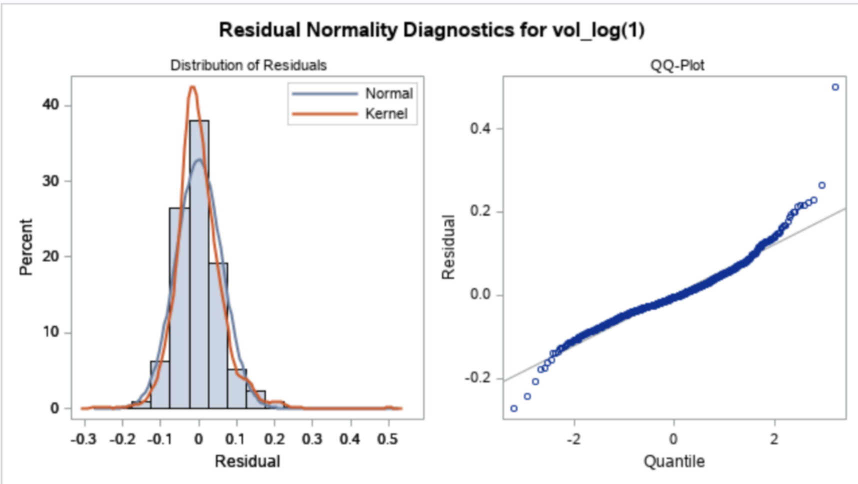
Above we can see that the ARMA (p=0,d=1,q=(1,2,4,8,10)) model has fit quite well on our Log transform data. Looking at the individual p-values for the Conditional Least Squares estimates we can observe that the p-value for all the parameters are <0.05. That means that all the MA coefficients have fit quite well.

Going ahead what we see is the AIC is the lowest among all the other models that we fit.

AIC = -2(log-likelihood) + 2K

We go ahead and look at the Autocorrelation check of Residuals. What we see is that the p value is greater than 0.05 and that shows that white noise test plots show that you cannot reject the hypothesis that the residuals are uncorrelated and hence we can conclude the model is adequate.





Model for variable vol_log	
Estimated Mean	0.000197
Period(s) of Differencing	1

Moving Average Factors	
Factor 1:	1 - 0.15848 B**(1) - 0.09847 B**(2) - 0.07364 B**(6) - 0.08591 B**(8) + 0.08098 B**(10)

Thus, our final model is

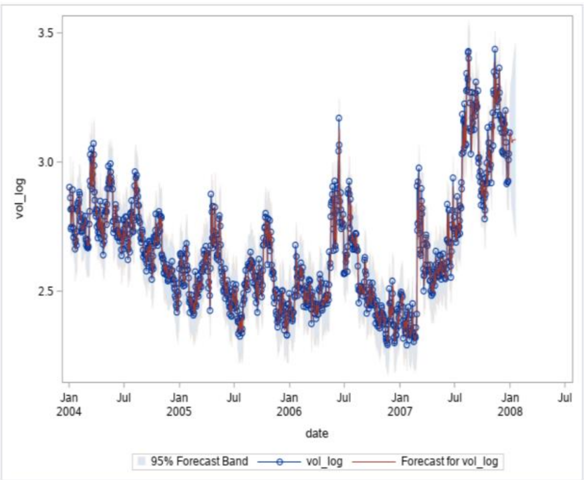
$$1-0.1548B - 0.0984 B^2 - 0.07364B^6 - 0.08591B^8 + 0.08098B^{10}$$

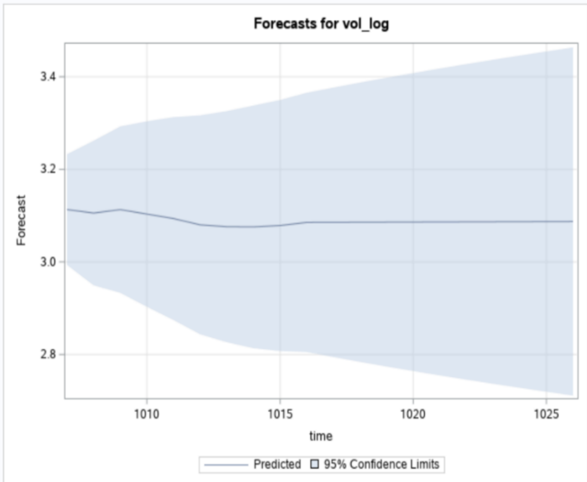
Model	**AIC**
ARMA(p=0, d=1,q=1)	-2745.1
ARMA(p=0, d=1,q=(1,2))	-2753.05
ARMA(p=0, d=1,q=(1,2,6))	-2757.35
ARMA(p=0, d=1,q=(1,2.6.8))	-2761.86
ARMA(p=0,d=1,q=(1,2,6,8,10))	-2766.25

5. Forecasting

Obs	Forecast	Std Error	95% Confidence Limits	
1007	3.1130	0.0609	2.9936	3.2325
1008	3.1055	0.0796	2.9495	3.2616
1009	3.1129	0.0916	2.9334	3.2924
1010	3.1034	0.1022	2.9031	3.3036
1011	3.0936	0.1117	2.8746	3.3126
1012	3.0801	0.1206	2.8437	3.3164
1013	3.0761	0.1273	2.8266	3.3255
1014	3.0756	0.1337	2.8137	3.3376
1015	3.0786	0.1383	2.8075	3.3497
1016	3.0853	0.1428	2.8054	3.3652
1017	3.0855	0.1484	2.7946	3.3764
1018	3.0857	0.1538	2.7842	3.3872
1019	3.0859	0.1591	2.7741	3.3977
1020	3.0861	0.1642	2.7644	3.4078
1021	3.0863	0.1691	2.7549	3.4176
1022	3.0865	0.1738	2.7457	3.4272
1023	3.0867	0.1785	2.7368	3.4365
1024	3.0869	0.1830	2.7281	3.4456
1025	3.0871	0.1875	2.7197	3.4545
1026	3.0873	0.1918	2.7114	3.4631

Forecasts for variable vol_log





We can see the Volatility log transformed and the forecasted value (red) in the figure above. To get the original Vix forecasts, we need to apply the inverse of the log function. The figure below shows the actual v/s forecasted volatility. The predicted figures are below the actuals but within the bands, this shows that we may need to tweak our model or use a different one.

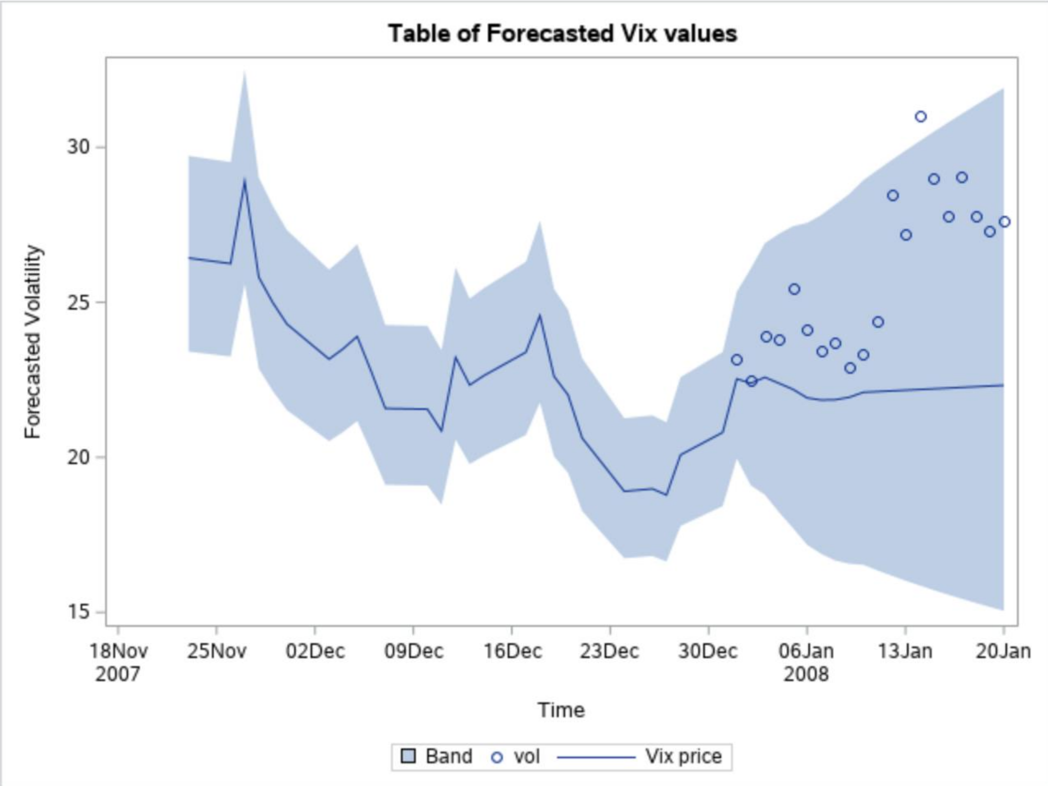


Table of Forecasted Vix values

Vix price	Lower 95% confidence limit	Upper 95% confidence limit
22.1206	16.35598	29.265109
22.1656	16.02415	29.894654
20.62378	18.268596	23.196341
22.58128	18.791606	26.908525
22.63269	20.048094	25.455837
23.50136	20.817563	26.432863
18.786046	16.640726	21.129372
18.903919	16.745139	21.261948
18.985391	16.817307	21.353583
20.079183	17.78619	22.583812
20.804441	18.428625	23.399536
20.851948	18.470708	23.45297
21.552899	19.091612	24.241355
21.577627	19.113516	24.269168
21.849182	16.887986	27.813588
21.858031	16.671208	28.151221
21.918392	17.180051	27.560035
21.936886	16.569195	28.493249
22.012505	19.498732	24.758292
22.098134	16.534119	28.938416
22.143089	16.186286	29.58356
22.188135	15.868831	30.199134
22.193998	17.717586	27.456384
22.210693	15.719699	30.497631
22.233273	15.576214	30.790684

SAS Code

```
*Reading in Data;
data vix;
	infile '/home/u48908401/sasuser.v94/Vix_Final.csv' firstobs=2 dsd missover;
	input date : mmddyy10. vol obs;
	format date ddmmyy10.;
	vol_log = log(vol);
run;

title "Univariate Analysis";
proc univariate data = vix;
  histogram vol_log / normal kernel;
  inset mean std normal / position = ne;
run;
title;

proc sgplot data = vix;
  series x = date y = vol;
run;

*looking for stationarity using Dickey Fuller test;
proc arima data = vix;
identify var = vol_log(1) stationarity=(adf=1);
run;

*Fitting a differenced model;
proc arima data = vix;
/* identify var = vol; */
/* identify var = vol_log; */
identify var = vol_log(1);
estimate q = (1);
estimate q = (1,2);
estimate q = (1,2,6);
estimate q = (1,2,6,8);
estimate q = (1,2,6,8,10);
forecast lead = 20 id = date out =out1;
outlier;
run;

*Plotting Log volatility vs forecasted Log vols;
proc sgplot data = work.out1;
band x =date lower = l95 upper = u95/
legendlabel="95% Forecast Band" fillattrs=graphconfidence
				transparency=0.5 fill outline;
series x = date y = vol_log / markers;
series x = date y = forecast / lineattrs=graphdata2;
run;

proc transreg data = vix;
```

```
model boxcox(vol) = identity(date);
run;

*Dataset made fror plotting volatility and Forecasted volatility;
data plot;
set out1;
y=exp(vol_log);
l95=exp( l95 );
u95=exp( u95 );
forecast = exp(forecast+ std*std/ 2 );
obs = _N_;
if _N_ > 980 ; *Number from where you would like to see the graph;
run ;

proc sgplot data =plot noautolegend ;
scatter x = obs y =forecast / yerrorlower =L95 yerrorupper =U95;
series x = obs y =forecast;
yaxis label = 'Forecasted Volatility' ;
xaxis label = 'Obsv number' ;
Run ;


data plot;
set plot;
label Forecast ="Vix price"
L95 = "Lower 95% confidence limit"
U95 = "Upper 95% confidence limit"
;
run;
proc report data=plot;
title1 "Table of Forecasted Vix values";
column Forecast L95 U95;
define forecast / order;
run;

data plot2;
  merge plot vix_test;
  by date;
  format date ddmmyy10.;
run;

proc sgplot data =plot2 ;
        band x =date lower = l95 upper = u95;
        scatter x = date y =vol ;
        *series x = date y =vol;
        series x = date y =forecasted;
        yaxis label = 'Forecasted Volatility' ;
        xaxis label = 'Time' ;
Run ;
```