**STAT 3301 Time Series Project**

Monthly measurements of carbon dioxide above Mauna Loa, Hawaii from Jan 1959 to Dec 1990

**HU CHENCHEN**

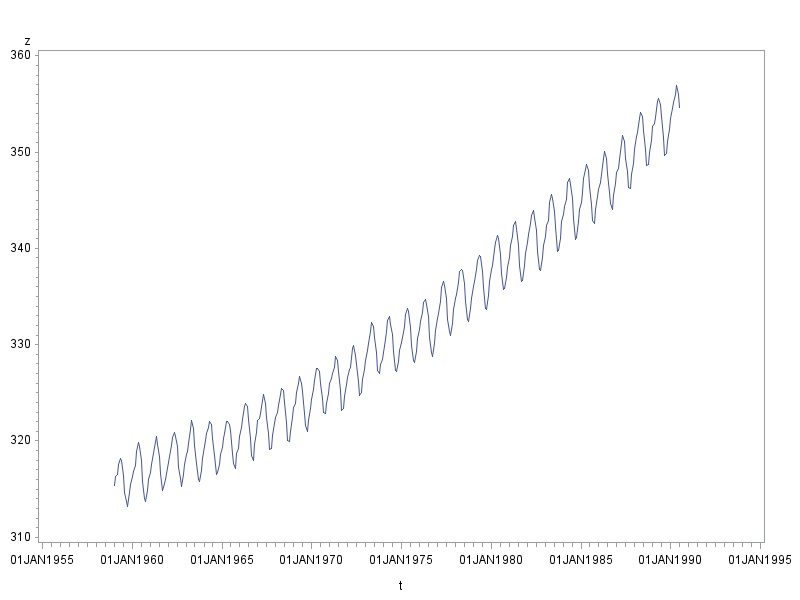
**3035021429**

1. **Introduction**

The carbon dioxide concentration in the atmosphere is closely related with the earth’s surface temperature regulation and ecological cycle. Furthermore, it’s the major contributor to greenhouse effect. Therefore, monitoring the carbon dioxide has significant meanings on the scientific study of present-day biosphere. In this report, the monthly measurements of carbon dioxide above Mauna Loa, Hawaii from Jan 1959 to Dec 1990 is studied by adopting a time series model, which could bring some insights to the analysis of global temperature and global warming issue.

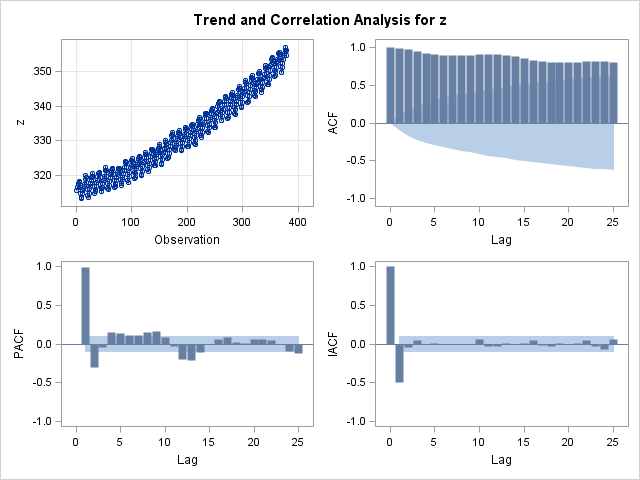
1. **Model Identification**
   1. **Time plot**

As shown in the time plot (the vertical axis represents the monthly carbon dioxide in ppm), the sample mean shows a significant increasing trend over the past 30 years, while the sample variance remains relevantly constant. Meantime, there’s a noticeable periodic pattern shown in the plot, which is consistent with the monthly nature of the data. These observations indicate a seasonal ARIMA model without log transformation.



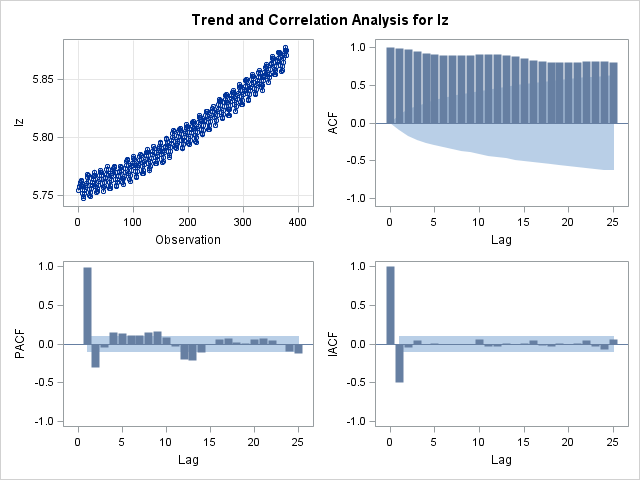
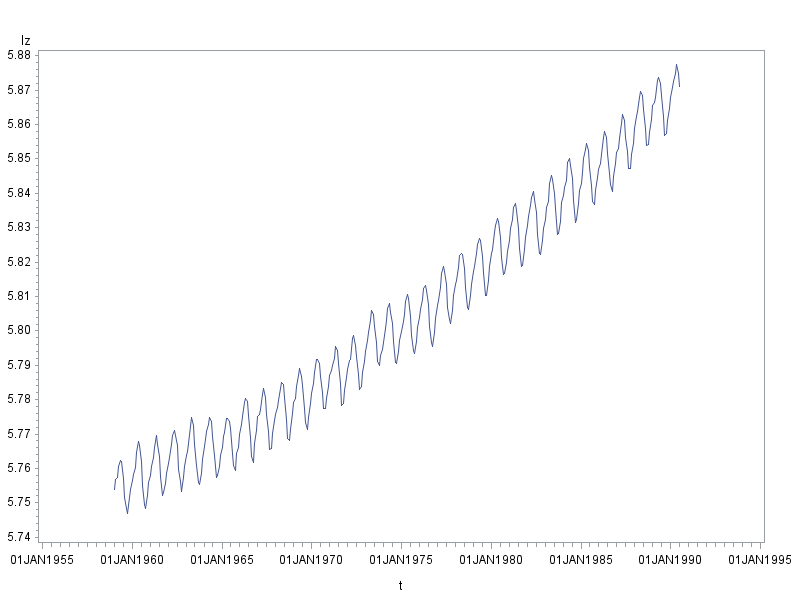
* 1. **Check Stationarity**

ARIMA model is applied to the raw data. The sample ACF decays slowly over all periods, which suggests that the raw data is non-stationary and transformation is needed.



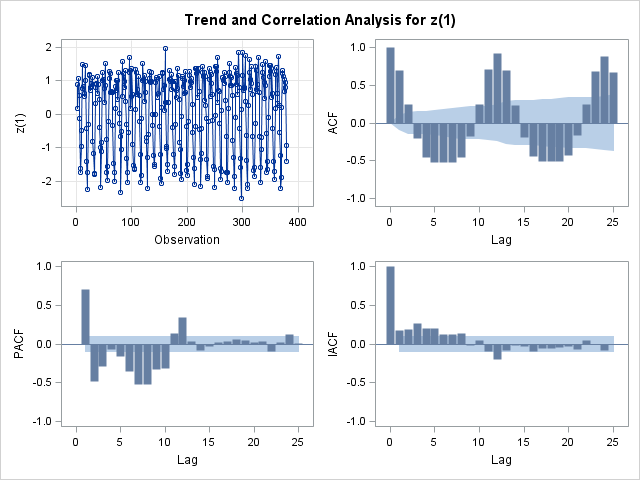
* 1. **Log Transformation**

Firstly, to check the possibility of inconstant variance, a log transformation is applied to the raw data. As shown in the time plot, the transformed data has a decreasing variance, and there’s no significant changes in the sample ACF. In this case, the log transformation is ruled out.



* 1. **First Difference (Normal Difference)**

Compared with sample ACF historgram of raw data, the ACF at non-seasonal lags decrease sharply. However, the ACF around seasonal lags are still significant and decay slowly. This indicates we still need to take seasonal difference.

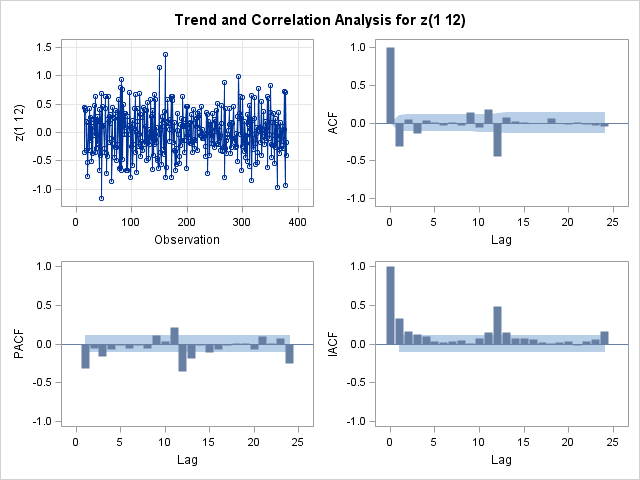


* 1. **Seasonal Difference**

From the perspective of sample ACF, there’s a clear cut-off pattern after lat =1 and lag =12. Meanwhile, there’s a noticable ACF at lag = 11. These features all consist with ARMA(0,1) X (0,1)12 model. This result verifies the previous conjecture of a seasonal model based on the monthly nature of the raw data.

From the perspective of sample PACF, it’s hard to determine a multiplicative seasonal ARMA model. Only pure seasonal models with no MA or AR components at common lags have recognizable patterns. Judging from the significant ACF at lag =1, the data is not likely to follow a pure seasonal model. Therefore the PACF histogram is not very informative in this case.

In sum, ARIMA(0,1,1) X (0,1,1)12 model is chosen.



* 1. **Augmented Dickey-Fuller unit root test**

To further verity the stationarity of the differenced data, the Augmented Dickey-Fuller unit root tests at lag = 1 to 4 are conducted. As shown in the table, The P-values of 3 test statistics at all time lags are significant, which indicates that after taking the common and seasonal difference, the time series is stationary.

| **Augmented Dickey-Fuller Unit Root Tests** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Type** | **Lags** | **Rho** | **Pr < Rho** | **Tau** | **Pr < Tau** | **F** | **Pr > F** |
| **Zero Mean** | **0** | -478.047 | 0.0001 | -26.30 | <.0001 |  |  |
|  | **1** | -536.675 | 0.0001 | -16.31 | <.0001 |  |  |
|  | **2** | -1006.23 | 0.0001 | -14.50 | <.0001 |  |  |
|  | **3** | -1960.44 | 0.0001 | -12.38 | <.0001 |  |  |
|  | **4** | -2595.62 | 0.0001 | -10.44 | <.0001 |  |  |
| **Single Mean** | **0** | -478.048 | 0.0001 | -26.27 | <.0001 | 345.03 | 0.0010 |
|  | **1** | -536.681 | 0.0001 | -16.29 | <.0001 | 132.67 | 0.0010 |
|  | **2** | -1006.17 | 0.0001 | -14.48 | <.0001 | 104.82 | 0.0010 |
|  | **3** | -1959.01 | 0.0001 | -12.36 | <.0001 | 76.44 | 0.0010 |
|  | **4** | -2591.31 | 0.0001 | -10.42 | <.0001 | 54.33 | 0.0010 |
| **Trend** | **0** | -478.089 | 0.0001 | -26.24 | <.0001 | 344.22 | 0.0010 |
|  | **1** | -536.822 | 0.0001 | -16.27 | <.0001 | 132.35 | 0.0010 |
|  | **2** | -1007.02 | 0.0001 | -14.46 | <.0001 | 104.54 | 0.0010 |
|  | **3** | -1959.93 | 0.0001 | -12.34 | <.0001 | 76.21 | 0.0010 |
|  | **4** | -2593.17 | 0.0001 | -10.41 | <.0001 | 54.16 | 0.0010 |

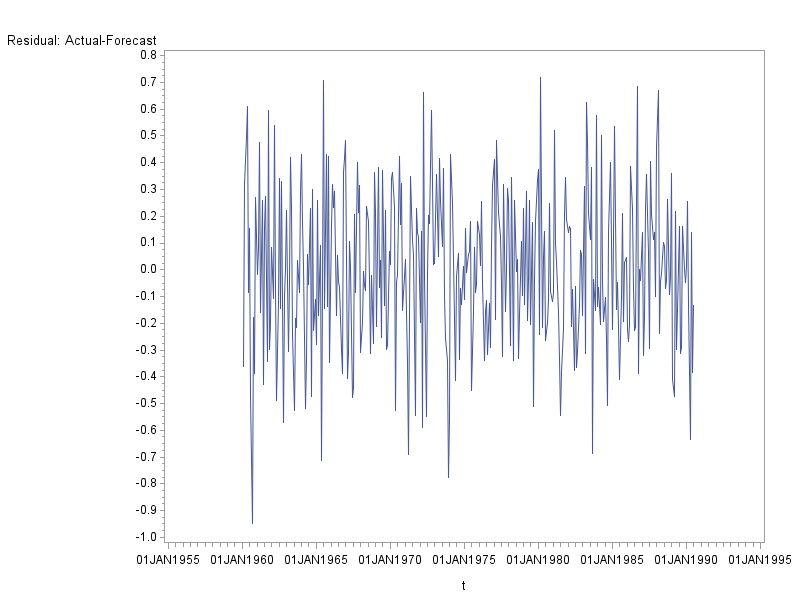
1. **Parameter Estimation**

To estimate the value of model coefficients and the variance, the ML method is adopted. Judging from the result, the coefficients of both common and seasonal factors are significant. The intercept is insignificant (P-value = 0.07), which indicates the mean of the model is zero.

| **Maximum Likelihood Estimation** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Approx Pr > |t|** | **Lag** |
| **MU** | 0.0027239 | 0.0015151 | 1.80 | 0.0722 | 0 |
| **MA1,1** | 0.37376 | 0.04781 | 7.82 | <.0001 | 1 |
| **MA2,1** | 0.87634 | 0.03211 | 27.29 | <.0001 | 12 |

|  |  |
| --- | --- |
| **Constant Estimate** | 0.002724 |
| **Variance Estimate** | 0.080389 |
| **Std Error Estimate** | 0.28353 |
| **AIC** | 136.6889 |
| **SBC** | 148.3968 |
| **Number of Residuals** | 366 |

1. **Diagnostic Checking** 
   1. **Time Plot of residuals**

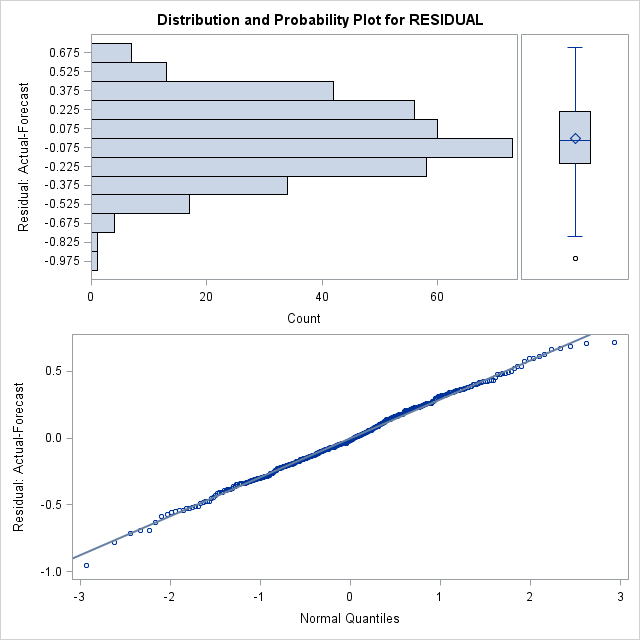
The time plot shows no discernable linear trend or variation in residual variance. The residual behaves similarly to the white noise.

* 1. **Normality Checking**

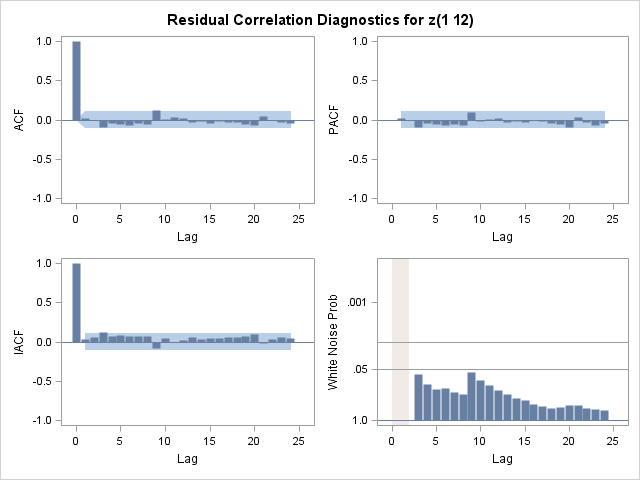
From an intuitive perspective, both the distribution histogram of residuals and the Q-Q plot suggests that the residuals fits the normal distribution quite well.

From an inferential perspective, the P-values of all four test statistics are insignificant. Therefore the normality assumption is accepted. We can safely conclude that the residuals follow normal distribution.

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.996824 | **Pr < W** | 0.6937 |
| **Kolmogorov-Smirnov** | **D** | 0.032653 | **Pr > D** | >0.1500 |
| **Cramer-von Mises** | **W-Sq** | 0.05197 | **Pr > W-Sq** | >0.2500 |
| **Anderson-Darling** | **A-Sq** | 0.298443 | **Pr > A-Sq** | >0.2500 |
|  |  |  |  |  |

****

* 1. **Residual ACF & PACF and Ljung-Box Test**

The sample ACF and PACF of residuals have typical characteristics of white noise since neither ACF nor PACF at any time lags are significant.

To further check whether the residuals contain any informative components such as AR or MA pattern, Ljung-Box tests at lag = 6 to 48 (step = 6) are conducted. The result consists with the graphical analysis. All P-values are insignificant, which consolidates the assumption that the residuals have no further information about the time series.

| **Autocorrelation Check of Residuals** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **To Lag** | **Chi-Square** | **DF** | **Pr > ChiSq** | **Autocorrelations** | | | | | |
| **6** | 6.66 | 4 | 0.1547 | 0.022 | -0.011 | -0.092 | -0.047 | -0.050 | -0.064 |
| **12** | 13.96 | 10 | 0.1747 | -0.045 | -0.050 | 0.117 | 0.002 | 0.030 | 0.017 |
| **18** | 15.60 | 16 | 0.4812 | -0.030 | -0.018 | -0.038 | -0.013 | -0.028 | -0.026 |
| **24** | 20.70 | 22 | 0.5393 | -0.055 | -0.070 | 0.051 | -0.006 | -0.030 | -0.040 |
| **30** | 25.59 | 28 | 0.5955 | 0.039 | -0.008 | 0.039 | 0.006 | -0.041 | -0.086 |
| **36** | 32.58 | 34 | 0.5371 | -0.059 | -0.045 | 0.027 | 0.095 | 0.039 | 0.021 |
| **42** | 44.35 | 40 | 0.2934 | 0.073 | 0.027 | -0.105 | -0.038 | -0.053 | -0.085 |
| **48** | 47.43 | 46 | 0.4140 | -0.026 | 0.043 | 0.014 | 0.050 | 0.046 | 0.000 |

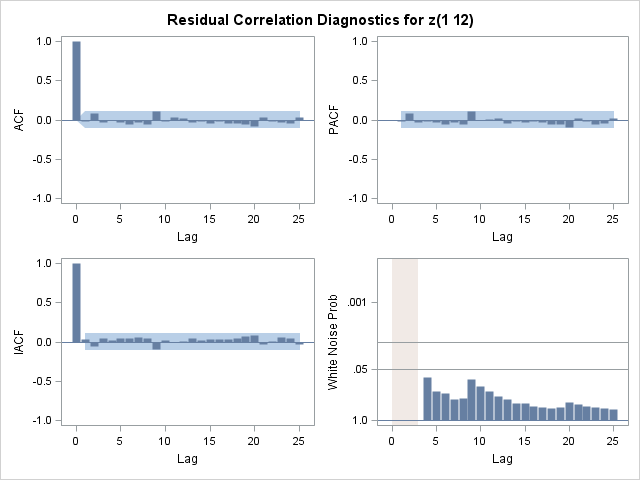
* 1. **Model Overfitting**
     1. ARIMA (1,1,1) X (0,1,1)12 model

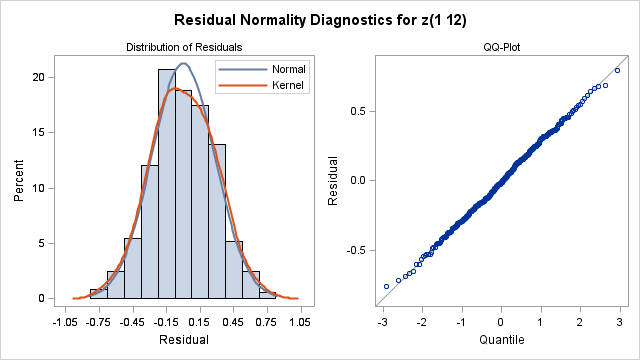
As shown in the table, the AR coefficient is significant at 0.01 level. Furthermore, the ACI of ARIMA (1,1,1) X (0,1,1)12 model (135.398) is smaller than the previous model (136.6889). This suggests that the previous model is not sufficient compared with new model.

| **Maximum Likelihood Estimation** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Approx Pr > |t|** | **Lag** |
| **MU** | 0.0027625 | 0.0011371 | 2.43 | 0.0151 | 0 |
| **MA1,1** | 0.67192 | 0.08655 | 7.76 | <.0001 | 1 |
| **MA2,1** | 0.88460 | 0.03204 | 27.61 | <.0001 | 12 |
| **AR1,1** | 0.32846 | 0.11042 | 2.97 | 0.0029 | 1 |

|  |  |
| --- | --- |
| **Constant Estimate** | 0.001855 |
| **Variance Estimate** | 0.079703 |
| **Std Error Estimate** | 0.282318 |
| **AIC** | 135.398 |
| **SBC** | 151.0086 |
| **Number of Residuals** | 366 |

To check the adequacy of the new model, the diagnostic checking is conducted as shown in the following charts. The distribution histogram and Q-Q plot show that residuals fit the normal distribution. Meanwhile, the sample ACF and PACF of residuals have clear cut-off pattern after lag = 0, which resembles the behavior of the white noise. The Ljung-Box test also corroborates the observation that sample ACF from lag = 6 to 48 (at step = 6) are insignificant. These conclusions indicate that ARIMA (1,1,1) X (0,1,1)12 is an plausible alternative model.





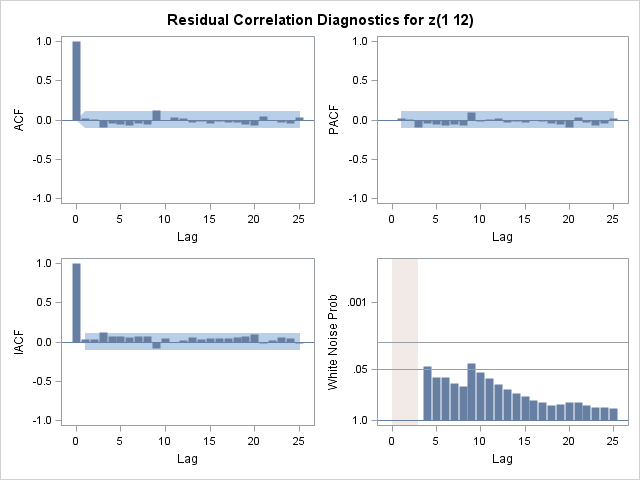
| **Autocorrelation Check of Residuals** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **To Lag** | **Chi-Square** | **DF** | **Pr > ChiSq** | **Autocorrelations** | | | | | |
| **6** | 4.59 | 3 | 0.2048 | -0.023 | 0.082 | -0.032 | -0.008 | -0.032 | -0.055 |
| **12** | 11.45 | 9 | 0.2462 | -0.031 | -0.060 | 0.111 | -0.017 | 0.027 | 0.017 |
| **18** | 14.54 | 15 | 0.4849 | -0.033 | -0.019 | -0.049 | -0.023 | -0.046 | -0.040 |
| **24** | 20.46 | 21 | 0.4925 | -0.060 | -0.084 | 0.038 | -0.023 | -0.029 | -0.042 |
| **30** | 24.83 | 27 | 0.5837 | 0.036 | -0.024 | 0.025 | -0.007 | -0.042 | -0.081 |
| **36** | 30.30 | 33 | 0.6021 | -0.047 | -0.041 | 0.020 | 0.087 | 0.035 | 0.021 |
| **42** | 41.22 | 39 | 0.3739 | 0.066 | 0.027 | -0.103 | -0.029 | -0.052 | -0.085 |
| **48** | 44.45 | 45 | 0.4949 | -0.029 | 0.038 | 0.006 | 0.050 | 0.053 | 0.003 |

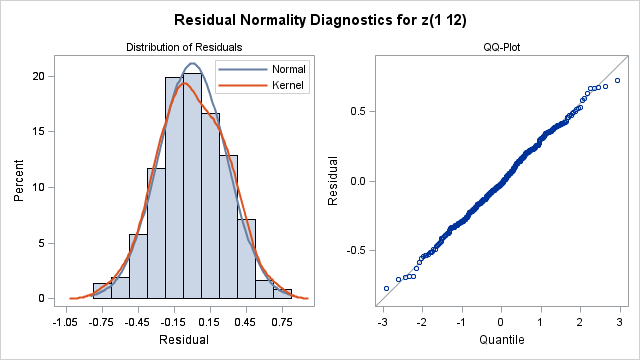
* + 1. ARIMA (0,1,2) X (0,1,1)12 model

The table shows that the newly added MA factor at lag = 2 in the non-seasonal component is insignificant, while MA1,1 and MA3,1 factors remain relatively unchanged. Meanwhile, AIC also increases from 136.6889 to 138.6182. In the diagnostic checking part, although the Q-Q plot suggests a satisfactory fitting to normal distribution, the Ljung-Box test at lag = 6 is significant at 0.1 level. This evidence indicates ARIMA (0,1,2) X (0,1,1)12 model is not adequate, and in the meanwhile prove the adequacy of the previous model.

| **Maximum Likelihood Estimation** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Approx Pr > |t|** | **Lag** |
| **MU** | 0.0027233 | 0.0015055 | 1.81 | 0.0705 | 0 |
| **MA1,1** | 0.36946 | 0.05120 | 7.22 | <.0001 | 1 |
| **MA2,1** | 0.01537 | 0.05471 | 0.28 | 0.7788 | 2 |
| **MA3,1** | 0.87629 | 0.03215 | 27.25 | <.0001 | 12 |

|  |  |
| --- | --- |
| **Constant Estimate** | 0.002723 |
| **Variance Estimate** | 0.080597 |
| **Std Error Estimate** | 0.283895 |
| **AIC** | 138.6182 |
| **SBC** | 154.2287 |
| **Number of Residuals** | 366 |



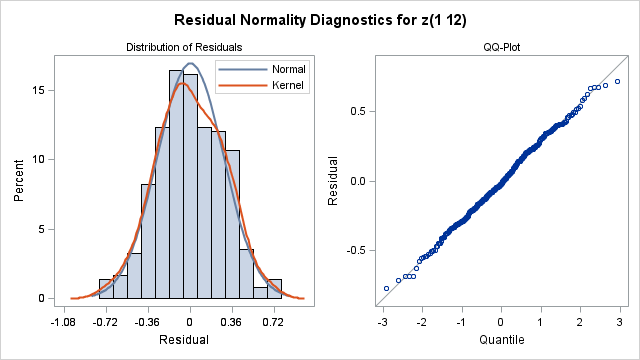


| **Autocorrelation Check of Residuals** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **To Lag** | **Chi-Square** | **DF** | **Pr > ChiSq** | **Autocorrelations** | | | | | |
| **6** | 6.72 | 3 | 0.0815 | 0.016 | 0.003 | -0.093 | -0.048 | -0.051 | -0.065 |
| **12** | 14.03 | 9 | 0.1212 | -0.043 | -0.051 | 0.117 | 0.000 | 0.031 | 0.017 |
| **18** | 15.73 | 15 | 0.4000 | -0.030 | -0.018 | -0.039 | -0.013 | -0.029 | -0.027 |
| **24** | 20.77 | 21 | 0.4730 | -0.054 | -0.070 | 0.051 | -0.008 | -0.029 | -0.040 |
| **30** | 25.68 | 27 | 0.5366 | 0.039 | -0.009 | 0.039 | 0.005 | -0.041 | -0.087 |
| **36** | 32.62 | 33 | 0.4857 | -0.059 | -0.045 | 0.026 | 0.095 | 0.039 | 0.022 |
| **42** | 44.32 | 39 | 0.2573 | 0.072 | 0.027 | -0.104 | -0.038 | -0.054 | -0.085 |
| **48** | 47.45 | 45 | 0.3731 | -0.026 | 0.042 | 0.014 | 0.051 | 0.047 | 0.001 |

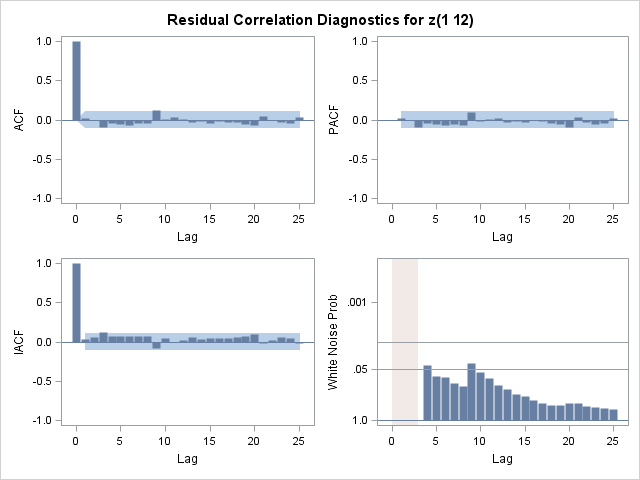
* + 1. ARIMA (0,1,1) X (1,1,1)12 model

The table shows that the newly added AR factor at lag = 12 in the seasonal component is insignificant, while MA1,1 and MA2,1 remain relatively unchanged. Meanwhile, AIC increases from 136.6889 to 138.6576. In the diagnostic checking part, although the Q-Q plot suggests a satisfactory fitting to normal distribution, the Ljung-Box test at lag = 6 is significant at 0.1 level. This evidence indicates ARIMA (0,1,1) X (1,1,1)12 model is not adequate, and in the meanwhile prove the adequacy of the previous model.

| **Maximum Likelihood Estimation** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Approx Pr > |t|** | **Lag** |
| **MU** | 0.0027219 | 0.0015149 | 1.80 | 0.0724 | 0 |
| **MA1,1** | 0.37263 | 0.04792 | 7.78 | <.0001 | 1 |
| **MA2,1** | 0.87925 | 0.03660 | 24.02 | <.0001 | 12 |
| **AR1,1** | 0.01076 | 0.06076 | 0.18 | 0.8595 | 12 |



|  |  |  |
| --- | --- | --- |
| **Constant Estimate** | | 0.002693 |
| **Variance Estimate** | | 0.080596 |
| **Std Error Estimate** | | 0.283894 |
| **AIC** | | 138.6576 |
| **SBC** | | 154.2681 |
| **Number of Residuals** | 366 |



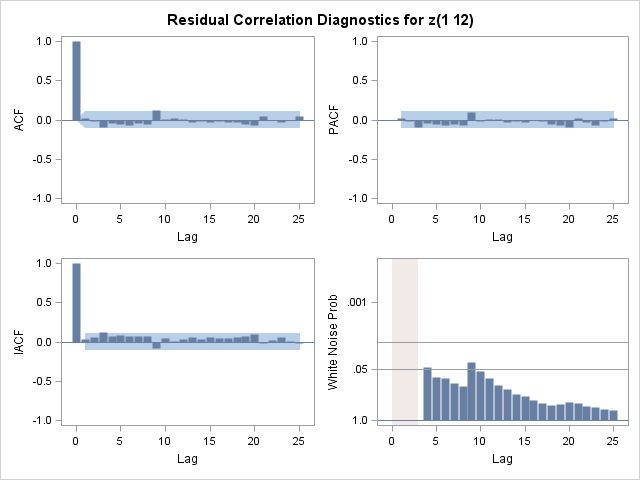
| **Autocorrelation Check of Residuals** | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **To Lag** | **Chi-Square** | **DF** | **Pr > ChiSq** | **Autocorrelations** | | | | | |
| **6** | 6.72 | 3 | 0.0815 | 0.022 | -0.010 | -0.093 | -0.047 | -0.050 | -0.065 |
| **12** | 13.94 | 9 | 0.1243 | -0.044 | -0.049 | 0.117 | 0.002 | 0.030 | 0.010 |
| **18** | 15.54 | 15 | 0.4132 | -0.029 | -0.017 | -0.037 | -0.013 | -0.028 | -0.025 |
| **24** | 20.47 | 21 | 0.4918 | -0.054 | -0.069 | 0.050 | -0.006 | -0.030 | -0.038 |
| **30** | 25.39 | 27 | 0.5526 | 0.039 | -0.008 | 0.040 | 0.006 | -0.041 | -0.086 |
| **36** | 32.38 | 33 | 0.4979 | -0.059 | -0.045 | 0.026 | 0.095 | 0.039 | 0.023 |
| **42** | 44.15 | 39 | 0.2629 | 0.072 | 0.027 | -0.104 | -0.039 | -0.054 | -0.085 |
| **48** | 47.21 | 45 | 0.3823 | -0.026 | 0.043 | 0.014 | 0.050 | 0.045 | 0.001 |

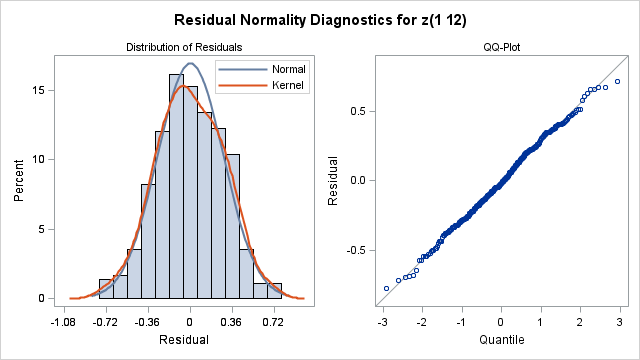
* + 1. ARIMA (0,1,1) X (0,1,2)12 model

The table shows that the newly added AR factor at lag = 24 in the seasonal component is insignificant, while MA1,1 and MA2,1 factors remain relatively unchanged. Meanwhile, AIC increases from 136.6889 to 137.9753. In the diagnostic checking part, although the Q-Q plot suggests a satisfactory fitting to normal distribution, the Ljung-Box test at lag = 6 is significant at 0.1 level. This evidence indicates ARIMA (0,1,1) X (0,1,2)12 model is not adequate, and in the meanwhile prove the adequacy of the previous model.

| **Maximum Likelihood Estimation** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Approx Pr > |t|** | **Lag** |
| **MU** | 0.0027417 | 0.0015298 | 1.79 | 0.0731 | 0 |
| **MA1,1** | 0.37627 | 0.04795 | 7.85 | <.0001 | 1 |
| **MA2,1** | 0.86365 | 0.03642 | 23.71 | <.0001 | 12 |
| **MA3,1** | 0.05202 | 0.05983 | 0.87 | 0.3846 | 24 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Constant Estimate** | | | 0.002742 | |
| **Variance Estimate** | | | 0.08047 | |
| **Std Error Estimate** | | | 0.283673 | |
| **AIC** | | | 137.9753 | |
| **SBC** | | | 153.5858 | |
| **Number of Residuals** | | | 366 | |
| **Autocorrelation Check of Residuals** | | | | | | | | | | |
| **To Lag** | **Chi-Square** | **DF** | **Pr > ChiSq** | | **Autocorrelations** | | | | | |
| **6** | 6.61 | 3 | 0.0854 | | 0.022 | -0.012 | -0.089 | -0.048 | -0.052 | -0.065 |
| **12** | 13.95 | 9 | 0.1242 | | -0.046 | -0.051 | 0.118 | 0.003 | 0.026 | 0.006 |
| **18** | 15.65 | 15 | 0.4057 | | -0.028 | -0.018 | -0.037 | -0.014 | -0.031 | -0.029 |
| **24** | 19.88 | 21 | 0.5286 | | -0.055 | -0.069 | 0.046 | -0.006 | -0.031 | -0.002 |
| **30** | 24.32 | 27 | 0.6126 | | 0.041 | -0.008 | 0.030 | 0.009 | -0.038 | -0.083 |
| **36** | 31.26 | 33 | 0.5538 | | -0.058 | -0.048 | 0.032 | 0.092 | 0.042 | 0.013 |
| **42** | 43.32 | 39 | 0.2922 | | 0.071 | 0.027 | -0.106 | -0.040 | -0.056 | -0.086 |
| **48** | 46.54 | 45 | 0.4087 | | -0.029 | 0.041 | 0.016 | 0.052 | 0.046 | -0.009 |





* 1. **Model Selection**

All together there are two candidate models, namely ARIMA(0,1,1) X (0,1,1)12 and ARIMA(1,1,1) X (0,1,1)12. Based on the AIC criteria, the ARIMA(1,1,1) X (0,1,1)12 model is selected as the better fitted model.

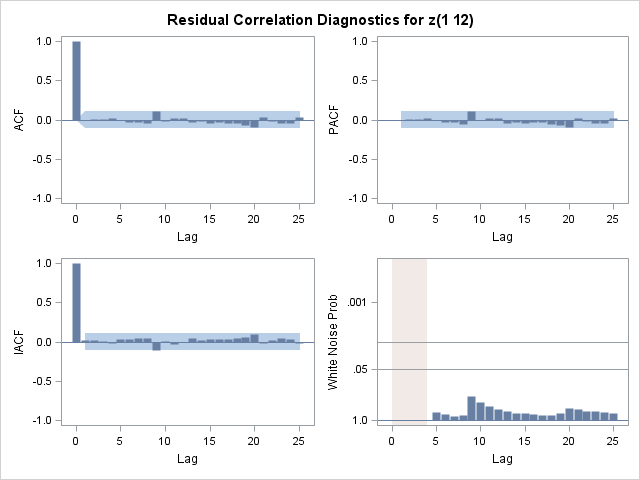
To further verify the adequacy of newly proposed model, further overfittings are conducted. Based on the results of previous overfittings, the newly added factors can be restricted to the non-seasonal component.

1. ARIMA(1,1,2) X (0,1,1)12

Firstly the order of MA factor (q) is increased to 2. Notice that the MA3,1 factor is significant at 0.05 level, but insignificant at 0.01 level, while other factors have slight changes (MA1,1 changes from 0.67 to 0.76; AR1,1 changes from 0.32 to 0.40). From the perspective of AIC, it decreases from 136.6889 to 133.7986, and the variance decreases from 0.079703 to 0.078943. This may due to the significant sample ACF at lag = 3 and 9. Also, the residuals have insignificant ACF and PACF over all periods and pass the normality check. Therefore the ARIMA(1,1,1) X (0,1,1)12 model still needs further verification.

| **Maximum Likelihood Estimation of ARIMA(1,1,2) X (0,1,1)12** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Approx Pr > |t|** | **Lag** |
| **MU** | 0.0028162 | 0.0009856 | 2.86 | 0.0043 | 0 |
| **MA1,1** | 0.76443 | 0.06834 | 11.19 | <.0001 | 1 |
| **MA2,1** | -0.11424 | 0.05730 | -1.99 | 0.0462 | 2 |
| **MA3,1** | 0.89281 | 0.03220 | 27.72 | <.0001 | 12 |
| **AR1,1** | 0.40531 | 0.09151 | 4.43 | <.0001 | 1 |

|  |  |
| --- | --- |
| **Constant Estimate** | 0.001675 |
| **Variance Estimate** | 0.078943 |
| **Std Error Estimate** | 0.280968 |
| **AIC** | 133.7986 |
| **SBC** | 153.3118 |
| **Number of Residuals** | 366 |



To check whether the ARIMA(1,1,1) X (0,1,1)12 model needs to be updated to a more aggressive model like ARIMA(1,1,2) X (0,1,1)12, two over-parameterized models are proposed by increasing the non-seasonal MA and AR factor by one step by step. As shown in the table, the newly proposed ARIMA (1,1,3) X (0,1,1)12 model has insignificant MA3,1 factor and larger AIC and variance. ARIMA (2,1,2) X (0,1,1)12 model has similar results. Therefore the newly proposed over-parameterized models are rejected, and ARIMA(1,1,2) X (0,1,1)12 model is kept as a plausible candidate.

| **Maximum Likelihood Estimation of ARIMA (1,1,3) X (0,1,1)12** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Approx Pr > |t|** | **Lag** |
| **MU** | 0.0028171 | 0.0009883 | 2.85 | 0.0044 | 0 |
| **MA1,1** | 0.76527 | 0.08140 | 9.40 | <.0001 | 1 |
| **MA2,1** | -0.11455 | 0.05968 | -1.92 | 0.0549 | 2 |
| **MA3,1** | -0.0012003 | 0.06213 | -0.02 | 0.9846 | 3 |
| **MA4,1** | 0.89285 | 0.03250 | 27.47 | <.0001 | 12 |
| **AR1,1** | 0.40603 | 0.09971 | 4.07 | <.0001 | 1 |

|  |  |
| --- | --- |
| **Constant Estimate** | 0.001673 |
| **Variance Estimate** | 0.079161 |
| **Std Error Estimate** | 0.281356 |
| **AIC** | 135.7983 |
| **SBC** | 159.2141 |
| **Number of Residuals** | 366 |

| **Maximum Likelihood Estimation of ARIMA (2,1,2) X (0,1,1)12** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Approx Pr > |t|** | **Lag** |
| **MU** | 0.0028326 | 0.0009548 | 2.97 | 0.0030 | 0 |
| **MA1,1** | 0.79286 | 0.07601 | 10.43 | <.0001 | 1 |
| **MA2,1** | 0.10742 | 0.46054 | 0.23 | 0.8156 | 2 |
| **MA3,1** | 0.89349 | 0.03235 | 27.62 | <.0001 | 12 |
| **AR1,1** | 0.43348 | 0.09644 | 4.49 | <.0001 | 1 |
| **AR2,1** | 0.23629 | 0.46596 | 0.51 | 0.6121 | 2 |

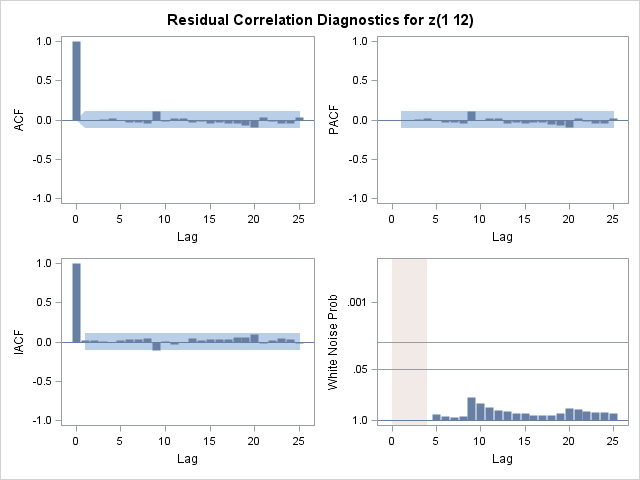
|  |  |
| --- | --- |
| **Constant Estimate** | 0.001226 |
| **Variance Estimate** | 0.079068 |
| **Std Error Estimate** | 0.28119 |
| **AIC** | 135.4832 |
| **SBC** | 158.899 |
| **Number of Residuals** | 366 |

ii. ARIMA(2,1,1) X (0,1,1)12

Secondly the order of AR factor (q) is increased to 2. Notice that the MA3,1 factor is significant at 0.05 level, but insignificant at 0.01 level, while other factors have slight changes (MA1,1 changes from 0.67 to 0.78; AR1,1 changes from 0.32 to 0.42). From the perspective of AIC, it decreases from 136.6889 to 133.5628, and the variance decreases from 0.079703 to 0.078871. Also, the residuals have insignificant ACF and PACF over all periods and pass the normality check. Therefore the ARIMA(2,1,1) X (0,1,1)12 model is proposed as an alternative model.

| **Maximum Likelihood Estimation of ARIMA(2,1,1) X (0,1,1)12** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Approx Pr > |t|** | **Lag** |
| **MU** | 0.0028259 | 0.0009647 | 2.93 | 0.0034 | 0 |
| **MA1,1** | 0.78038 | 0.06693 | 11.66 | <.0001 | 1 |
| **MA2,1** | 0.89348 | 0.03222 | 27.73 | <.0001 | 12 |
| **AR1,1** | 0.42020 | 0.08880 | 4.73 | <.0001 | 1 |
| **AR2,1** | 0.12565 | 0.05994 | 2.10 | 0.0360 | 2 |

|  |  |
| --- | --- |
| **Constant Estimate** | 0.001433 |
| **Variance Estimate** | 0.078871 |
| **Std Error Estimate** | 0.28084 |
| **AIC** | 133.5628 |
| **SBC** | 153.076 |
| **Number of Residuals** | 366 |



Similarly, two over-parameterized models are proposed to check whether the ARIMA(1,1,1) X (0,1,1)12 model needs to be updated to a more aggressive model like ARIMA(2,1,1) X (0,1,1)12. As shown in the table, the newly proposed ARIMA (2,1,2) X (0,1,1)12 model has insignificant MA2,1 factor and larger AIC and variance. ARIMA (3,1,1) X (0,1,1)12 model has similar results. Therefore the newly proposed over-parameterized models are rejected, and ARIMA(2,1,1) X (0,1,1)12 model is kept as a plausible candidate.

| **Maximum Likelihood Estimation of ARIMA(2,1,2) X (0,1,1)12** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Approx Pr > |t|** | **Lag** |
| **MU** | 0.0028326 | 0.0009548 | 2.97 | 0.0030 | 0 |
| **MA1,1** | 0.79286 | 0.07601 | 10.43 | <.0001 | 1 |
| **MA2,1** | 0.10742 | 0.46054 | 0.23 | 0.8156 | 2 |
| **MA3,1** | 0.89349 | 0.03235 | 27.62 | <.0001 | 12 |
| **AR1,1** | 0.43348 | 0.09644 | 4.49 | <.0001 | 1 |
| **AR2,1** | 0.23629 | 0.46596 | 0.51 | 0.6121 | 2 |

|  |  |
| --- | --- |
| **Constant Estimate** | 0.001226 |
| **Variance Estimate** | 0.079068 |
| **Std Error Estimate** | 0.28119 |
| **AIC** | 135.4832 |
| **SBC** | 158.899 |
| **Number of Residuals** | 366 |

| **Maximum Likelihood Estimation of ARIMA (3,1,1) X (0,1,1)12** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Parameter** | **Estimate** | **Standard Error** | **t Value** | **Approx Pr > |t|** | **Lag** |
| **MU** | 0.0028371 | 0.0009476 | 2.99 | 0.0028 | 0 |
| **MA1,1** | 0.79233 | 0.07527 | 10.53 | <.0001 | 1 |
| **MA2,1** | 0.89432 | 0.03248 | 27.54 | <.0001 | 12 |
| **AR1,1** | 0.43090 | 0.09239 | 4.66 | <.0001 | 1 |
| **AR2,1** | 0.13105 | 0.06278 | 2.09 | 0.0368 | 2 |
| **AR3,1** | 0.01310 | 0.06190 | 0.21 | 0.8324 | 3 |

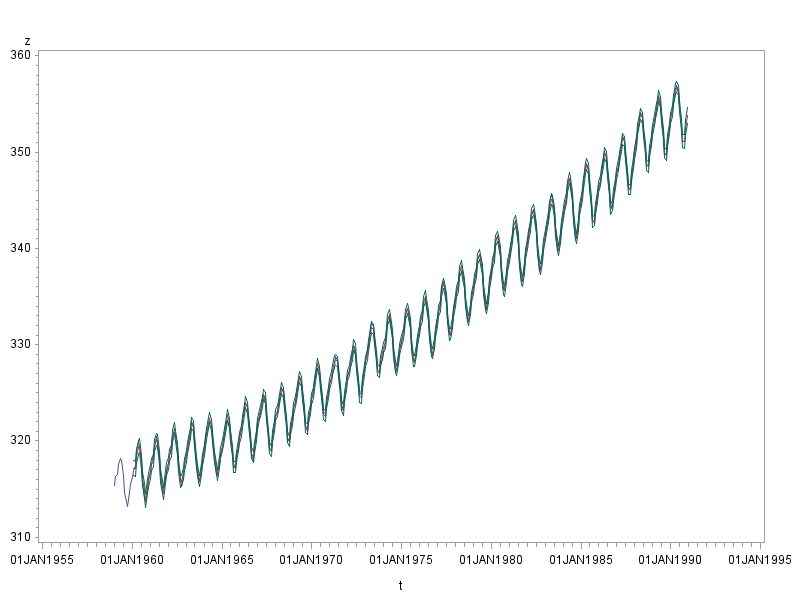
|  |  |
| --- | --- |
| **Constant Estimate** | 0.001385 |
| **Variance Estimate** | 0.079058 |
| **Std Error Estimate** | 0.281173 |
| **AIC** | 135.5265 |
| **SBC** | 158.9423 |
| **Number of Residuals** | 366 |

In Sum, based on the AIC criteria, we can conclude that ARIMA (2,1,1) X (0,1,1)12 is the most adequate model based on the data.

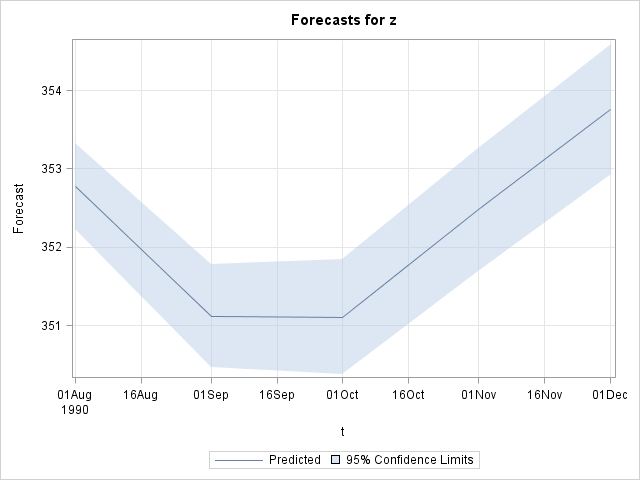
|  |  |
| --- | --- |
| Model | AIC |
| ARIMA(1,1,1) X (0,1,1)12 | 135.398 |
| ARIMA(1,1,2) X (0,1,1)12 | 133.7986 |
| ARIMA(2,1,1) X (0,1,1)12 | 133.5628 |

1. **Forecasting**

The ARIMA(2,1,1) X (0,1,1)12 model is adopted to forecast the next 5 months’ carbon dioxide concentration. From the table it’s clear that all true values fall within the 95% confidence interval, which demonstrates the good performance of the selected model in terms of forecasting.



|  |  |  |  |
| --- | --- | --- | --- |
| True Value | Forecast Value | 95% Lower Bound | 95% Upper Bound |
| 352.68 | 352.7776 | 352.2272 | 353.3281 |
| 350.72 | 351.1238 | 350.4704 | 351.7773 |
| 350.92 | 351.1114 | 350.3757 | 351.8471 |
| 352.55 | 352.4793 | 351.6928 | 353.2659 |
| 353.91 | 353.759 | 352.9301 | 354.588 |



1. **Further Discussion**

The seasonal ARIMA model affirms people’s conjecture that the monthly carbon dioxide concentration is increasing steadily with a relative constant variation, which indicates that the global warming problem is still unsolved and developing with growing severity. Furthermore, this increasing trend is primarily contributed by 3 factors. The first one comes from yearly-based variation (seasonal MA factor), including the changes in population, boom of private transportation and so on. The second factor comes from monthly dependency (non-seasonal AR factor), which can be interpreted as the prolonged effect of the last month carbon dioxide concentration. The third factor comes from the seasonal variation within a year (non-seasonal MA factor), such as the seasonal changes in the vegetation coverage rate and meteorological features. By analyzing the individual and joint effects of these factors, scientists may come up with new solutions to the global warming issue.

1. **Code & Data Source**

/\* Draw the time plot \*/

symbol i=join v=none;

**proc** **gplot** data=tsnew;

plot z\*t;

**run**;

**quit**;

/\* Identify arima models on raw data \*/

**proc** **arima** data=tsnew;

identify alpha=**0.05** var=z nlag=**25**;

**run**;

/\*check log transformation\*/

**data** tslog;

set tsnew;

lz=log(z);

**run**;

symbol i=join v=none;

**proc** **gplot** data=tslog;

plot lz\*t;

**run**;

**quit**;

**proc** **arima** data=tslog;

identify alpha=**0.05** var=lz nlag=**25**;

**run**;

/\* Identify arima models on d=1 data \*/

**proc** **arima** data=tsnew;

identify alpha=**0.05** var=z(**1**) nlag=**25**;

**run**;

/\* Take seasonal differencing since the sample ACF decays slowly especially after periods \*/

identify alpha=**0.05** var=z(**1**,**12**) stationarity=(dickey=**4**);

**run**;

/\* Estimate the ARIMA(0,1,1)X(0,1,1)12 model to the data \*/

estimate method=ml q=(**1**)(**12**) plot;

**run**;

forecast out=fore0 lead=**0** id=t;

**run**;

/\* Draw the time plot of residual\*/

symbol i=join v=none;

**proc** **gplot** data=fore0;

plot residual\*t;

**run**;

**quit**;

/\* Perform the normality test on residuals\*/

**proc** **univariate** data=fore0 normal plot;

var residual;

**run**;

/\*overfitting\*/

**proc** **arima** data=tsnew;

identify var=z(**1**,**12**) nlag=**25**;

**run**;

/\*ARIMA(1,1,1)X(0,1,1)12 model \*/

estimate method=ml p=(**1**) q=(**1**)(**12**) plot;

**run**;

/\*ARIMA(0,1,2)X(0,1,1)12 model \*/

estimate method=ml q=(**1**)(**2**)(**12**) plot;

**run**;

/\*ARIMA(0,1,1)X(1,1,1)12 model \*/

estimate method=ml p=(**12**) q=(**1**)(**12**) plot;

**run**;

/\*ARIMA(0,1,1)X(0,1,2)12 model \*/

estimate method=ml q=(**1**)(**12**)(**24**) plot;

**run**;

/\*overfitting to new model: ARIMA(1,1,1)X(0,1,1)12 model \*/

**proc** **arima** data=tsnew;

identify var=z(**1**,**12**) nlag=**25**;

**run**;

/\*ARIMA(1,1,2)X(0,1,1)12 model \*/

estimate method=ml p=(**1**) q=(**1**)(**2**)(**12**) plot;

**run**;

estimate method=ml p=(**1**) q=(**1**)(**2**)(**3**)(**12**) plot;

**run**;

estimate method=ml p=(**1**)(**2**) q=(**1**)(**2**)(**12**) plot;

**run**;

/\*ARIMA(2,1,1)X(0,1,1)12 model \*/

estimate method=ml p=(**1**)(**2**) q=(**1**)(**12**) plot;

**run**;

estimate method=ml p=(**1**)(**2**) q=(**1**)(**2**)(**12**) plot;

**run**;

estimate method=ml p=(**1**)(**2**)(**3**) q=(**1**)(**12**) plot;

**run**;

/\* Do forecasting by using the fitted ARIMA(2,1,1)x (0,1,1)12 model \*/

**proc** **arima** data=tsnew;

identify var=z(**1**,**12**) nlag=**25**;

**run**;

estimate method=ml p=(**1**)(**2**) q=(**1**)(**12**) plot;

**run**;

forecast out=fore5 lead=**5** id=t interval=month;

**run**;

**quit**;

/\* Draw the forecasting time plot \*/

symbol i=join v=none;

**proc** **gplot** data=fore5;

plot z\*t=**1** forecast\*t=**2** l95\*t=**3** u95\*t=**3**/overlay;

**run**;

**quit**;

**Data Source**

https://datamarket.com/data/set/22qs/monthly-measurements-of-carbon-dioxide-above-mauna-loa-hawaii-jan-1959-dec-1990-units-parts-per-million-ppm-missing-values-have-been-filled-in-by-linear-interpolation-the-data-were-collected-by-scripps-institute-of-oceanography-la-jolla-california#!ds=22qs&display=line