**DNSC 6219 Time Series Final Report**

**Time Series Analysis of Delhi Climate**

Team Member: Yihang Zhao, Qian Xie, Chengkai Ji, Leqi Yin

1. [Introduction](https://docs.google.com/document/d/1KprRLuqKwjrAsc3fz8i5BJ-ZdrNLpAJ_qM8mvtK5FLA/edit#heading=h.jdgvft67a4ur) and Overview [2](https://docs.google.com/document/d/1KprRLuqKwjrAsc3fz8i5BJ-ZdrNLpAJ_qM8mvtK5FLA/edit#heading=h.jdgvft67a4ur)
2. Univariate Time-series models 3-16

2.1 Deterministic Time Series Models and Error model

2.2 ARIMA models

2.3 Comparison of models

1. Multivariate Time Series Models 16-23

3.1 Regression model and analysis of regression residuals

3.2 Error model using regression residuals

3.3 Cross correlation analysis

1. Conclusion 23-24

# 

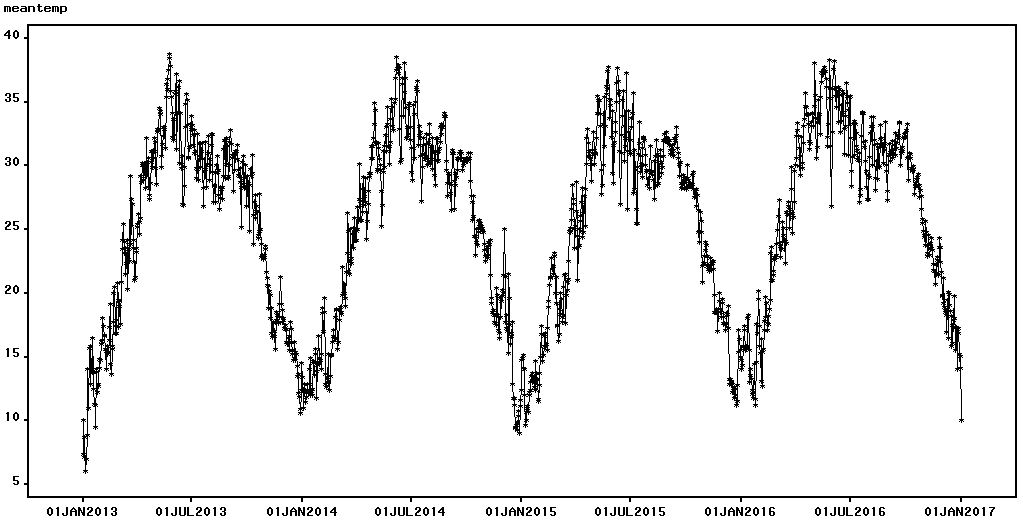
# Introduction and Overview

Weather is important but hard to predict. The complexity of the weather gives us more predictability to this intriguing system. Many variables within the Earth’s atmosphere, for example temperature, pressure, wind velocity, humidity, and precipitation, are not linear but interacting which aroused our interest in in-depth research and prediction.

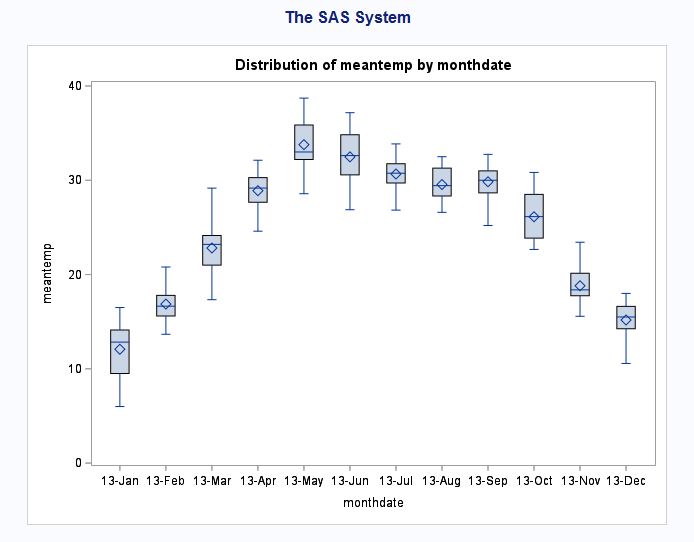
We find the daily climate time series dataset on Kaggle and decide to use it as the main data for the project. It contains the general climate information from 1st January 2013 to 24th April 2017 for the city Delhi, India provided by the Weather Underground API. We will use this data to forecast the weather and gain further understanding on the time series models.

The data contains 1462 observations, including 4 variables that are the basic attributes of the climate, which are mean temperature, humidity, wind speed, and mean pressure. Our goal is to test the dataset through different models we learned in class and compare the models to find the best one in predicting the mean temperature for the city Delhi.

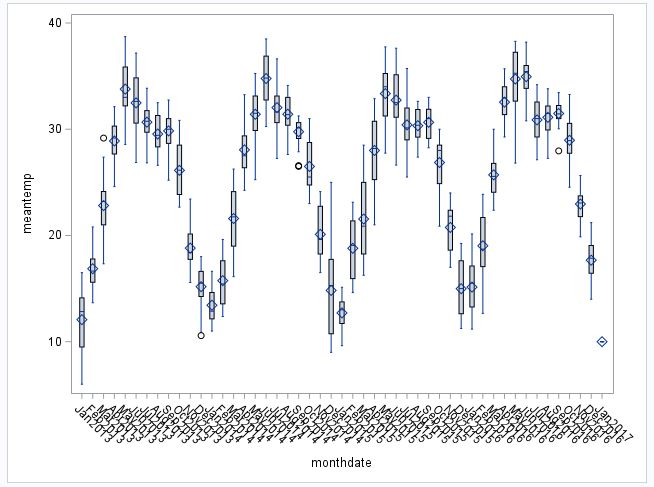
*Figure1.1 Time Series Plot view*



*Figure 1.2.1 Box Plot View of the Mean Temperature in Year 2013*



*Figure 1.2.2 Box Plot View of the Mean Temperature in Year 2013-2017*



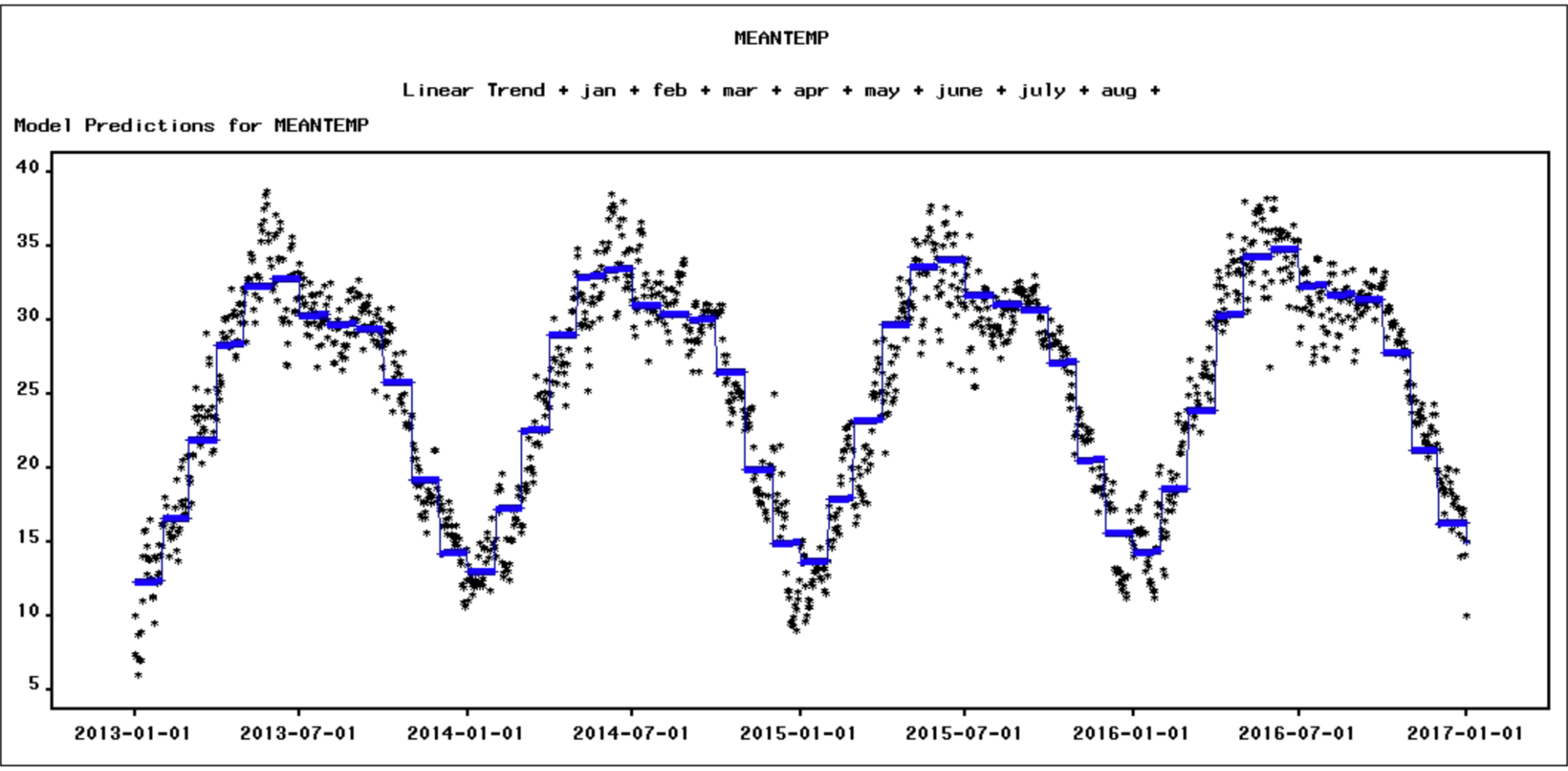
Based on the Time Series Plot and the box plot above, there is a strong seasonality and cyclical behavior of the climate in Delhi. The seasonality for this dataset is that the mean temperature trend moves upward from January to July every year and goes downward from July to December. Also, The trend of the mean temperature repeats in regular intervals every year. Therefore, we can conclude cyclical behavior is an annual circle. Furthermore, we are using 200 hold-out samples in our model fitting procedures.

2.Univariate Time-series models

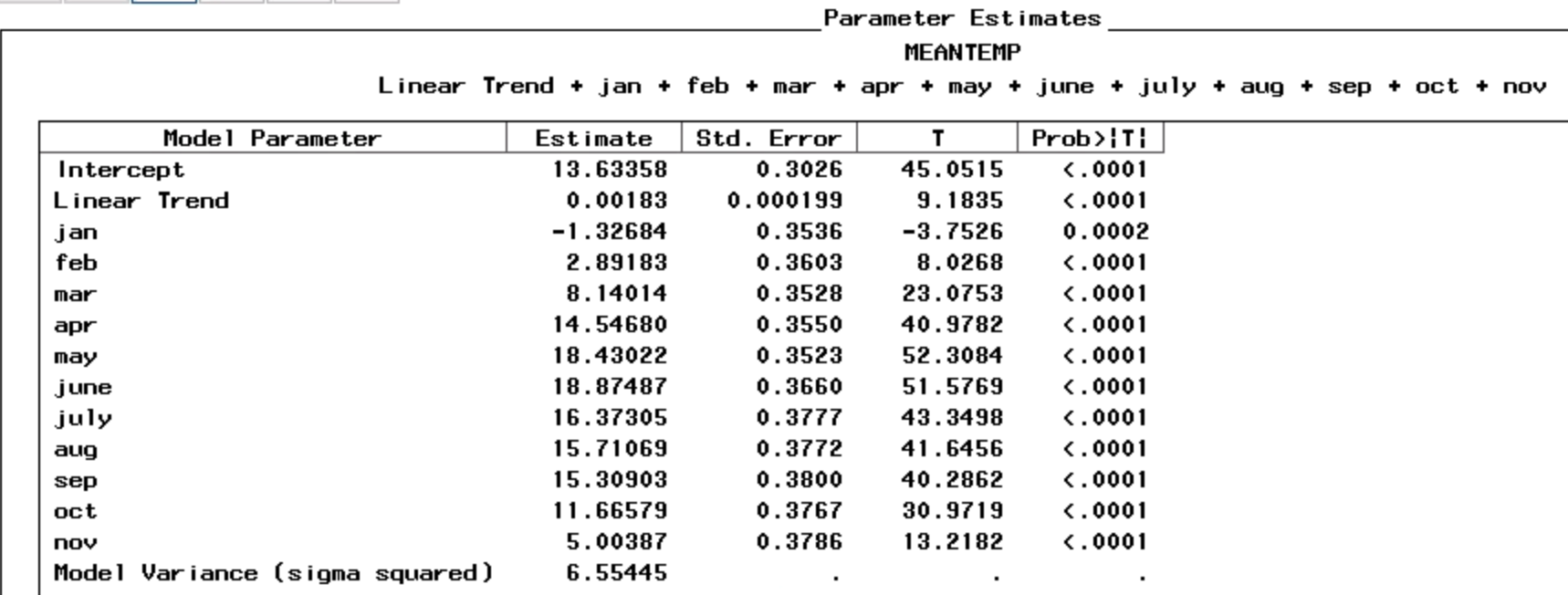
2.1 Deterministic Time Series Models and Error model

* Seasonal Dummies and Trend Model:

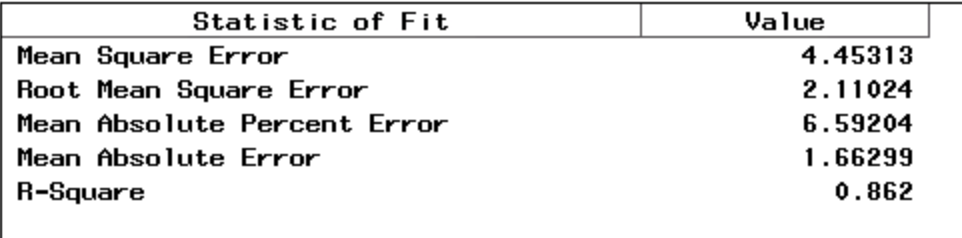
*Figure 2.1.1 Fit of model*



*Figure 2.1.2 Parameters of seasonal dummies and trend model*



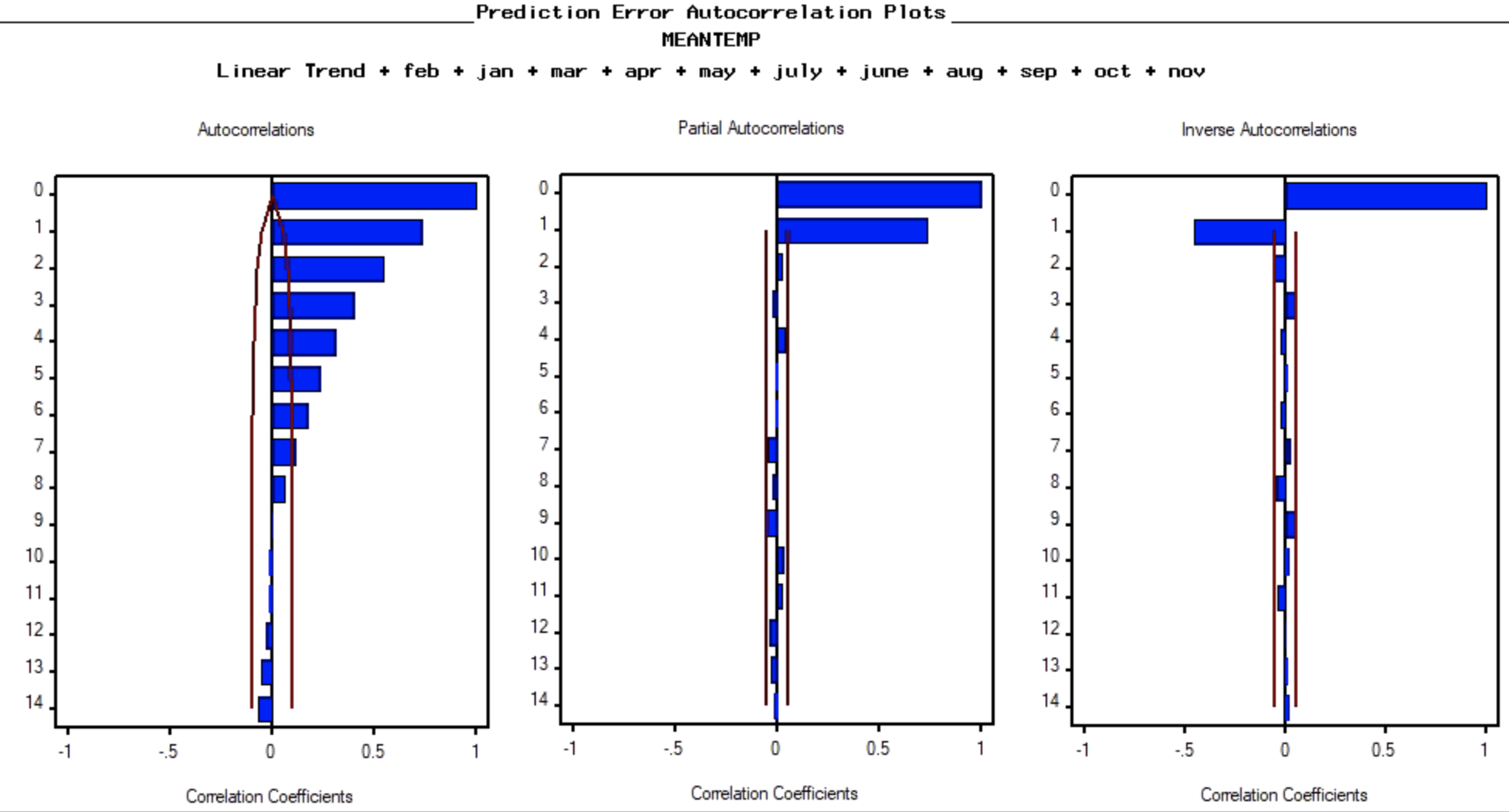
*Figure 2.1.3 Statistic of fit*



From the fit of the model(Figure 2.1.1), we can see that the model follows the pattern of the data, and the MAE and MAPE(Figure 2.1.3) of the model are 1.66 and 6.59, which is not too bad. The t-statistic of parameters of the model(Figure 2.1.2) are all less than 0.05, which indicates that all seasonal dummies and trend parameters are significant.

* Error Model of Seasonal Dummies and Trend Model

*Figure 2.1.4 ACF of residual*



According to figure 2.1.4, the ACF of residual decays exponentially and PACF drops to 0 after lag 1, so we believe AR(1) is a good model to model residuals.

*Figure 2.1.5 ACF of residual of error model AR(1)*

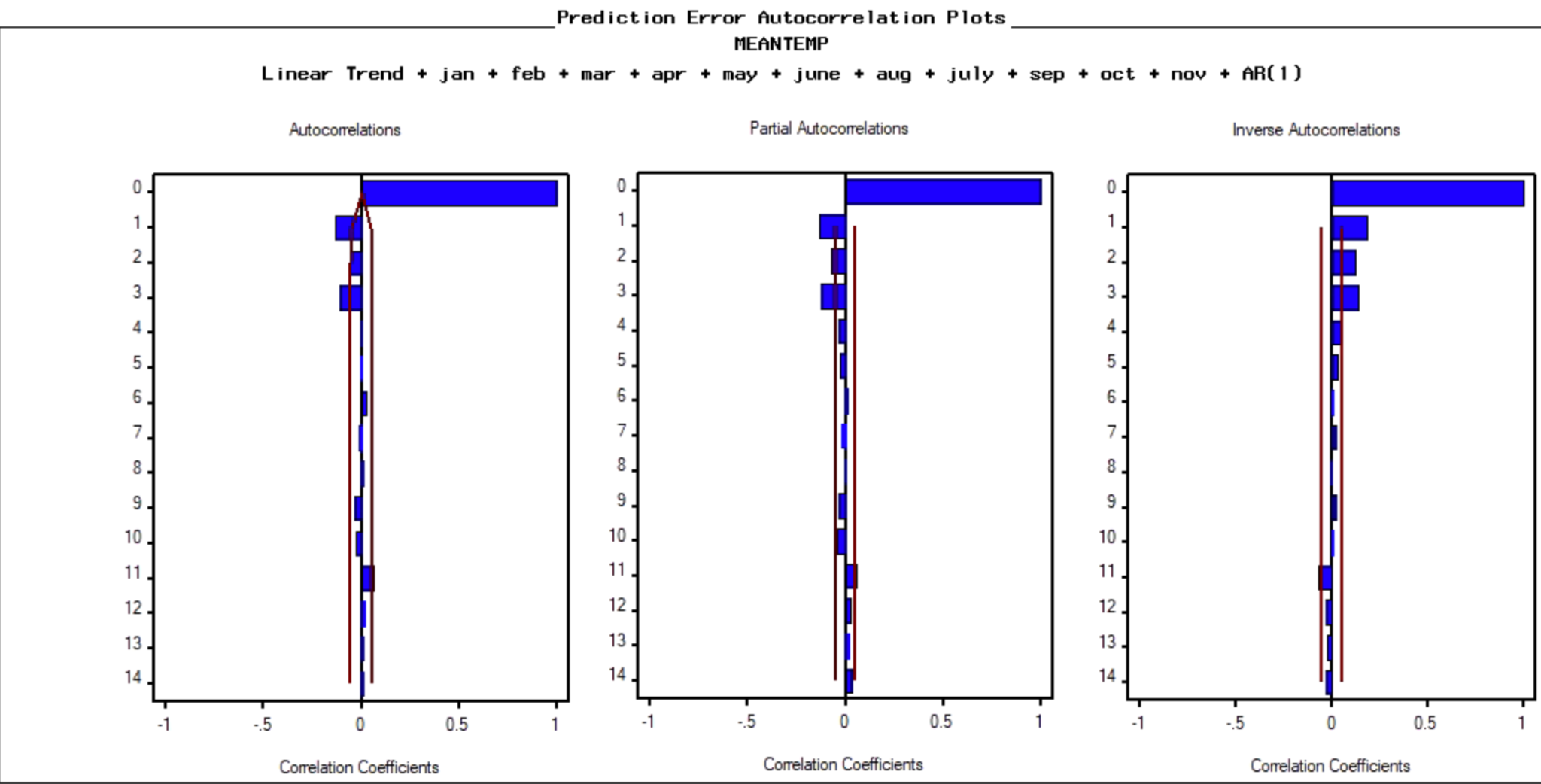


Figure 2.1.5 shows that the ACF drops to 0 after lag 3 and IACF decays, so we can fit an MA(3) to model the residuals of the error model.

*Figure 2.1.6 ACF and white noise test of residual of error model ARMA(1,3)*

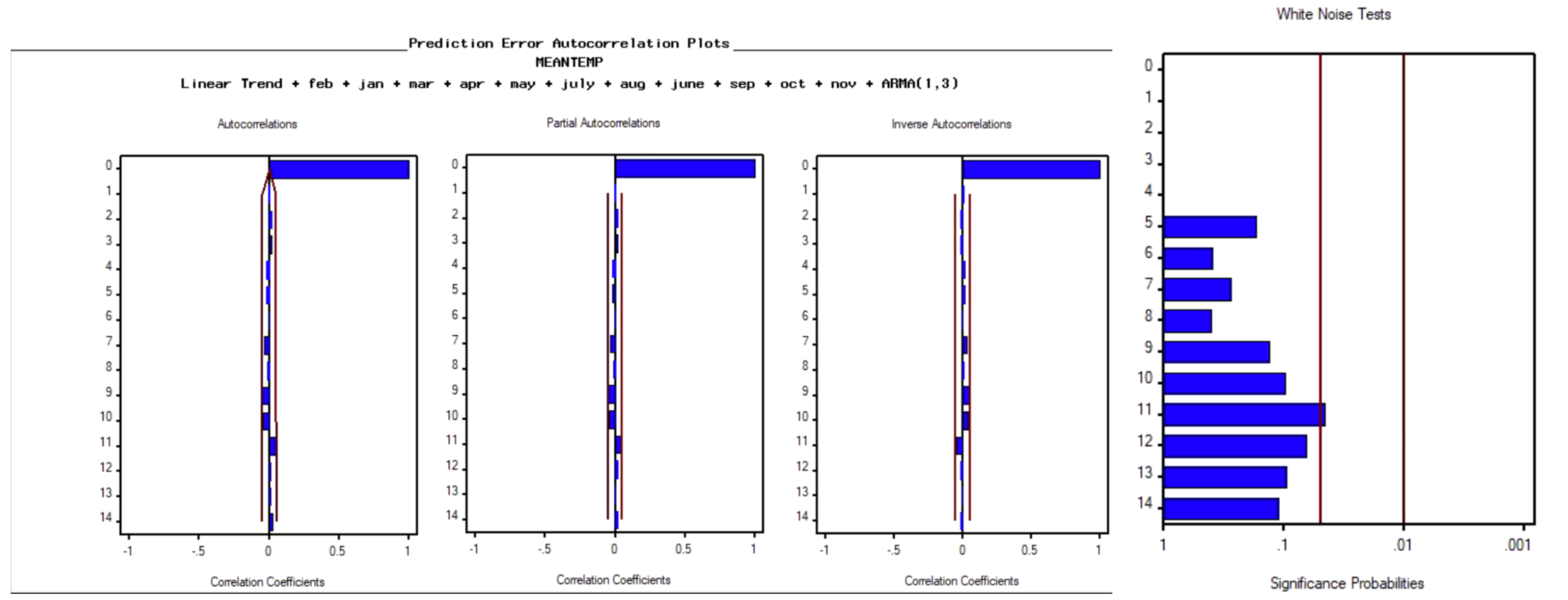
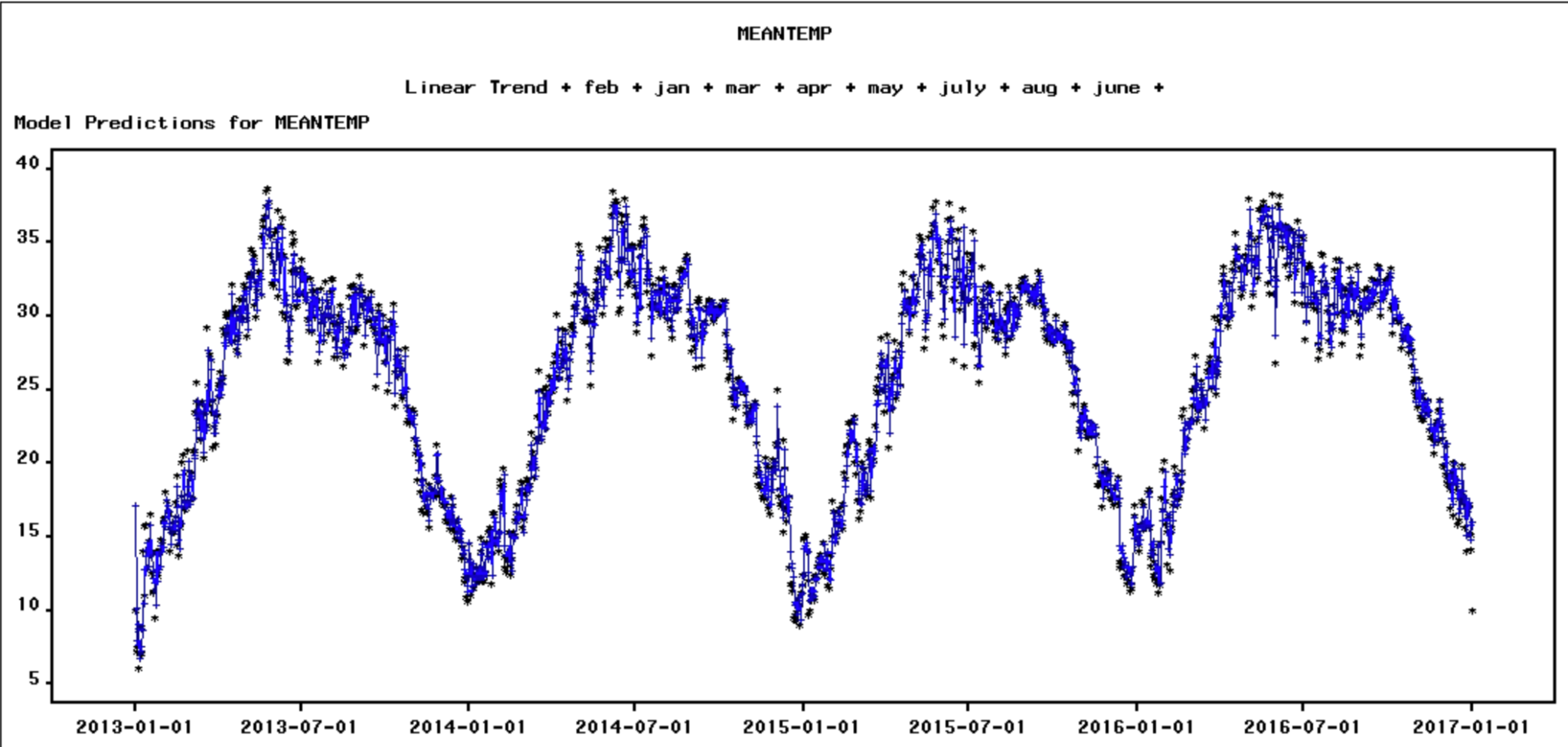
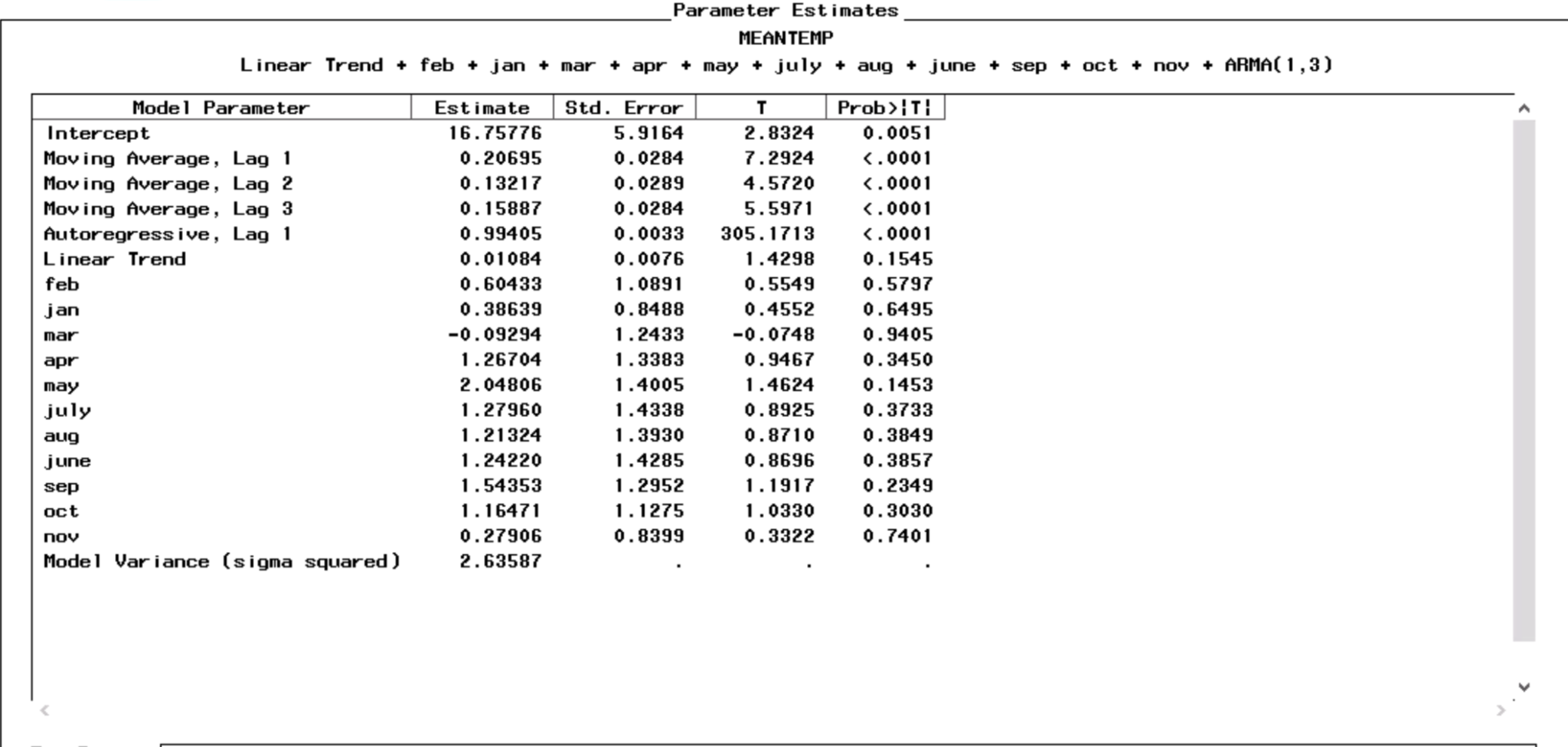
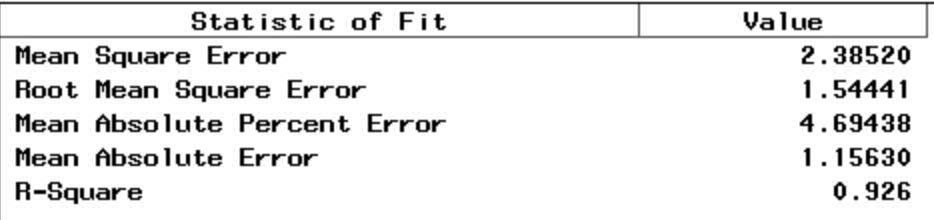


Figure 2.1.6 shows ACF, PACF and IACF drops to 0 after lag 0 and the white noise test indicates that the residuals are white noise, therefore we can stop fitting error at this point.

*Figure 2.1.7 Fit of error model ARMA(1,3)*



*Figure 2.1.8 Parameters of ARMA(1,3)*  
*Figure 2.1.9 Statistic of fit of ARMA(1,3)*

The MAE and MAPE of error model(Figure 2.1.9) are 1.15 and 4.69, which is lower than simple seasonal dummies and trend model at 1.66 and 6.59. On the other hand, the t-statistics of parameters of error model(Figure 2.1.8) show that only the parameters of ARMA are significant. Seasonal dummies and trend are no longer showing significance. Therefore, the performance may be better to use a simple ARMA or ARIMA model.

* Cyclical Model

Hold-out sample 200

*Figure 2.1.10 Periodogram*

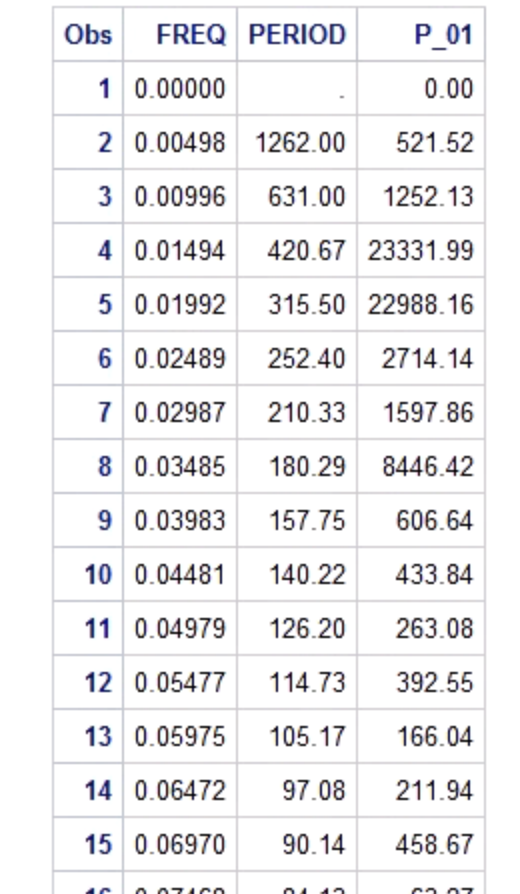
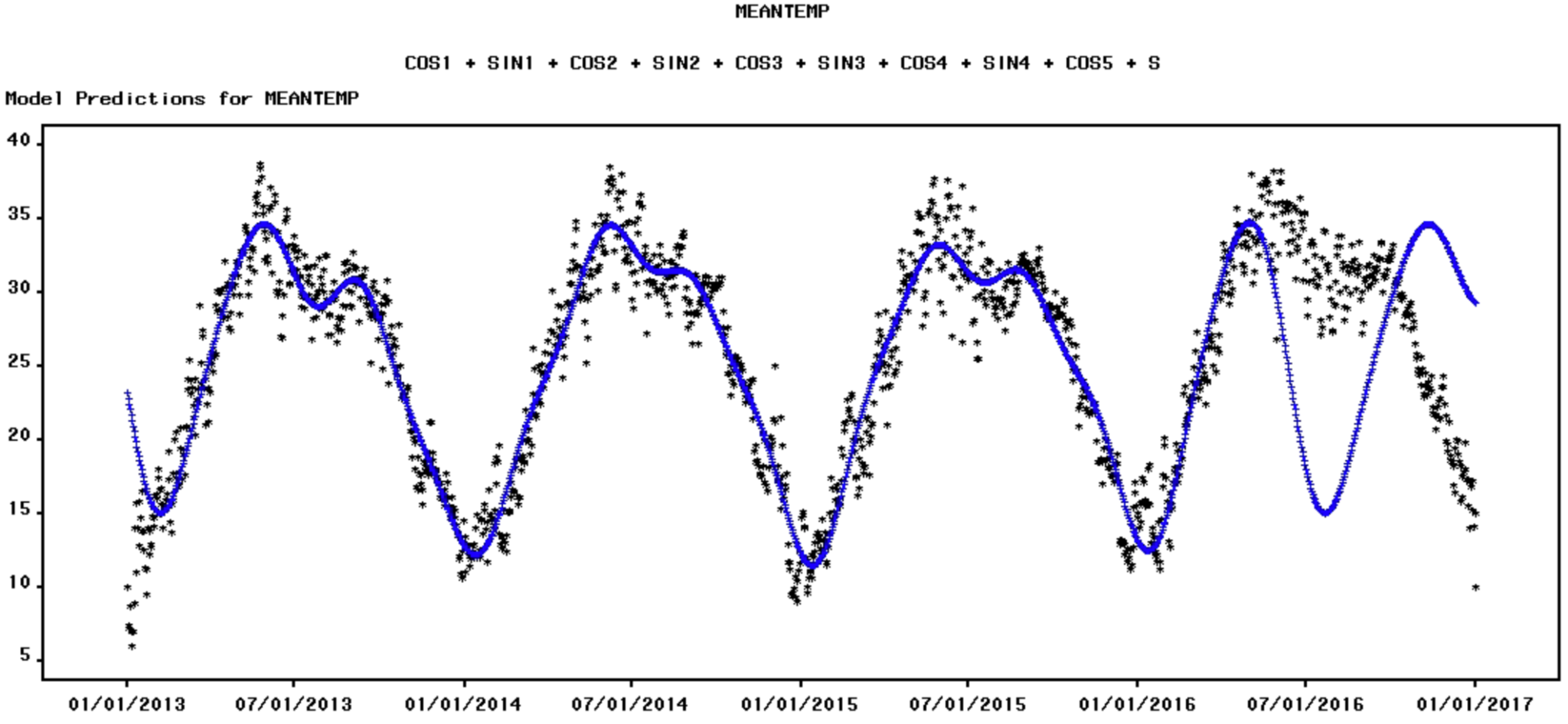
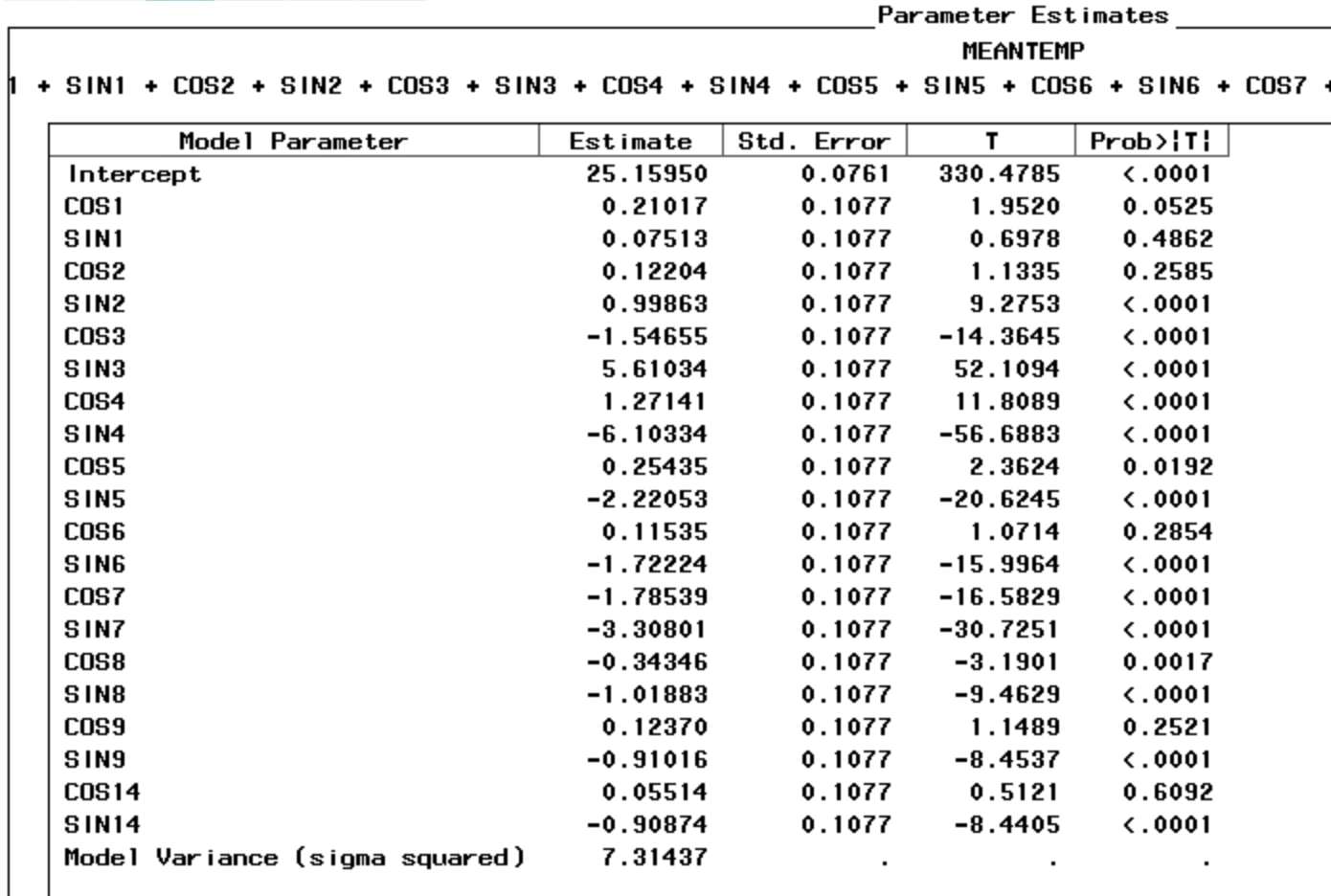


Figure 2.1.10 shows that the 10 harmonics with highest amplitude are: 1,2,3,4,5,6,7,8,9,14, so we created 20 sin and cos terms.

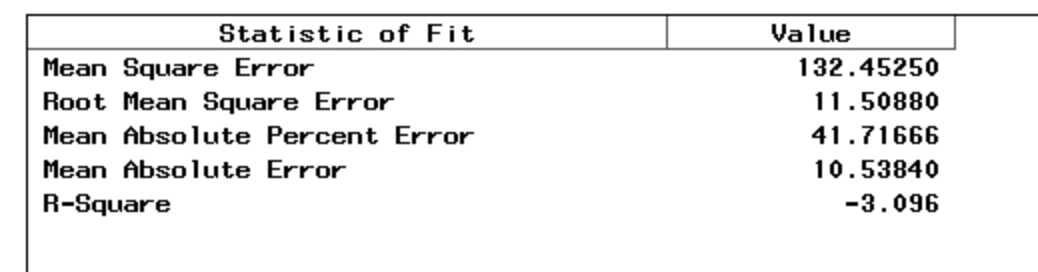
*Figure 2.1.11 Fit of cyclical model*



*Figure 2.1.12 Parameters of cyclical model*



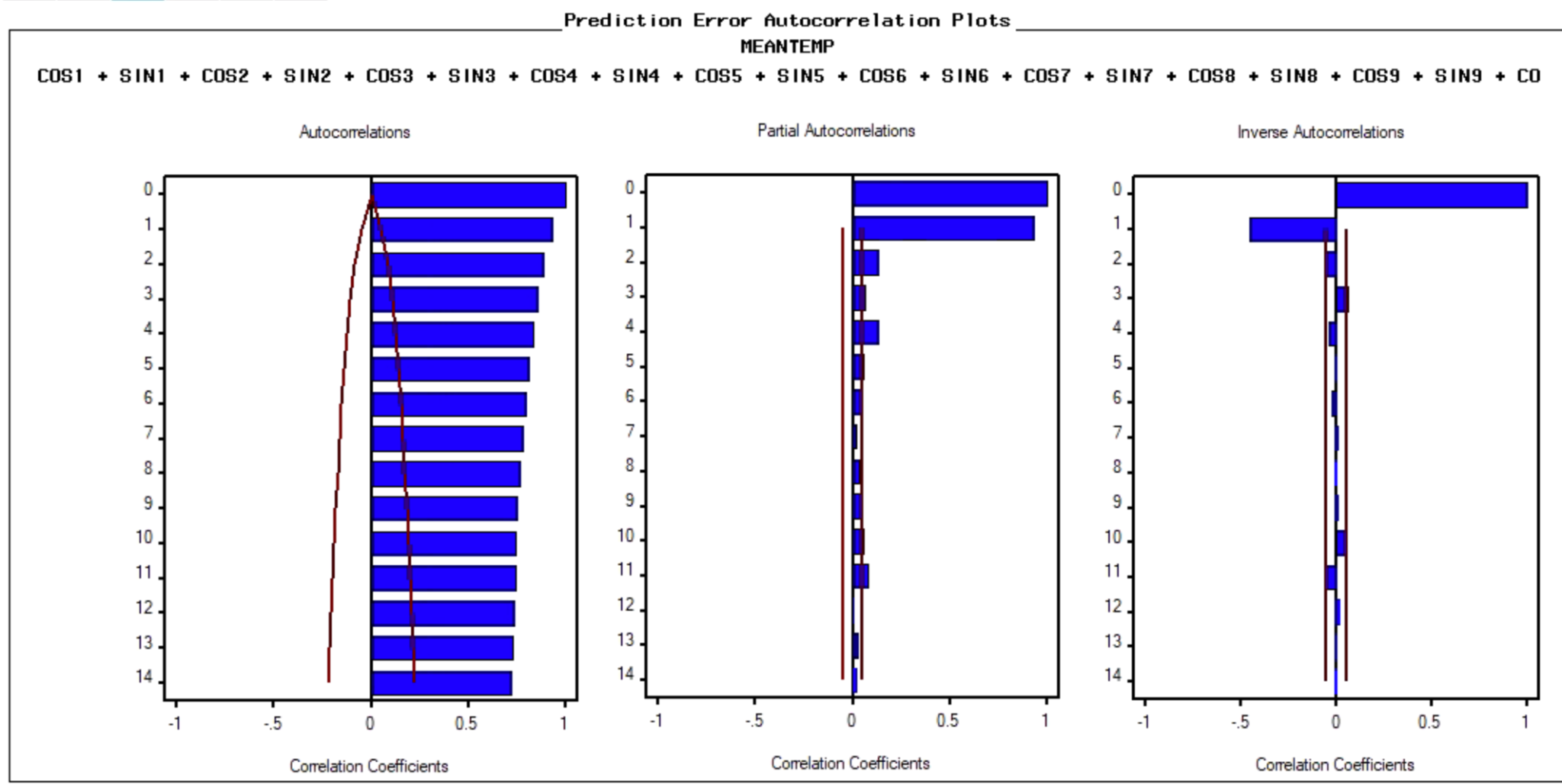
*Figure 2.1.13 Statistic of cyclical model*



From the fit of the cyclical model(Figure 2.1.11) and the statistic of fit(Figure 2.1.13), we can see that the cyclical model fits well on existing data but doesn't perform well on prediction. Most of the parameters of the model(Figure 2.1.12) are significant except for cos1, sin1, cos2, cos6 and cos14.

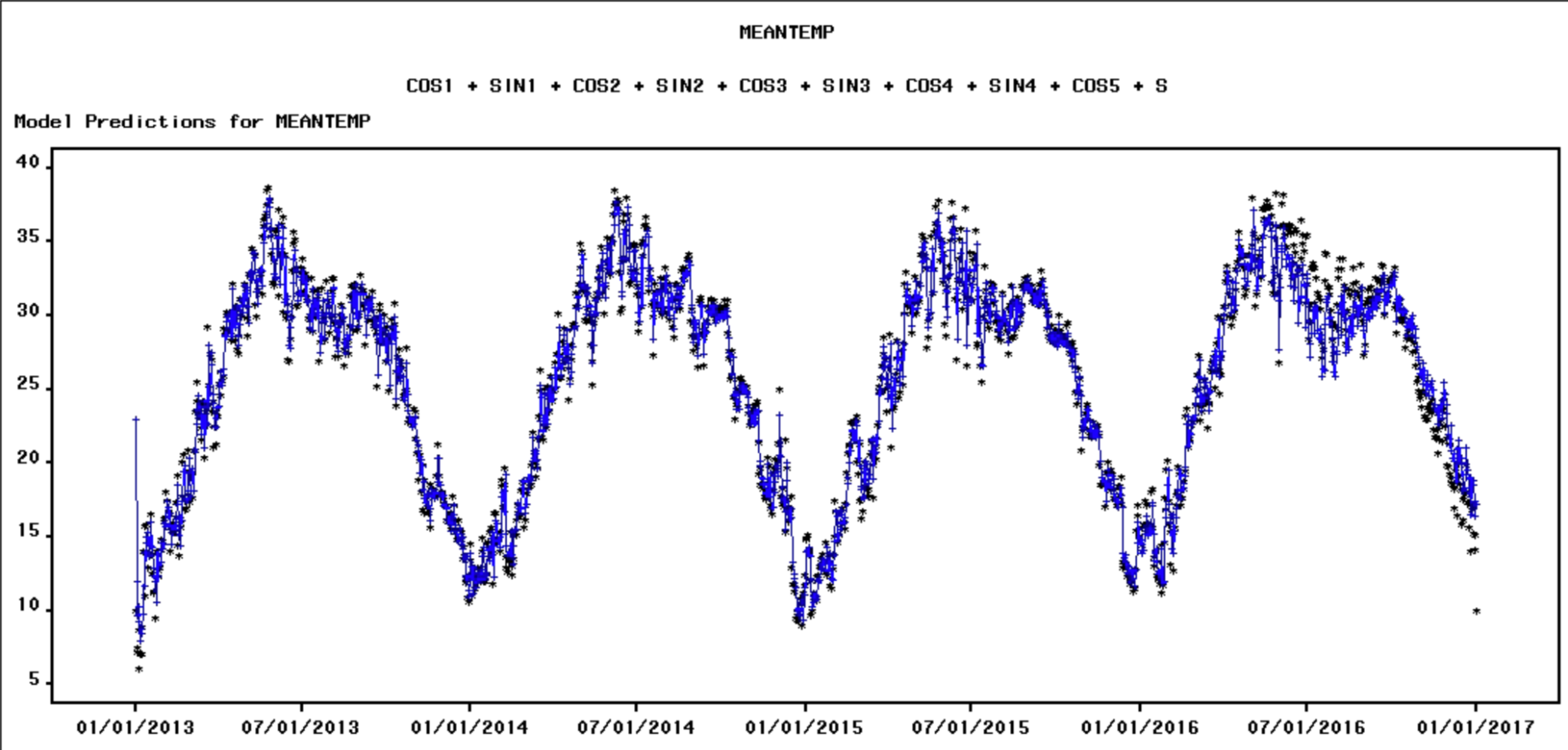
* Error Model of Cyclical Model

*Figure 2.1.14 ACF of residual of cyclical model*

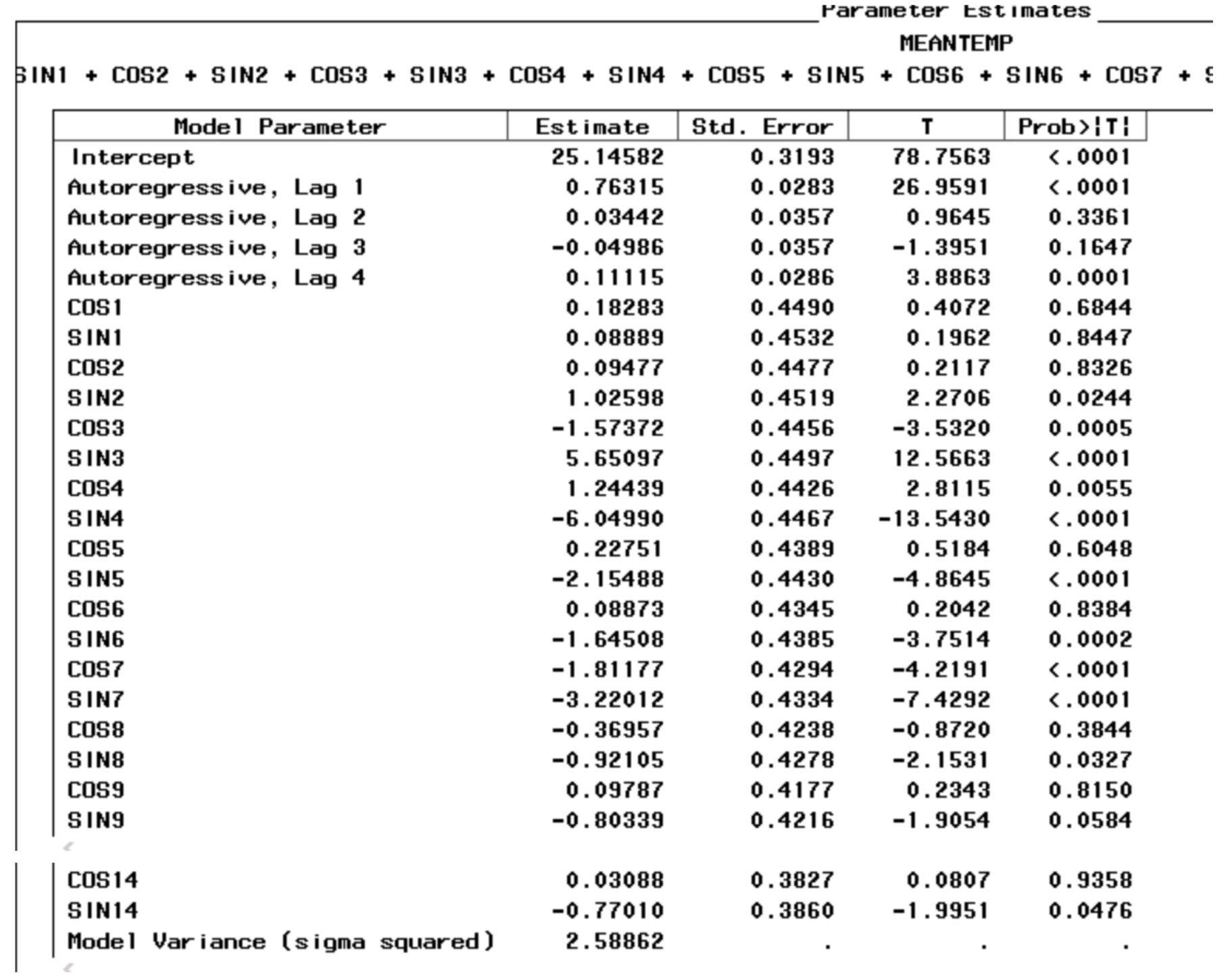


From figure 2.1.14, we can see that the ACF decays slowly, PACF drops to 0 after lag 4 and IACF drops quickly. So we decided to fit an AR(4) error model.

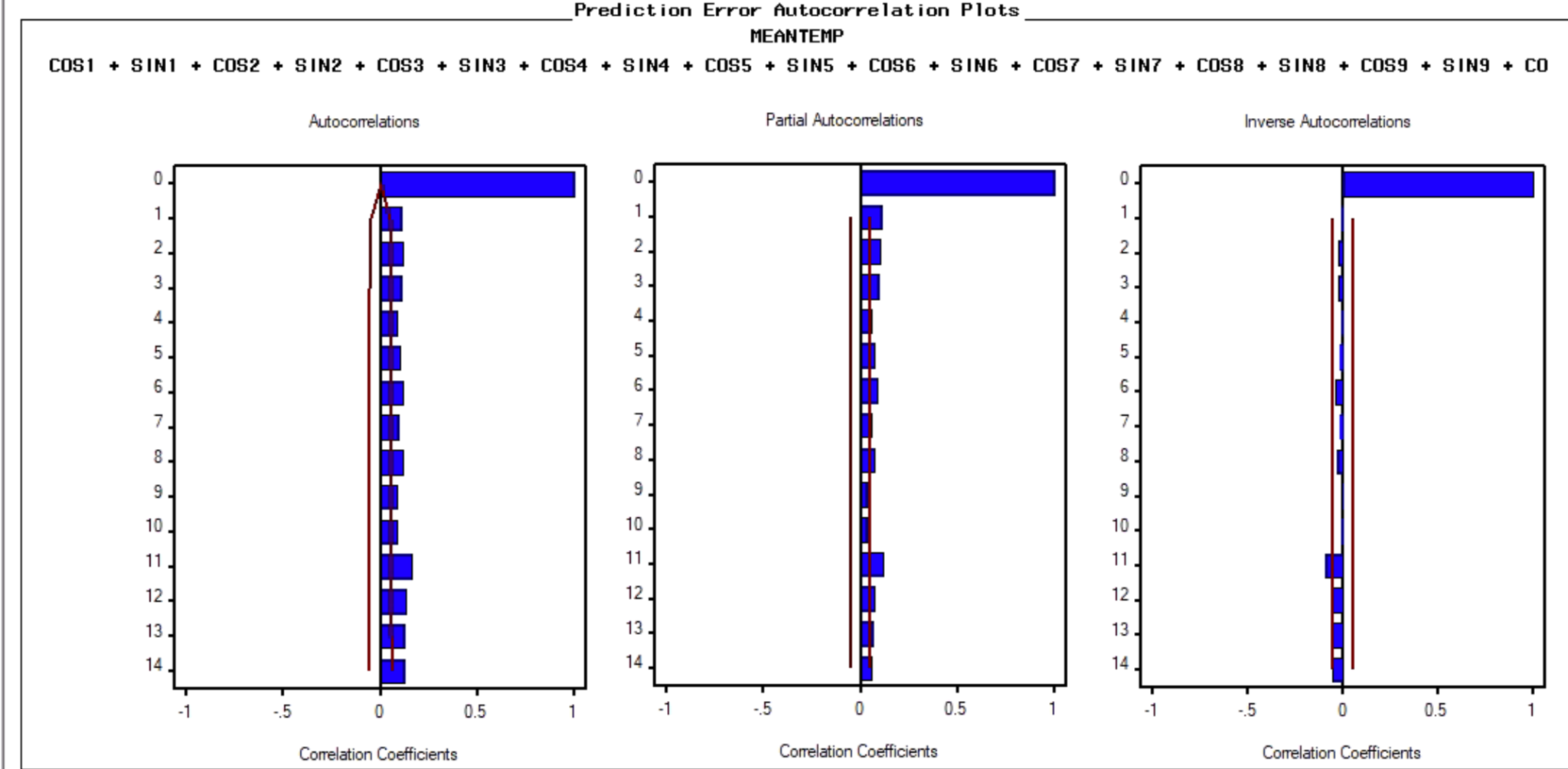
*Figure 2.1.15 Fit of AR(4) error model*



*Figure 2.1.16 Parameters of error model*



*Figure 2.1.17 ACF of residual of error model*



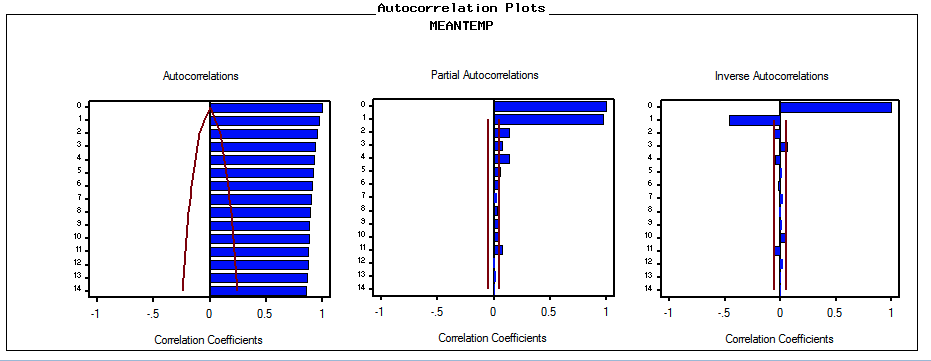
*Figure 2.1.18 Statistic of fit of error model*



From the fit of the model(Figure 2.1.15) and statistics of fit(Figure 2.1.18), we can see that the error model has a much better performance than the original model. Only a few parameters remain significant in the model(Figure 2.1.16). The ACF, PACF and IACF of residuals of the error model becomes very small after lag0, so we assume they’re white noise.

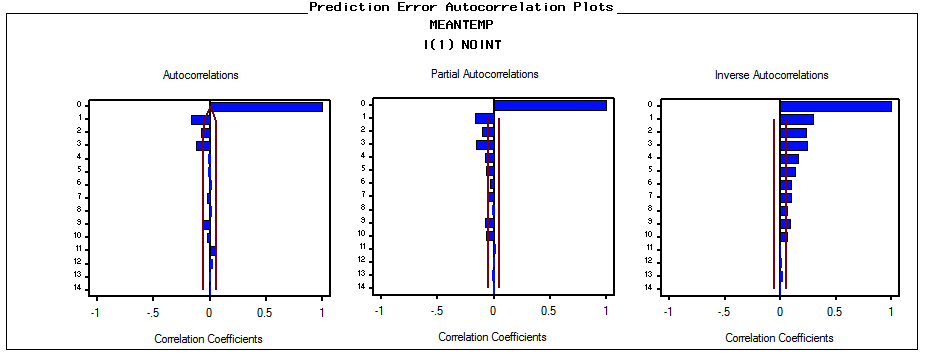
2.2 ARIMA models

*Figure 2.2.1 (ACF, PACF and Inverse Autocorrelation)*



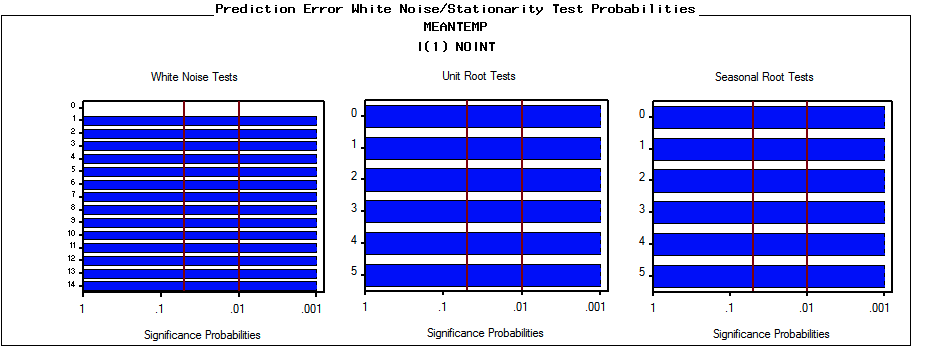
After graphing the series, we can see the obvious pattern of seasonality within the series. Based on autocorrelation plots, we found that ACF decays slowly which indicates that the series is nonstationary. Therefore, we fit the differencing model first.

*Figure 2.2.2 (ACF, PACF and Inverse Autocorrelation using non-seasonal differencing)*



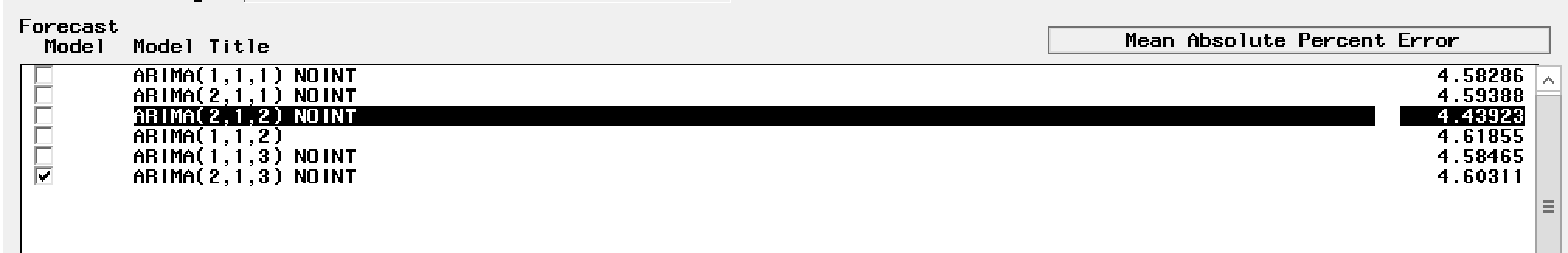
From figure (2.2.2), both ACF and PACF decays quickly, the first difference makes the series stationary. There is no seasonal pattern after the first difference, so we don’t need to include seasonal ARIMA components.

*Figure 2.2.3 (White Noise Test and Unit Root Test using non-seasonal differencing)*



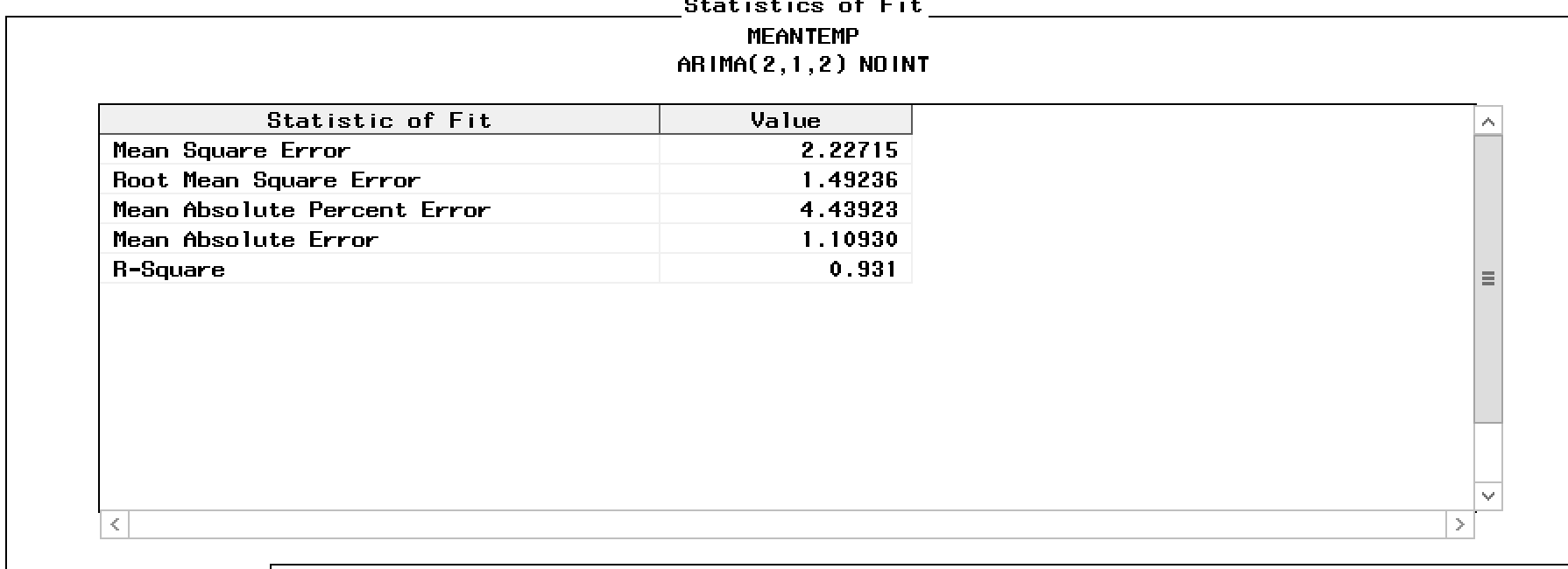
From figure (2.2.3), the white noise test indicates all lags have values less than 0.05, we can reject the null and conclude with reasonable confidence that the series is not White Noise. Also, the unit root test indicates the values are less than 0.05, we can reject the null and conclude with reasonable confidence that this series is stationary (it doesn’t have a unit root).

*Figure 2.2.4 (Fitting ARIMA model)*

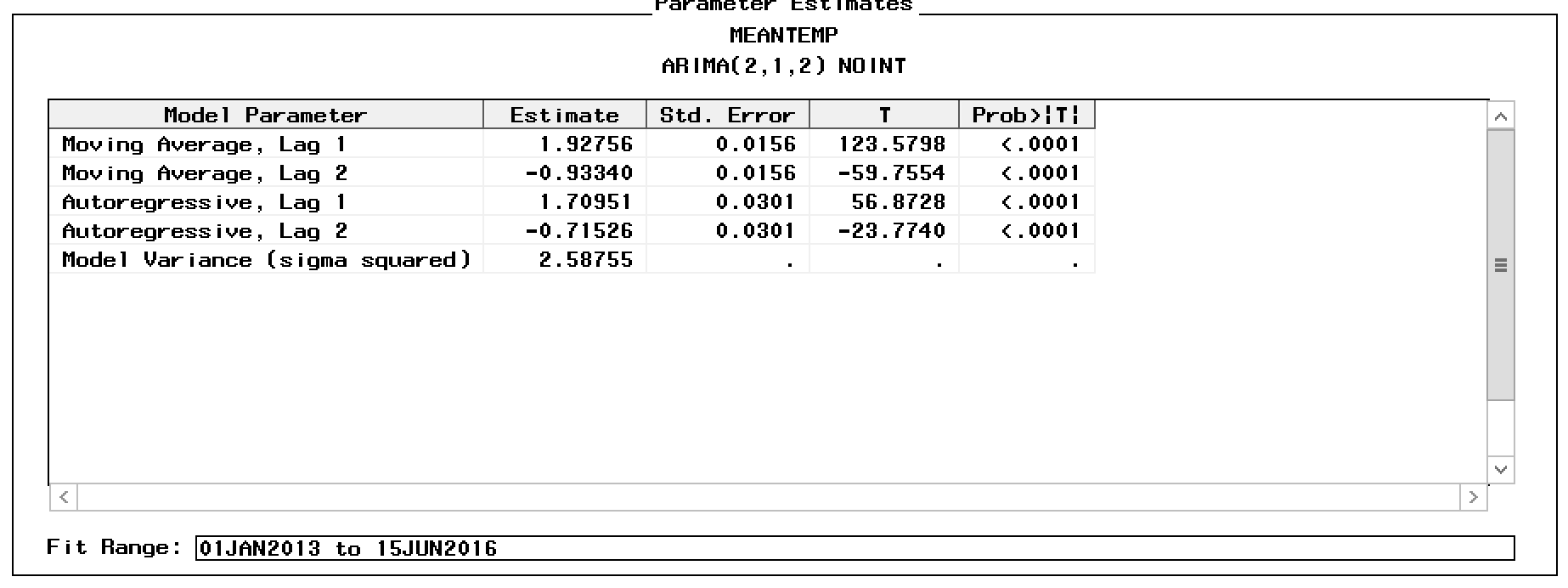


Based on the Mean Absolute Percent Error on figure (2.2.4), we found these models have similar predictive performance since their Mean Absolute Percent Error are close to each other and all around 4.5. We pick ARIMA(2,1,2) which has the lowest error to look into other information.

*Figure 2.2.5 ARIMA(2, 1, 2) Error*

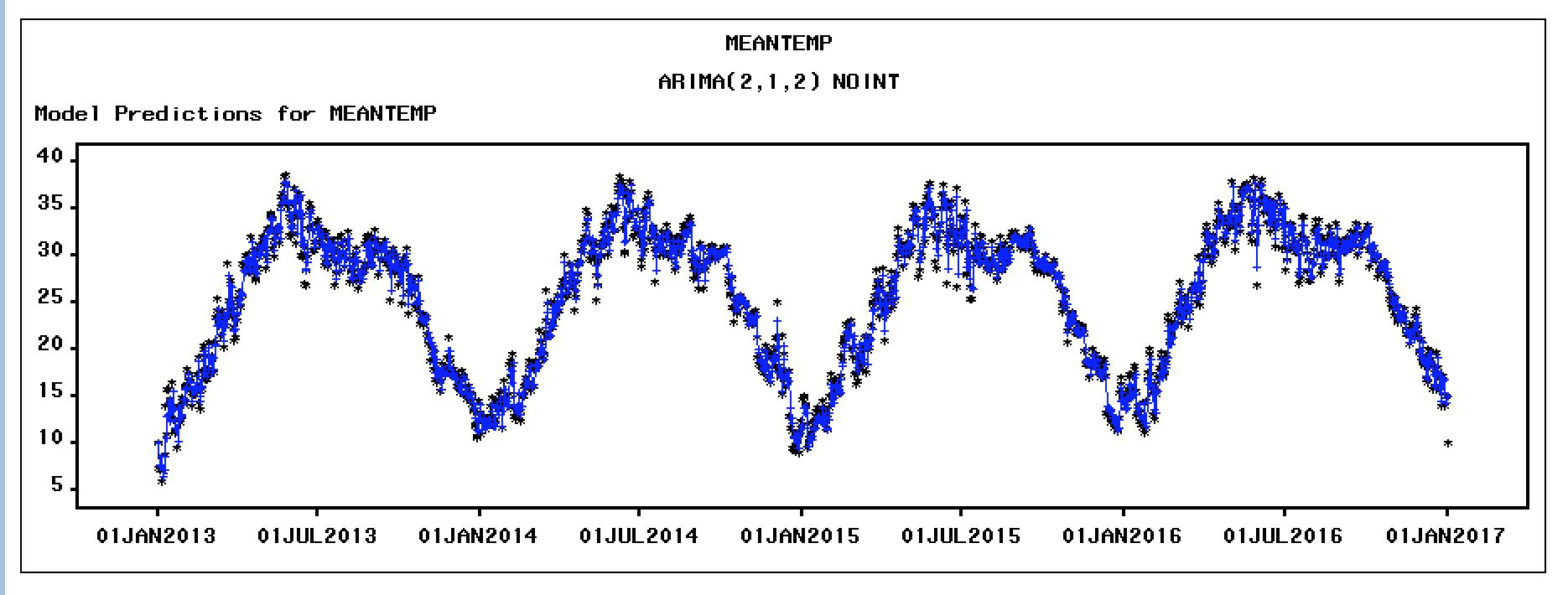


*Figure 2.2.6 ARIMA(2, 1, 2) parameters*



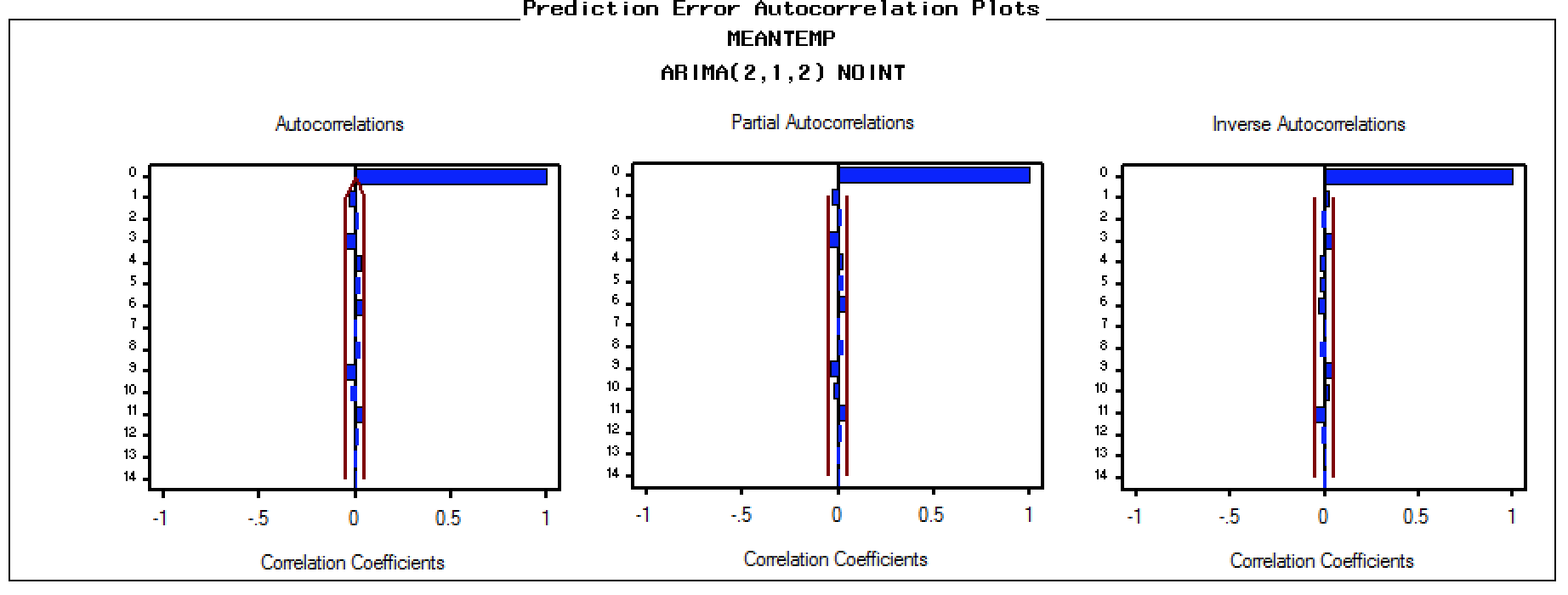
Based on the figure (2.2.6), we can find that both MA and AR coefficients at lag 1 and lag 2 have p-values less than 0.05, which indicate they are significant. Also, because the intercept is not significant, we delete the intercept.

*Figure 2.2.7 (Actual vs. predicted values of the series) --- ARIMA(2,1,2)*



Looking at the Figure (2.2.7) actual vs. predicted graph, we can see ARIMA(2,1,2) model has a good performance in fitting the actual values and the moving trend.

*Figure 2.2.8 (ACF, PACF and Inverse Autocorrelation) --- ARIMA(2,1,2)*



Based on figure 2.2.8, after fitting the ARIMA(2,1,2) model, the series becomes stationary. Both ACF and PACF decays quickly and within the bound after lag 0.

2.3 Comparison of models

Comparing the models above, the mean absolute percent error of Cyclical Model with AR(4) and seasonal dummies and trend model with MA(3) is 7.22 and 6.59. It is higher than the ARIMA model which has the mean absolute percent error of 4.44. Although the difference between the values is not very significant which means they all did a good job in predicting the values, the ARIMA model still has better performance than the other. Thus, the ARIMA(2,1,2) can predict the data more accurately than seasonal dummies and trend models with MA(3).

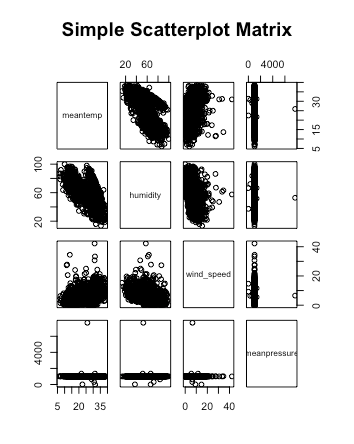
|  |  |  |
| --- | --- | --- |
| Model | Root Mean Square Error | Mean Absolute Percent Error |
| Seasonal Dummies and trend | 2.11 | 6.59 |
| Seasonal Dummies and trend + ARMA(1,3) | 1.54 | 4.69 |
| ARIMA(2,1,2) | 1.49 | 4.44 |
| Cyclical Model | 11.50 | 41.71 |
| Cyclical Model with AR(4) | 2.21 | 7.22 |

3. Multivariate Time Series Models

3.1 Regression model and analysis of regression residuals

Because we use 3 variables: humidity, wind\_speed and mean pressure to predict the mean temperature in Delhi, we decided to fit the model with these variables as predictors. Like the last draft, we choose the hold-out sample as 200 observations.

*Figure 3.1.1 (Scatter Plot of correlations between variables)*



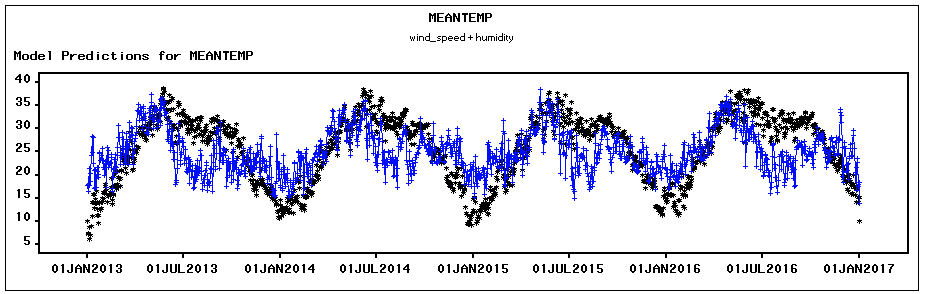
Based on the scatterplot of dependent variable ‘meantemp’ and other predictors, we can see that predictor ‘meanpressure’ doesn’t have much correlation with the ‘meantime’. later, we look into more information about the regression outputs.

*Figure 3.1.2 (Error of Regression model)*



According to Figure 3.1.2, although ‘meanpressure’ is not a great predictor, the two models above have very similar predictive performance and Mean Absolute Percent Error. Then, we will look at predictive vs. actual values graphs.

*Figure 3.1.3 (Predicted values vs. actual values of regression model)*



Based on the figure (3.1.3), we can see that the predicted values have a very similar trend compared to the actual values. But there are some errors between these two values.

*Figure 3.1.4 (ACF and PACF of regression model)*



From the figure (3.1.4), we can find the ACF of the regression model is not stationary, so we take the first difference of this regression model to make it stationary.

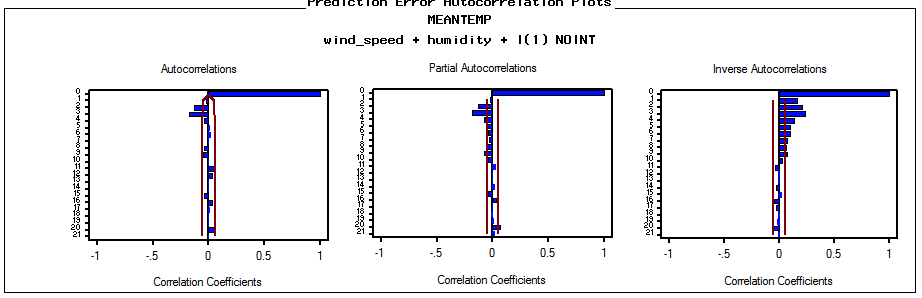
*Figure 3.1.5 (First difference on regression model)*



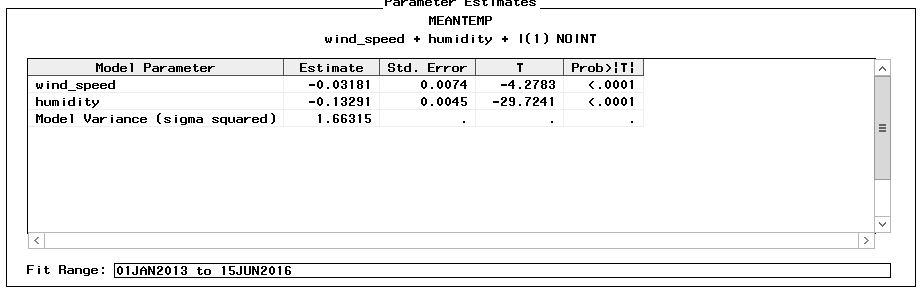
After taking the first difference, we can see the Mean Absolute Percent Error is reduced from 23.14179 to 3.11296. Next, we are going to develop error models based on the regression model with first differencing.

3.2 Error model using regression residuals

*Figure 3.2.1 (ACF and PACF of differencing model)*

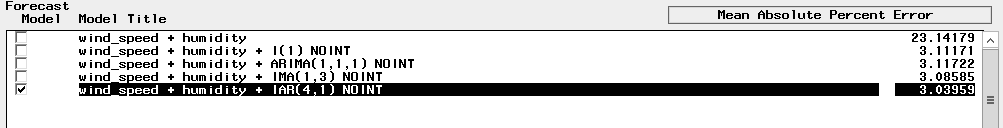


*Figure 3.2.2 (Parameters of differencing model)*

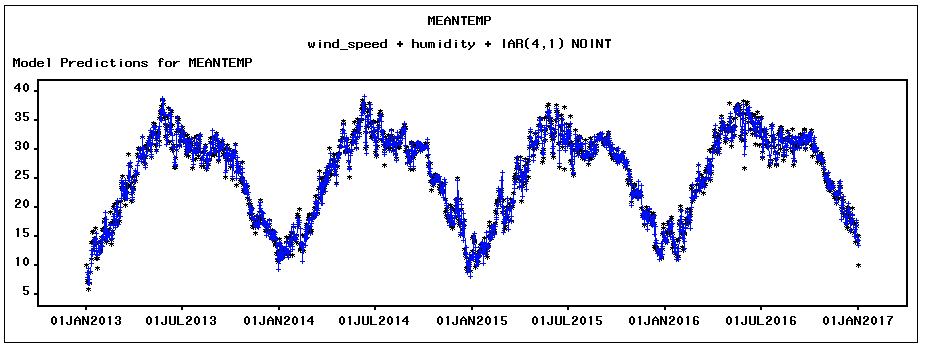


After taking the first difference, the IACF decays quickly, ACF chops off after lag 3 and PACF chops off after lag 4. Therefore, This is very likely to apply to MA(4), AR(3) or ARIMA (1,1,1) models. Because the p-value of the intercept is not significant, so we decide not to add the intercept to our model.

*Figure 3.2.3 (Performance of ARIMA models)*

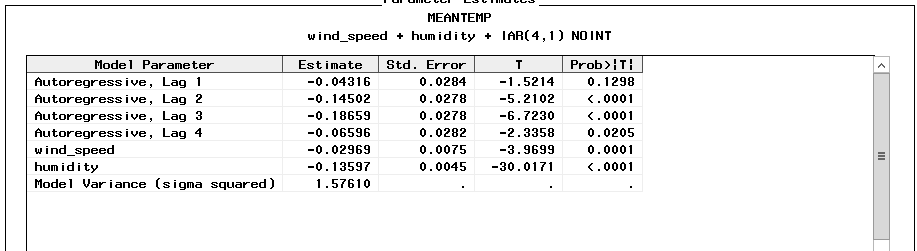


*Figure 3.2.4 (Actual values vs. predicted values of IAR(4,1) )*



After fitting the above models, we found they have similar Mean Absolute Percent Error. Then, we pick the IAR(4,1) model to find the graph of actual vs. predicted values. Figure (3.2.4) indicates the predicted values have a similar pattern as the actual values.

*Figure 3.2.5 (Parameters of IAR (4,1) )*

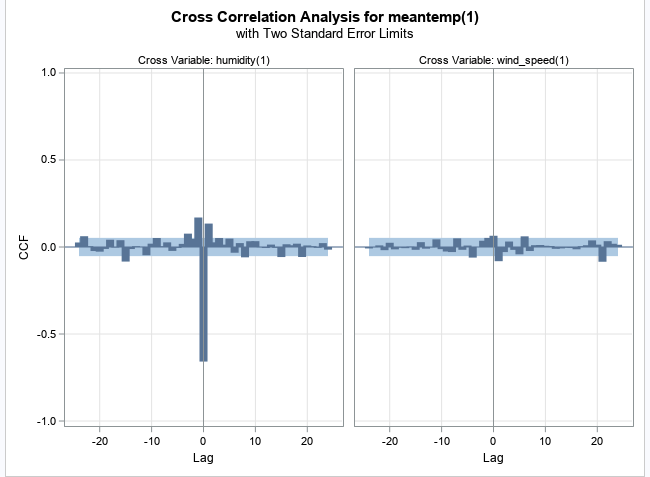


Based on these parameters in the figure (3.2.5), we can be certain that predictors ‘wind\_speed’ and ‘humidity’ are useful in predicting the dependent variable ‘meantemp’.

3.3 Cross correlation analysis

Since only predictors ‘wind\_speed’ and ‘humidity’ are useful in predicting dependent variable ‘meantemp’ and these predictors are not stationary before first differencing, we look at their Cross Correlation Analysis after the first differencing.

*Figure 3.3.1 (Cross Correlation plot)*

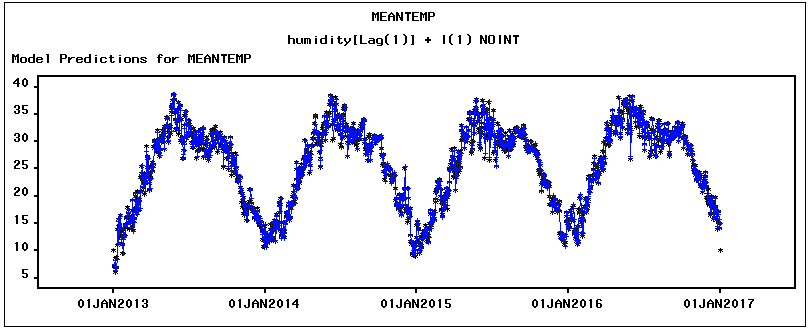


Based on the figure (3.3.1), we can find that variable ‘humidity’ at lag 0, lag 1 and lag -1 have some correlations with the ‘meantemp’ after the first difference. Moreover, we can conclude that we can use the ‘humidity’ value at today and yesterday to predict the dependent variable ‘meantemp’ at today. Later, we are going to fit a cross correlation model based on the results.

*Figure 3.3.2(Performance of error models)*

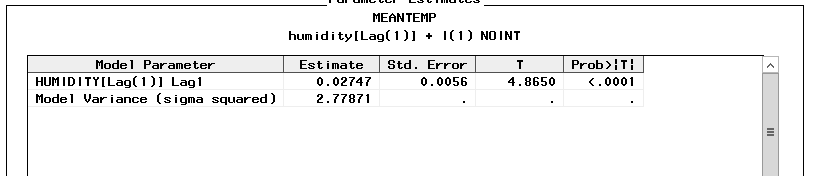
In figure (3.3.2), we fitted 3 cross correlation models. Since only one predictor ‘humidity’ at lag 1 is useful in predicting the dependent variable, we apply ‘humidity’ in the model. Moreover, we added the error terms to the model, but the predicting performance MAPE does not improve much.

*Figure 3.3.3 (Predicted values vs. actual values of first difference)*



Based on the figure (3.3.3), we can see the predicted values and actual values are close and their trends also match each other.

*Figure 3.3.4 (Parameters of first difference)*



On the figure 3.3.4, the predator ‘humidity’ at lag 1 has a significant p-value (less than 0.05). Therefore, we can conclude ‘humidity’ at lag 1 is useful in predicting the dependent variable ‘menatemp’.

4. Conclusion

Finally, we compare the model fit using the square root of model variance estimate. And we compare model predictive performance using MAPE.

|  |  |  |
| --- | --- | --- |
| Models | Square root of model variance | MAPE |
| Seasonal dummies and trend | 2.5602 | 6.59204 |
| Seasonal dummies and trend + ARMA(1,3) | 1.62354 | 4.69438 |
| Cyclical Model | 2.7045 | 41.7666 |
| Cyclical Model + AR(4) | 1.6089 | 7.22072 |
| ARIMA (2,1,2) | 1.6085 | 4.43923 |
| Regression Model | 5.8189 | 23.09708 |
| Regression Model + I(1) | 1.2896 | 3.11296 |
| Regression Model + IAR(4,1) | 1.2554 | 3.03959 |
| Cross-correlation (humidity) | 6.0035 | 21.74557 |
| Cross-correlation (humidity)+ I(1) | 1.6669 | 4.74112 |

Based on the square root of variance, the regression model with variables ‘wind\_speed’ and ‘humidity’ after differencing has the best model fit. Besides the regression model; cyclical model + AR(4), ARIMA(2,1,2) and Cross-correlation (humidity) after differencing all have pretty good model fit. Moreover, the regression model after differencing has the best predictive performance with MAPE (3.03959), which indicates we could receive the best results in predicting mean temperature in Delhi India by using variables ‘wind\_speed’ and ‘humidity’.