Forecasting Industrial Production:
A Study of Materials, Business Equipment, and Nondurable Consumer Goods

by

Shuwei Deng, Austin Frazer, Mengqi Li, and Yinian Lyu

Prepared for

Professor Refik Soyer's DNSC 6219: Time Series Forecasting for Analytics Course, Spring 2018

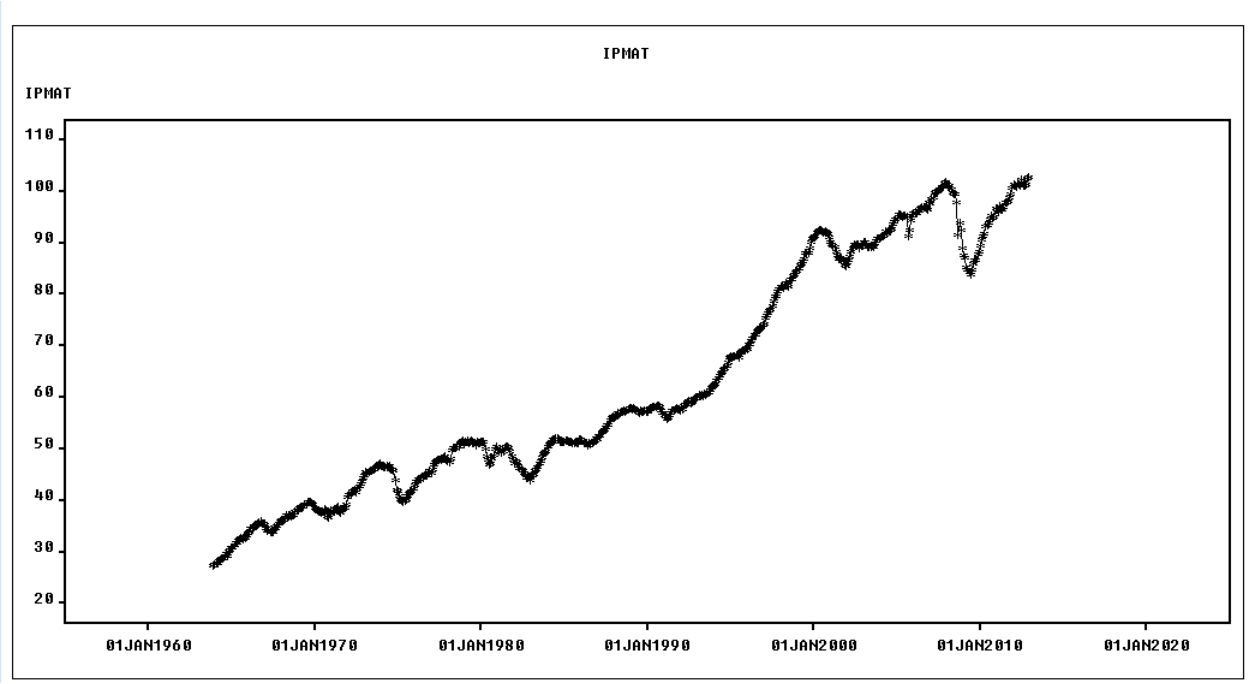
May 15th, 2018

1. Project data background:

This project is about three components of the U.S. monthly industrial production index from December 1963 to December 2012. There are 5 columns in the data, production_index.csv. These columns are year, month, nondurable consumer goods, business equipment, and materials. The data are obtained from the Federal Reserve Bank of St. Louis.

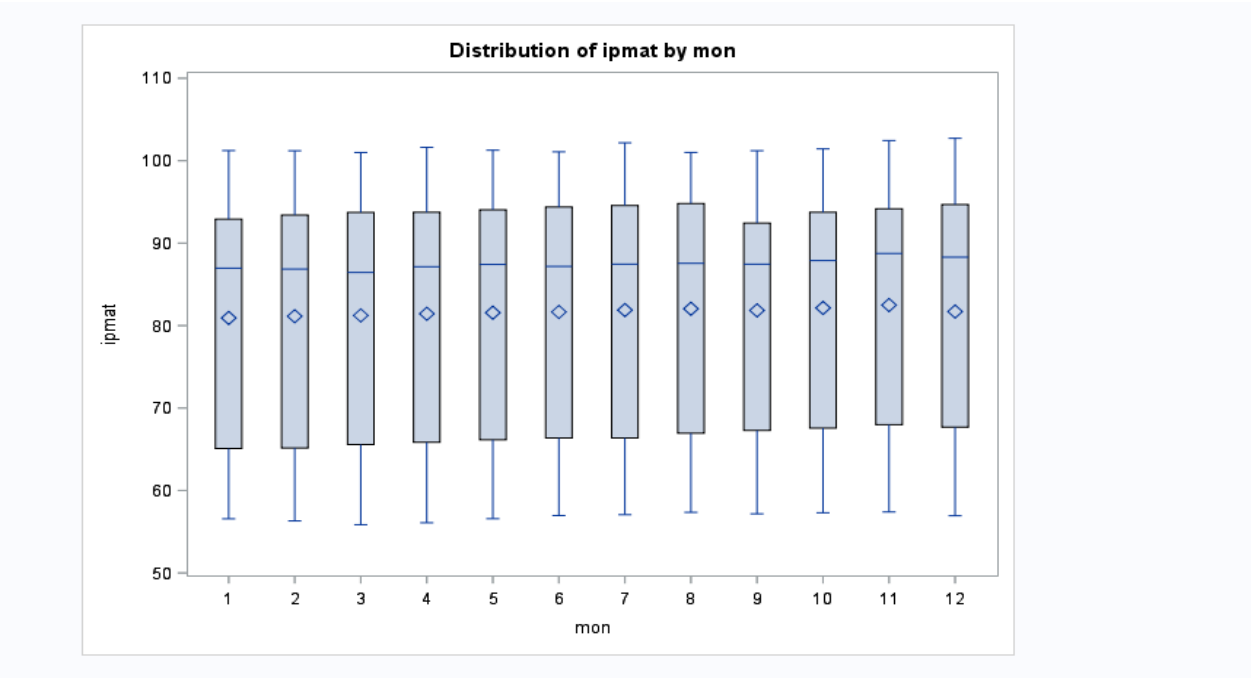
The main series we are interested in the dataset is the Industry Production index with regard to materials (IPMAT) as it could reflect the basic industry conditions within the specified period of time.

Figure 1-1-1: Plot of the main series: IPMAT



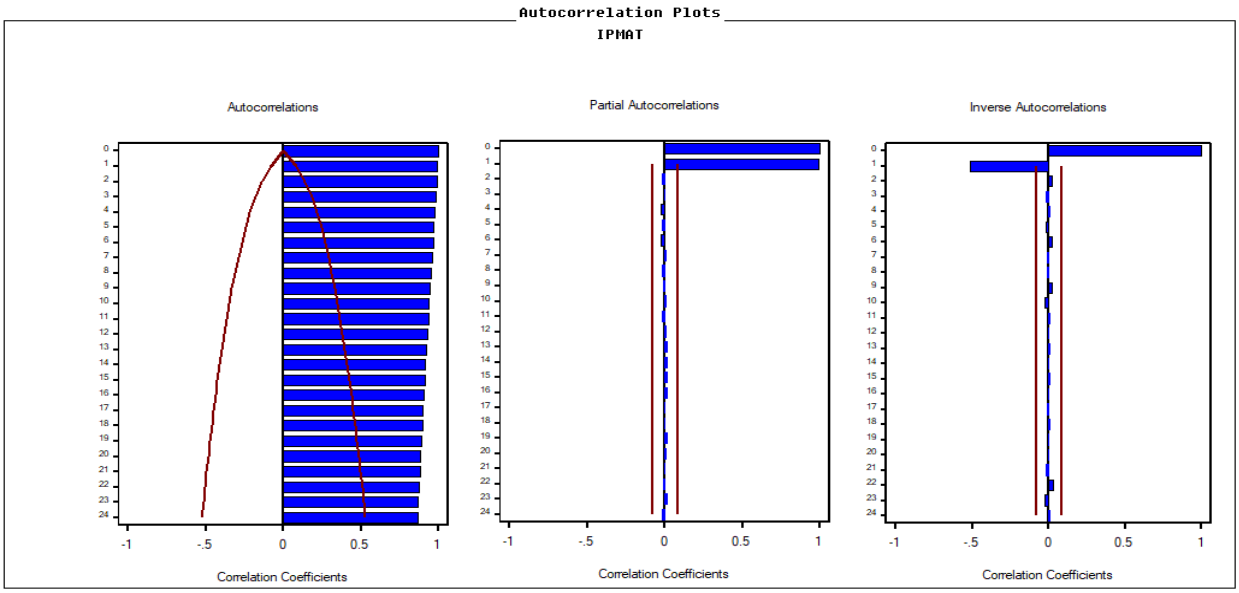
From the plot of the series depicted in Figure 1-1-1, we could easily discover that an upward trend existed in the series as time increased. For the seasonal behavior, we could employ box plots to find more information.

Figure 1-1-2: Box plot of the main series: IPMAT



From the box plot of the main series shown in Figure 1-1-2, we find no obvious seasonality in IPMAT. This is likely due to the fact that the series we downloaded was already deseasonalized.

Figure 1-1-3: Autocorrelation Plots of the main series: IPMAT



The sample ACF plots of the main series here in Figure 1-1-3 displayed a nonstationary feature for the main series.

2. Univariate Time Series Model:

We are beginning our analysis from December of 1988. We noticed in office hours that our model seemed to perform better when just looking at more recent data. So we discarded the first half of the series. We used a 30-month period for the holdout sample for our analysis.

2.1 Deterministic Time Series Models (Linear Trend) and Error model

We started with the Time series models with Linear trend since there is no obvious seasonality in our data as shown in the Figure 1-1-2. Note that the Linear Trend model shown in Figure 2-1-1 does not track very well with the series. The residual of the series is definitely not stationary as the ACF is slowly decaying, as depicted in Figure 2-1-2.

Figure 2-1-1: Plot of the series using Linear Trend Model

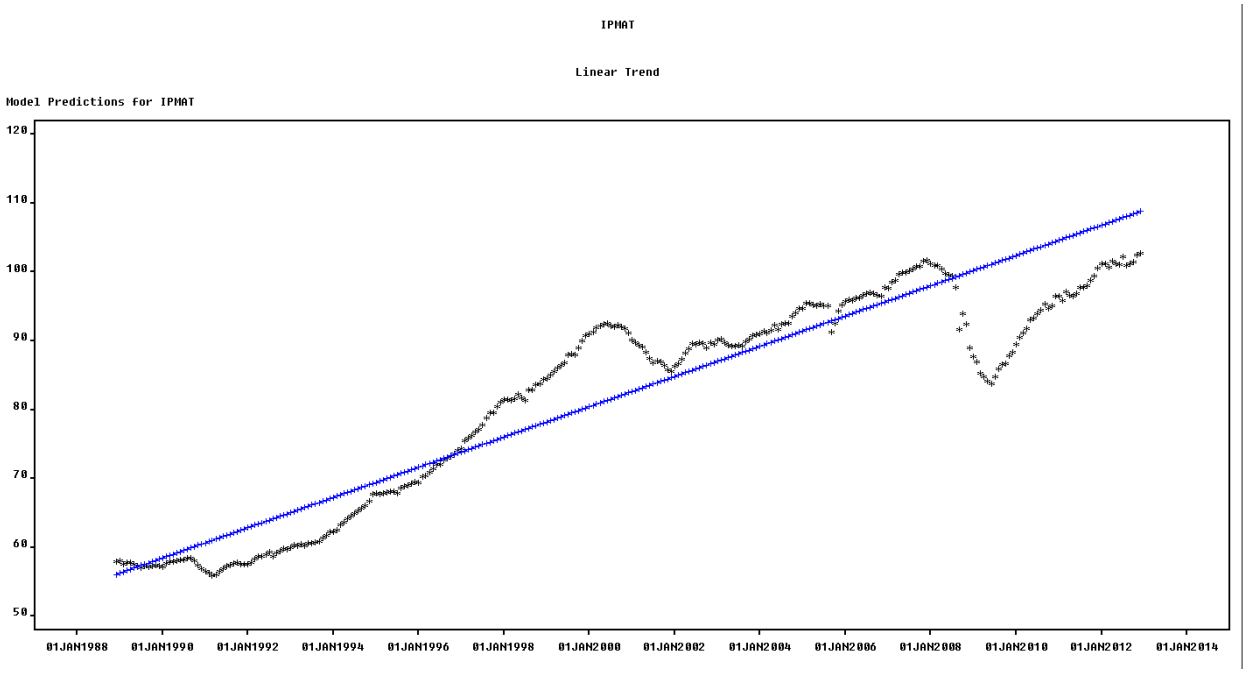


Figure 2-1-2: Prediction Error Autocorrelation Plots of the series using Linear Trend

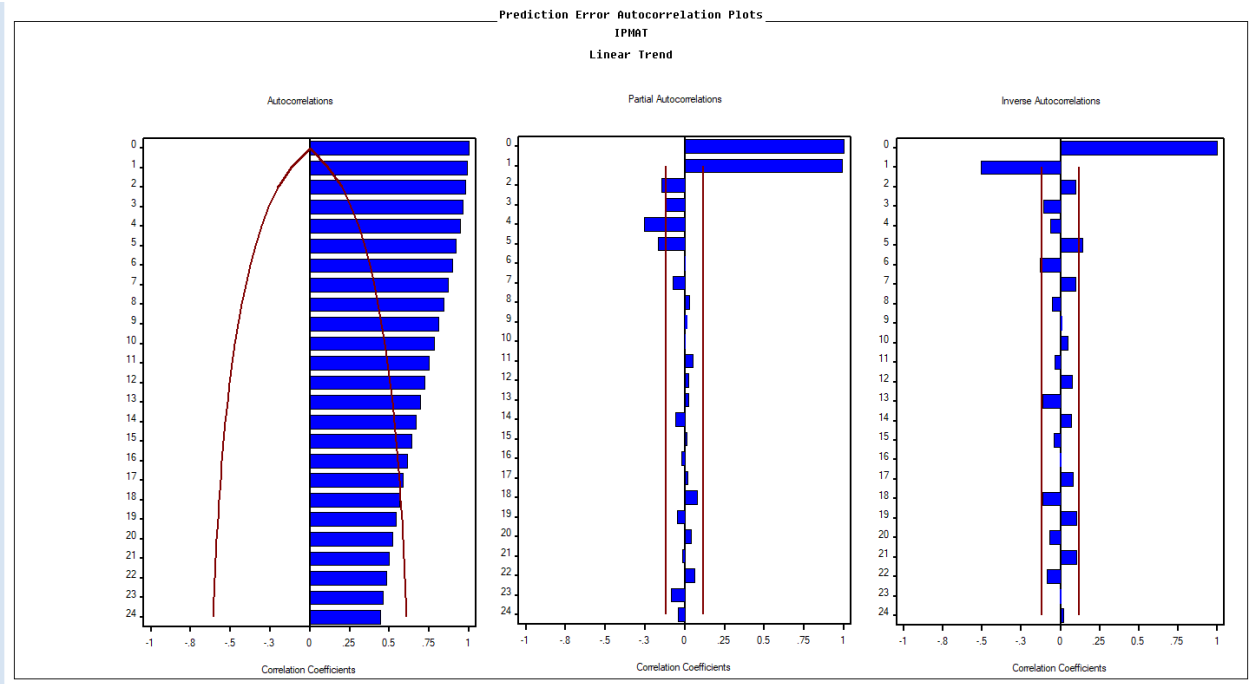


Figure 2-1-3: Parameter Estimates of the series using Linear Trend

Parameter Estimates
IPMAT
Linear Trend

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	-53.92175	3.4755	-15.5150	<.0001
Linear Trend	0.18305	0.0047	38.6491	<.0001
Model Variance (sigma squared)	32.47584	.	.	.

Here is the plot of the series in Quadratic Trend as a comparison. The parameters here are all significant and the model performs better in terms of MAPE. This better performance is visually apparent in Figure 2-1-4 below. The residual of the series is not stationary, as depicted in Figure 2-1-5. The parameters do all turn out to be significant, as shown in Figure 2-1-6.

Figure 2-1-4: Plot of the series using Quadratic Trend

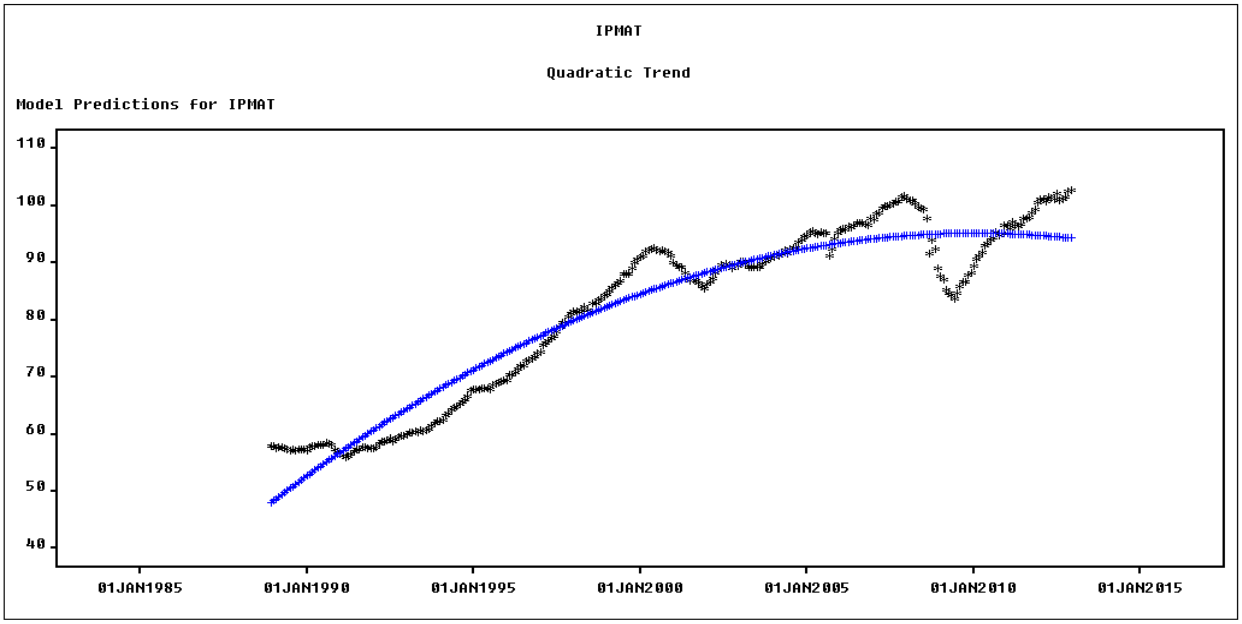


Figure 2-1-5: Prediction Error Autocorrelation Plots of the series using Quadratic Trend

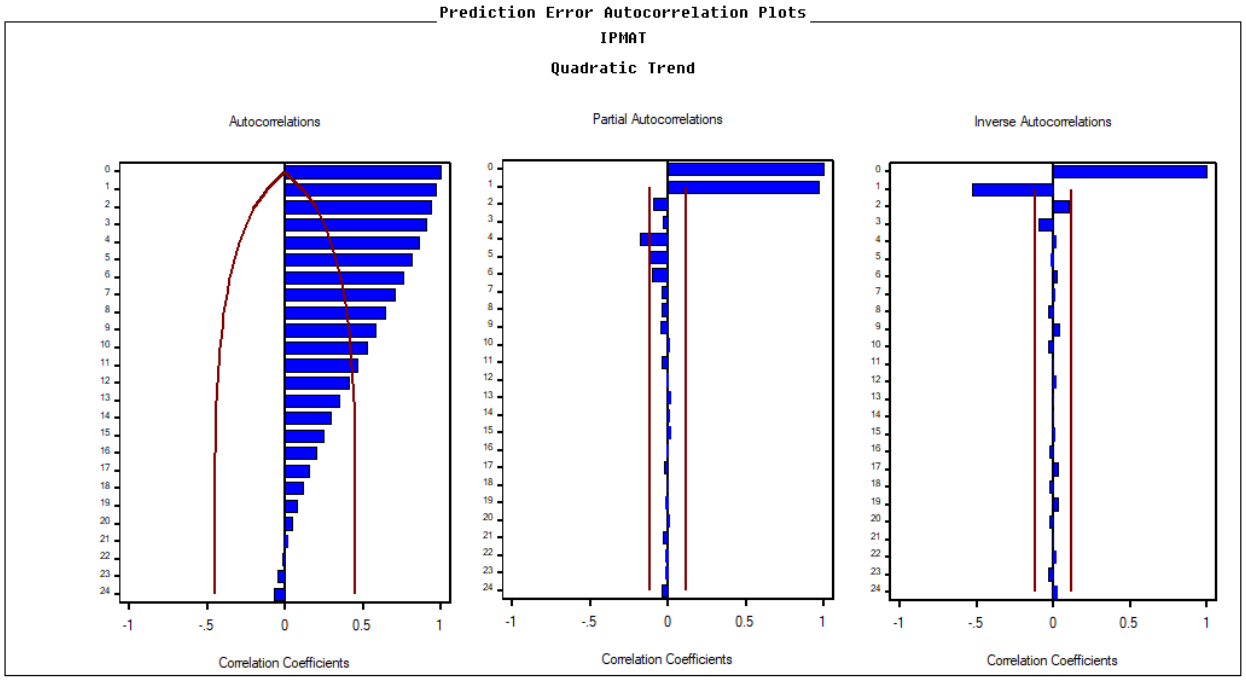


Figure 2-1-6: Parameter Estimates of the series using Quadratic Trend

Parameter Estimates

IPMAT

Quadratic Trend

Model Parameter	Estimate	Std. Error	T	Prob>|T|
Intercept	-379.96542	24.7245	-15.3680	<.0001
Linear Trend	1.18303	0.0734	16.1087	<.0001
Quadratic Trend	-0.0007364	0.000054	-13.6327	<.0001
Model Variance (sigma squared)	18.88937	.	.	.

Fit Range: DEC1988 to JUN2010

2.2 Exponential Smoothing models (only the relevant ones)

Here, we are using the Linear (Holt) Exponential Smoothing model since the series does not have a seasonal trend. The Linear (Holt) Exponential Smoothing model performs well in terms of MAPE among other smoothing models. Note how well the model conforms to the series in Figure 2-2-1. In Figure 2-2-2, the series appears to be stationary or mostly stationary at least. The parameters all seem to be significant in Figure 2-2-3.

Figure 2-2-1: Plot of the series using Linear (Holt) Exponential Smoothing model

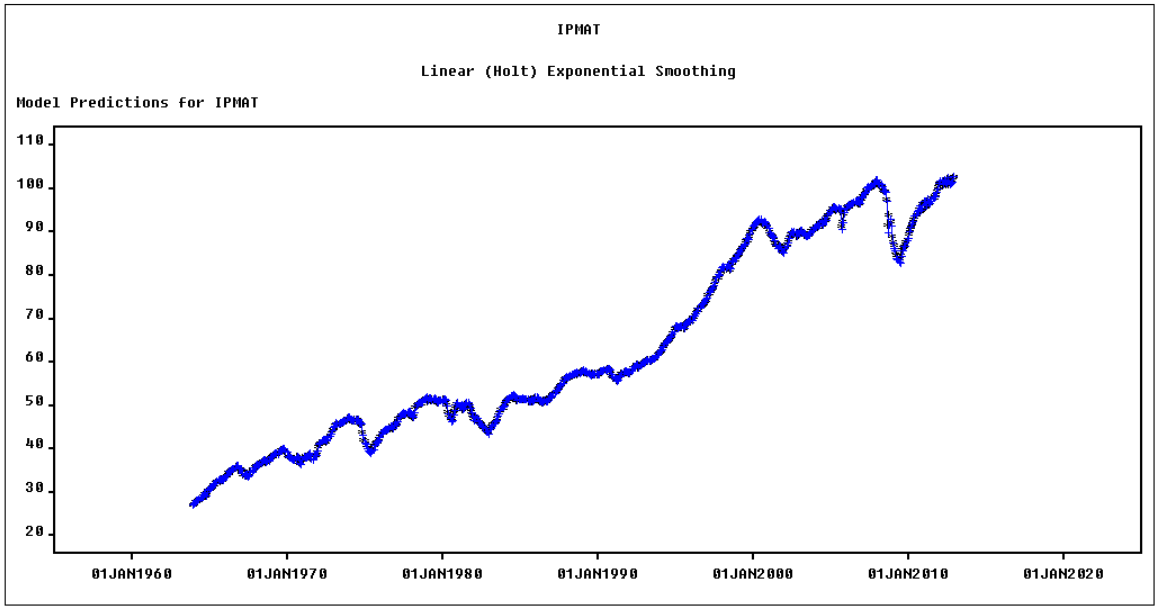


Figure 2-2-2: Prediction Error Autocorrelation Plots of the series using Linear (Holt) Exponential Smoothing model

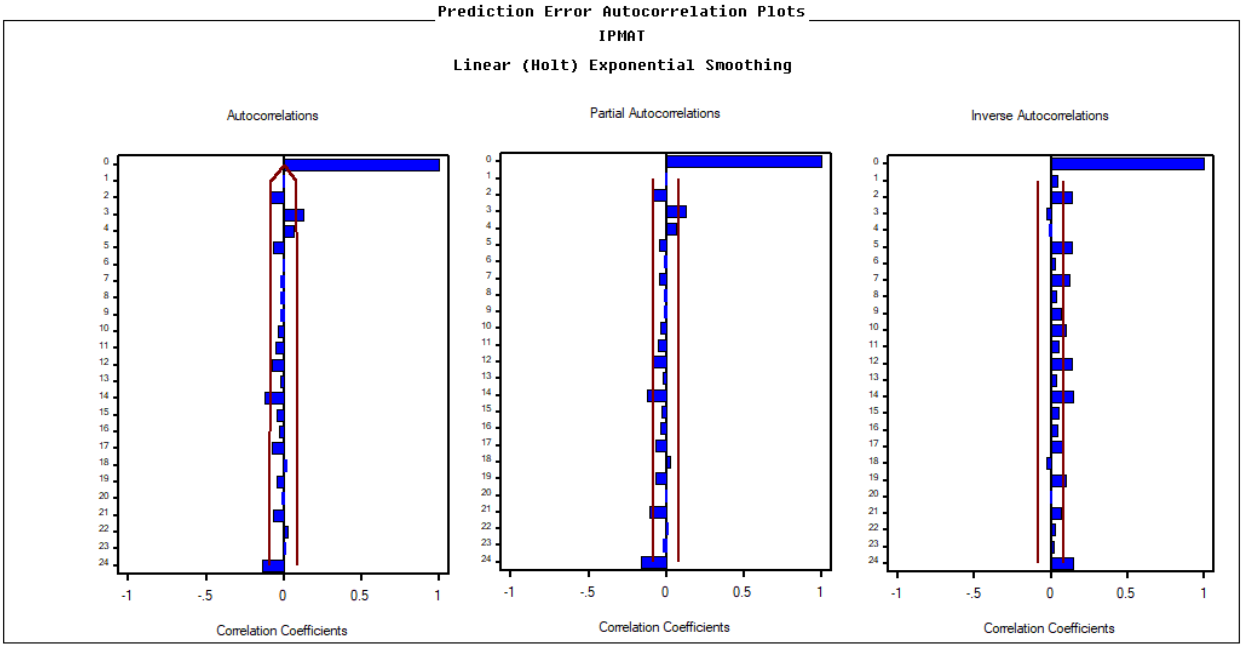


Figure 2-2-3: Parameter estimates of the series using Linear (Holt) Exponential Smoothing model

Parameter Estimates

IPMAT

Linear (Holt) Exponential Smoothing

Model Parameter	Estimate	Std. Error	T	Prob>|T|
LEVEL Smoothing Weight	0.83358	0.0436	19.1280	<.0001
TREND Smoothing Weight	0.30896	0.0429	7.2066	<.0001
Residual Variance (sigma squared)	0.49092	.	.	.
Smoothed Level	93.46109	.	.	.
Smoothed Trend	0.74734	.	.	.

From the plot of the series, we can see that the Linear (Holt) Exponential Smoothing model does a good job in prediction. The sample ACF plot of the series also shows that the residuals of the series are almost stationary.

Given the parameter estimates above, we find the formula for the series to be:

$\alpha = 0.999$and $\beta = 0.2238$

The high value of α implies there is no smoothing in the level estimates. The relatively small value of βimplies not very smooth trend factors. And p-values are all less than 0.05, so they are all significant.

$$L_t = \alpha * Y_t + (1 - \alpha) * (L_{t-1} + T_{t-1})$$

$$T_t = \beta * (L_t - L_{t-1}) + (1 - \beta) * T_{t-1}$$

2.3 ARIMA models (with seasonal ARIMA components if relevant)

We experimented with a few different models. When we first looked at the ACF and PACF, we noticed that there was a gradual descent in the factors. So, we were motivated to take the first difference. We also took the log of the series at the professor's suggestion.

We experimented with a few different ARIMA models and looked at the PACF and ACF of each. We also compared the RMSE of each. We found good performance from the higher-order models, but with them, we had issues with non-significant coefficients on lag 2. So, with the help of Professor Soyer, we implemented a factored ARIMA model. With it, we were able to keep lags 1 and 3 but to drop the unwanted lag 2. Based on the feedback from the first draft, we are using the ARIMA (1,1,(1,3)) factored model, where the (1,3) represents the factors for q in the (p,d,q) ARIMA syntax. Note the good performance of the factored arima model in Figure 2-3-1. The coefficients look good in Figure 2-3-2. And the series looks stationary in Figure 2-3-3.

Figure 2-3-1: Plot of Series with Factored Arima Model

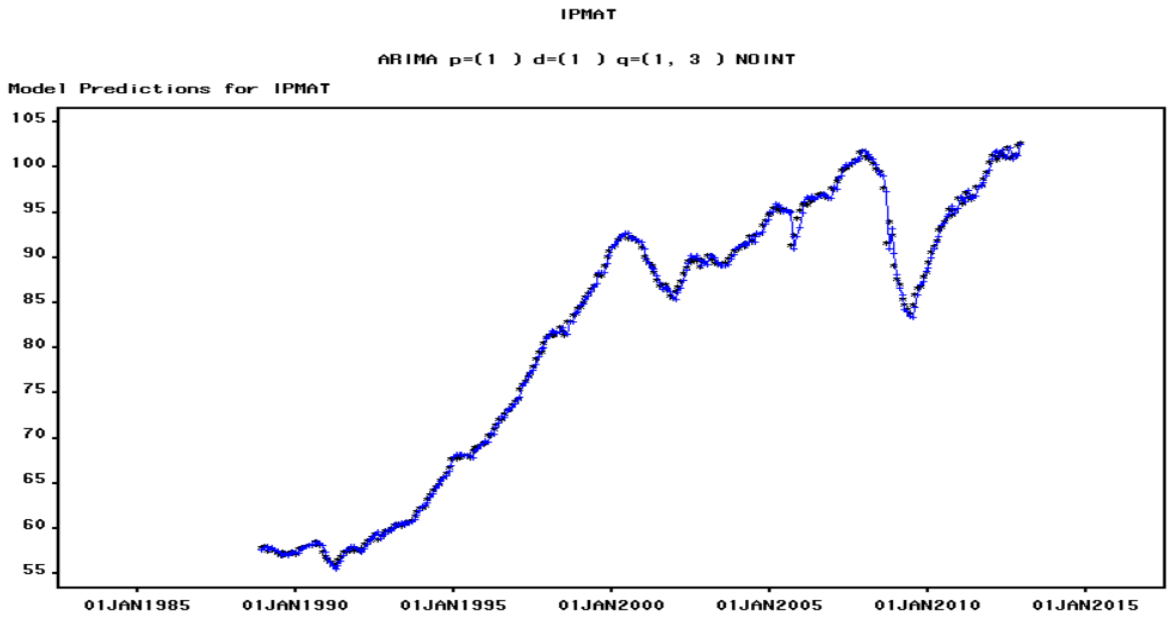


Figure 2-3-2: Prediction Error Autocorrelation plots of Factored ARIMA Model

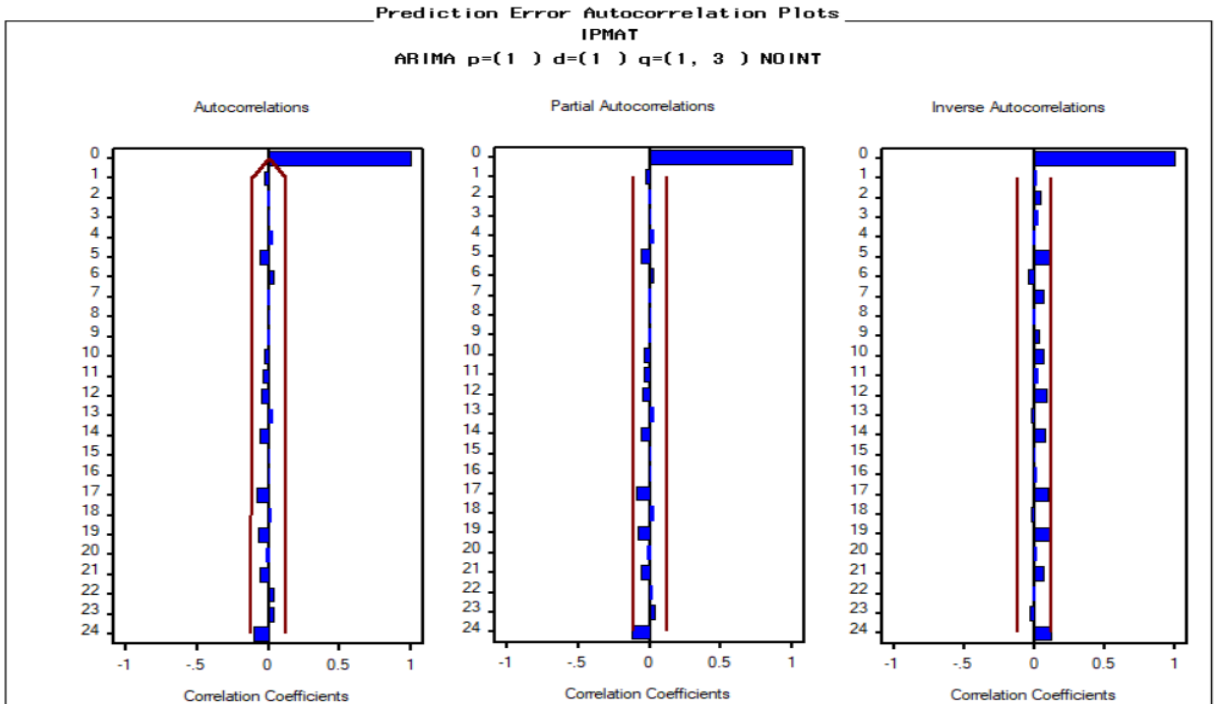


Figure 2-3-3: Parameter Estimates of Factored Arima Model

Parameter Estimates

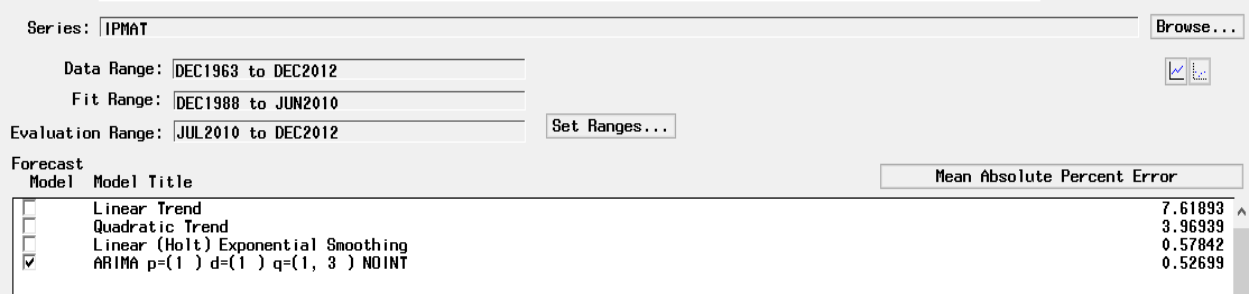
IPMAT

ARIMA p=(1) d=(1) q=(1, 3) NOINT

Model Parameter	Estimate	Std. Error	T	Prob>|T|
MA factor 1 lag 1	0.74441	0.0744	10.0029	<.0001
MA factor 1 lag 3	-0.24586	0.0531	-4.6267	<.0001
AR factor 1 lag 1	0.81095	0.0675	12.0068	<.0001
Model Variance (sigma squared)	0.44770	.	.	.

2.4 Comparison of models (in terms of fit and validation)

Figure 2-4-1: Comparison of models in terms of MAPE



From the figure above, we could compare all the models we have finally used to predict our data. In terms of MAPE, the Factored ARIMA model was the best as the error figure, Figure 2-4-1, shows above.

With regard to validation, the parameter estimates for all four of the attempted models have demonstrated that the coefficients are all statistically significant. Their p-values are smaller than 0.05. We could have enough confidence to believe in each model's prediction. Based on the model variance, (found in the following figures: Figure 2-1-3, 2-1-6, 2-2-3, and 2-3-3; which are the parameter estimates of the four models we employed to predict the data), the Factored ARIMA model has the smallest variance which further proves it is the most accurate model we have.

3. Multivariate Time Series Model

3.1 Identify the transfer function model using Cross Correlation Function

For the analysis of the data using Multivariate time series model, we utilize the cross correlation function to explore the relationship between the two other independent variables, IPBUSEQ and IPNCONGD, and the dependent variable, IPMAT.

First of all, we need to prewhiten the input data since both independent variables are not white noise based on the sample ACF plot. Here, based on the sample ACF and PACF of both series, we decided to use the ARMA(1,1) model to prewhiten the series of IPBUSEQ and IMA(1,(1,9)) to prewhiten the series IPNCONGD..

Finally, we have the two cross correlation function plot between the difference of two independent variables and one dependent variable. These two cross correlation plots can be seen in Figure 3-1-1 and Figure 3-1-2 below.

Figure 3-1-1 Cross Correlations of the differenced series of IPMAT and of IPNCONGD.

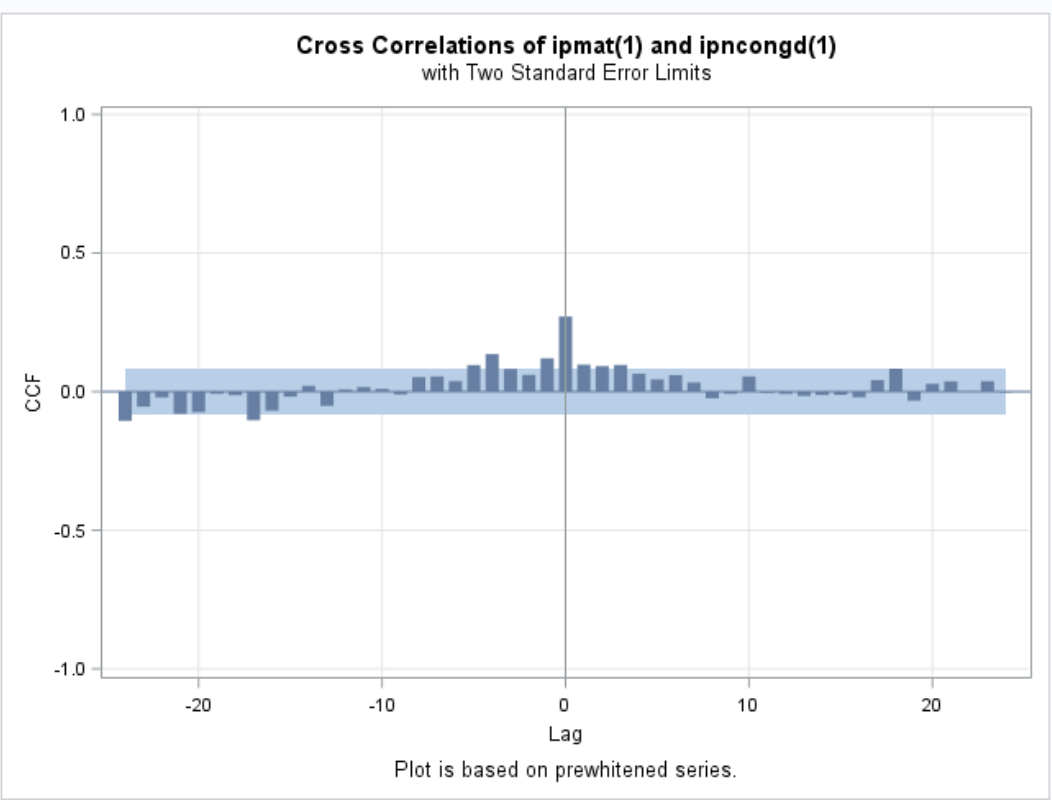
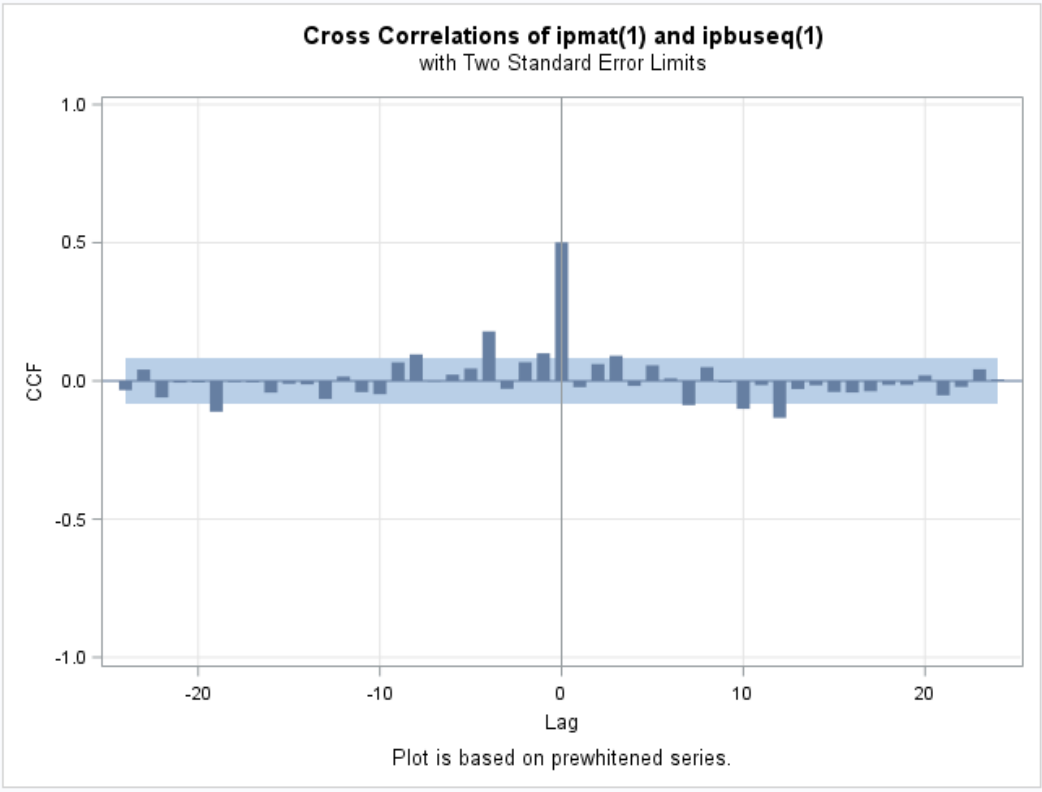


Figure 3-1-2 Cross Correlations of the differenced series of IPMAT and of IPBUSEQ.



Based on the result of these two cross correlation plots, we have identified the parameters for our transfer function model as b=0, r=0 and s=0 as only the contemporary periods are significantly correlated.

3.2 Estimate the series using the transfer function model

With the transfer function model identified as b=0,r=0 and s=0, we next analyze the series using the transfer function model in SAS. The parameter estimates can be viewed in Figure 3-2-1. The Autocorrelation check of the residuals can be seen in Figure 3-2-2. The cross-correlation check is in Figure 3-2-3.

Figure 3-2-1 ULS Estimation of parameter with regard to the difference of ipbuseq

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	0.05764	0.02388	2.41	0.0158	0	Y	0
MA1,1	-0.19735	0.04070	-4.85	<.0001	3	Y	0
NUM1	0.43891	0.02626	16.71	<.0001	0	X2	0

Figure 3-2-2 Autocorrelation Check of Residuals

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	12.34	5	0.0304	0.079	-0.079	0.016	0.047	-0.001	0.075
12	16.19	11	0.1341	0.033	-0.051	-0.015	0.041	-0.001	-0.028
18	28.43	17	0.0401	0.005	-0.125	-0.047	-0.022	-0.043	-0.002
24	49.54	23	0.0011	-0.025	-0.025	-0.044	0.030	-0.046	-0.168
30	64.10	29	0.0002	-0.036	0.073	-0.038	-0.001	0.123	-0.019
36	69.12	35	0.0005	-0.015	0.083	0.008	-0.018	0.021	-0.006
42	78.84	41	0.0003	-0.057	0.074	0.051	0.024	-0.057	0.011
48	85.76	47	0.0005	-0.012	0.036	0.027	-0.068	0.001	-0.063

Figure 3-2-3 Crosscorrelation Check of Residuals with Input ipbuseq

Crosscorrelation Check of Residuals with Input X2									
To Lag	Chi-Square	DF	Pr > ChiSq	Crosscorrelations					
5	21.42	6	0.0015	0.004	0.008	0.117	0.143	0.036	0.025
11	26.89	12	0.0080	-0.045	-0.013	0.030	-0.027	-0.074	0.006
17	40.02	18	0.0021	-0.106	-0.012	-0.048	-0.028	-0.030	-0.084
23	44.81	24	0.0061	-0.006	0.010	-0.004	-0.078	0.008	-0.043
29	51.02	30	0.0097	0.018	0.028	-0.068	-0.025	0.065	0.000
35	59.23	36	0.0087	0.020	-0.011	-0.034	0.074	-0.072	-0.041
41	93.91	42	<.0001	0.171	-0.124	-0.085	-0.009	-0.062	0.059
47	99.63	48	<.0001	0.059	-0.032	-0.000	0.072	0.006	-0.003

For the transfer function model between IPBUSEQ and IPMAT, the crosscorrelation check of residuals with input here is smaller than 0.2 in absolute values such that the adequacy of the transfer function model is recognized. The formula is $(Yt - Yt - 1) = 0.058 + 0.439(Xt - Xt - 1) + \xi t \quad \xi t = \epsilon t + 0.197\epsilon t - 3$. The autocorrelation checks for the residual also proves the adequacy of the model since most residuals are not correlated. The RMSE is 9.39 for this model.

Next, we look at the transfer function model between IPNCONGD (Industrial Non-Durable Consumer Goods) and IPMAT (Industrial Materials). The parameter estimates can be viewed in Figure 3-2-4 below. The Autocorrelation check of the residuals can be seen in Figure 3-2-5. The cross-correlation check is found in Figure 3-2-6.

Figure 3-2-4 ULS Estimation of parameters with regard to the difference of ipncongd

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
MU	0.10564	0.02425	4.36	<.0001	0	Y	0
NUM1	0.26135	0.04133	6.32	<.0001	0	X1	0

Figure 3-2-5 Autocorrelation Check of Residuals

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	132.92	6	<.0001	0.226	0.160	0.289	0.208	0.085	0.116
12	144.42	12	<.0001	0.083	0.065	0.045	-0.018	-0.019	-0.074
18	168.16	18	<.0001	-0.022	-0.135	-0.074	-0.050	-0.091	-0.065
24	194.63	24	<.0001	-0.076	-0.057	-0.108	-0.024	-0.050	-0.139
30	200.33	30	<.0001	-0.018	0.006	-0.047	0.035	0.071	-0.018
36	218.55	36	<.0001	0.015	0.006	-0.005	-0.108	0.000	0.131
42	230.51	42	<.0001	-0.130	-0.029	-0.002	-0.010	-0.017	0.029
48	243.10	48	<.0001	-0.050	0.078	0.053	-0.079	-0.041	-0.022

Figure 3-2-6 Crosscorrelation Check of Residuals with Input ipncongd

Crosscorrelation Check of Residuals with Input X1									
To Lag	Chi-Square	DF	Pr > ChiSq	Crosscorrelations					
5	18.71	6	0.0047	0.026	0.102	0.090	0.080	0.057	0.055
11	25.29	12	0.0135	0.061	0.015	-0.017	0.010	0.080	-0.019
17	27.44	18	0.0711	0.023	-0.002	-0.024	0.008	-0.014	0.048
23	35.98	24	0.0552	0.080	-0.059	0.032	0.042	-0.023	0.038
29	41.23	30	0.0831	0.019	-0.071	-0.029	-0.049	-0.011	-0.014
35	43.61	36	0.1793	-0.001	-0.047	-0.011	-0.023	-0.000	0.035
41	51.12	42	0.1581	-0.028	0.021	-0.035	-0.056	-0.071	-0.046
47	81.44	48	0.0018	-0.061	-0.057	-0.173	-0.099	0.038	-0.056

The cross-correlation check of residuals with input ipncongd is bigger than 0.05 for lags from 17 to 41 which could prove the adequacy of the transfer function model. The formula is $(Yt - Yt - 1) = 0.106 + 0.261(Xt - Xt - 1) + \xi t$. $\xi t = \epsilon t - 0.272\epsilon t - 1 + 0.143\epsilon t - 9$. The residual of the model is not white noise since the p-value related to the ACF is smaller than 0.05 for all lags. The RMSE for the model is 21.804.

3.3 Bivariate Input Transfer Function Model

We did some basic explorations with regard to the bivariate input transfer function model, the two inputs here are the IPNCONGD and IPBUSEQ, the output is IPMAT.

Figure 3-3-1 Parameter Estimates with regard to Bivariate Input TF Model

Maximum Likelihood Estimation							
Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag	Variable	Shift
NUM1	0.01781	0.04240	0.42	0.6744	0	X1	1
NUM1,1	-0.04472	0.04242	-1.05	0.2918	1	X1	1
NUM2	0.14369	0.03051	4.71	<.0001	0	X2	1
DEN1,1	0.62979	0.08742	7.20	<.0001	1	X2	1

Figure 3-3-2 Autocorrelation Check of Residuals

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	Autocorrelations					
6	39.09	6	<.0001	0.084	0.009	0.190	0.141	0.010	0.054
12	43.55	12	<.0001	0.021	0.031	0.015	-0.027	-0.026	-0.066
18	57.19	18	<.0001	-0.035	-0.103	-0.049	-0.050	-0.077	-0.010
24	73.02	24	<.0001	-0.059	-0.015	-0.075	0.001	0.001	-0.129
30	86.57	30	<.0001	0.022	0.042	-0.060	0.058	0.110	0.027
36	111.59	36	<.0001	0.030	0.040	0.042	-0.062	0.031	0.176
42	120.65	42	<.0001	-0.104	-0.012	0.028	0.007	-0.015	0.048
48	125.34	48	<.0001	-0.030	0.038	0.028	-0.052	-0.030	-0.025

Based on the autocorrelation check for the residuals, we could conclude the adequacy of the model. The numerator factors for the input X1 is 0.018+0.045*B and the denominator factors for the input X2 is 1-0.630*B. Using SAS, we could find out that this model has a RMSE of 0.250861.

3.4 Comparison between Transfer Function Models

The three distinct transfer function models above have been validated to be able to predict the data in terms of autocorrelation checks and cross-correlation checks. But the bivariate transfer function model was the most accurate model with regard to the RMSE as a estimator of fit.