Ruoheng Du

Regression and Multivariate Data Analysis

Prof. Jeff Simonoff

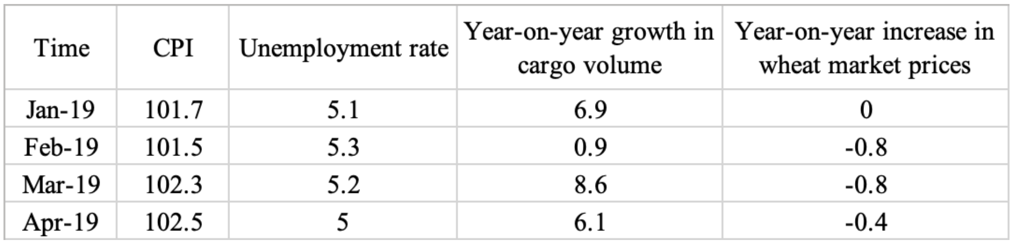
Apr. 11, 2023

Time Series Analysis of Consumer Price Index

The consumer price index (CPI) is a relative economic indicator that reflects the pattern and magnitude of price changes in consumer goods and services purchased by urban and rural populations over a given time period. In general, the CPI for a specific month in the prior year is set to 100, serving as the base year or reference point for computing inflation rates. For example, if the CPI for a particular month is 110, it signifies that the cost of living has risen by 10% since the same month the previous year.

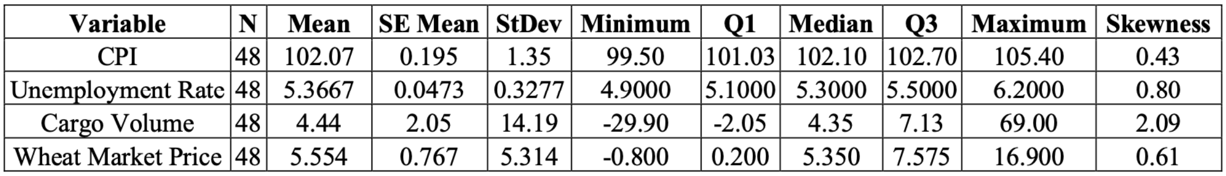
Accurately forecasting CPI is critical for a wide range of stakeholders, including governments, businesses, investors, and consumers, because it can provide insight into the direction of inflation, allowing policymakers to make informed decisions about monetary policy, interest rates, and fiscal policies, and companies and investors to make informed investment decisions. It is appropriate to examine the unemployment rate, cargo volume growth year over year (i.e., cargo volume in my following analysis), and wheat market price increase year over year (i.e., wheat market price in my following analysis) as potential CPI predicting variables because they are indications of economic growth, supply and demand dynamics, and market forces that might influence prices. To be more specific, here, the unemployment rate is calculated using aggregated data from the sample survey of urban labor force situation, specifically the ratio of the number of urban surveyed unemployed to the sum of the number of urban surveyed employed and the number of urban surveyed unemployed. Moreover, the year-over-year cargo volume growth is a measure of the change in the amount of cargo shifted by multiple forms of transportation, such as railways, roadways, waterways, and airplanes, over a year. It is derived by comparing the volume of cargo moved in a specific period to the volume transported in the prior year during the same period. Finally, because wheat is an important grain crop in China, the market price of wheat is added as a predicting variable. This price is transacted in the medium-sized marketplace. The marketplace is a set area where the market manager operates and maintains the market, and where the surrounding urban and rural populations assemble at specific times and locations to trade spot commodities, and it is an important venue for trading grain crops in China.

The data for this report comes from the National Bureau of Statistics of China (http://www.stats.gov.cn). I downloaded monthly data from the website from January 2019 to December 2022, for a total sample of 48. Here are the first four rows of data.

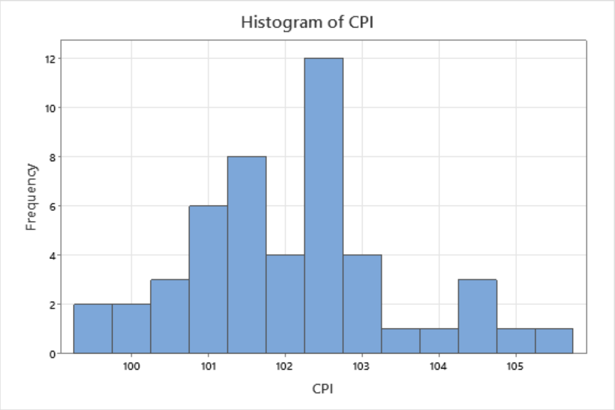


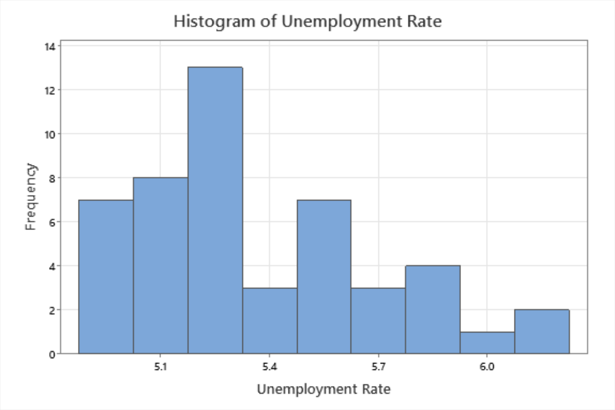
Time	CPI	Unemployment rate	Year-on-year growth in cargo volume	Year-on-year increase in wheat market prices
Jan-19	101.7	5.1	6.9	0
Feb-19	101.5	5.3	0.9	-0.8
Mar-19	102.3	5.2	8.6	-0.8
Apr-19	102.5	5	6.1	-0.4

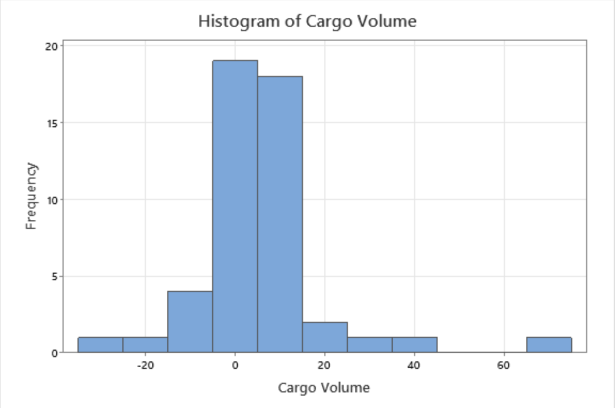
Let’s first look at some basic statistics of the data.

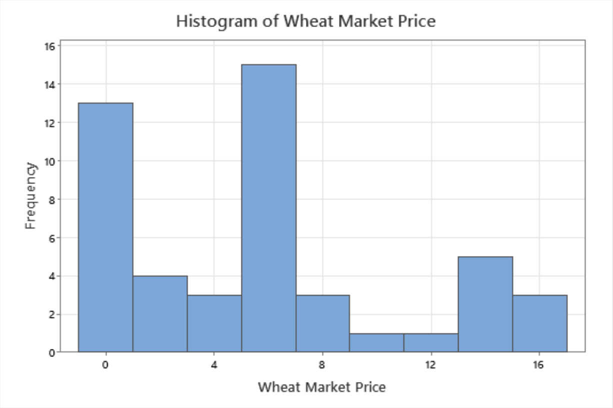


Variable	N	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum	Skewness
CPI	48	102.07	0.195	1.35	99.50	101.03	102.10	102.70	105.40	0.43
Unemployment Rate	48	5.3667	0.0473	0.3277	4.9000	5.1000	5.3000	5.5000	6.2000	0.80
Cargo Volume	48	4.44	2.05	14.19	-29.90	-2.05	4.35	7.13	69.00	2.09
Wheat Market Price	48	5.554	0.767	5.314	-0.800	0.200	5.350	7.575	16.900	0.61



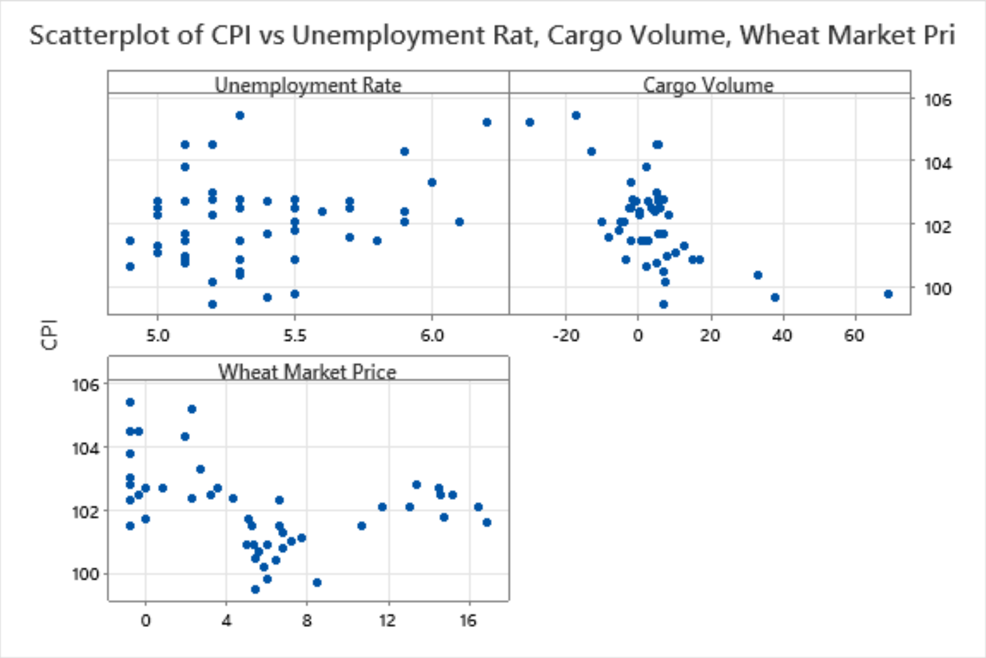






The target variable, CPI, has a mean of 102.07 and a standard deviation of 0.195. For predicting variables, the mean of unemployment rate, cargo volume, and wheat market price are 5.3667%, 4.44%, and 5.554% respectively. Furthermore, as seen by the histogram and skewness value of these variables, cargo volume is somewhat skewed to the right. However, because the cargo volume data has negative values, I fail to utilize logarithm to deal with the right tail. Besides, although the CPI and unemployment rate are slightly tilted to the right and do not have negative values, my attempt to take the natural logarithm of these two variables is inconsequential.

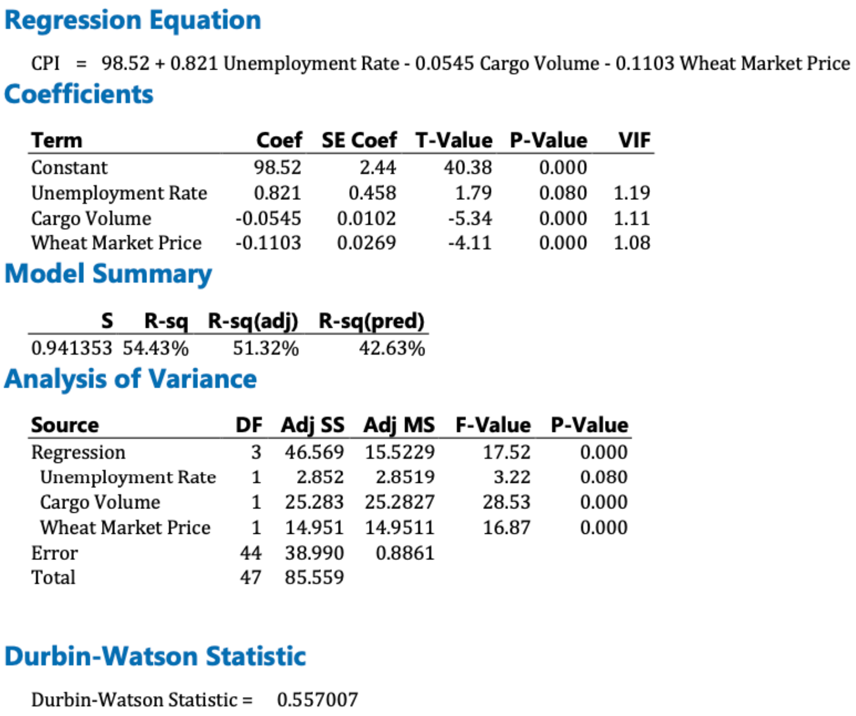
Here is the scatterplot of the target variable versus all predicting variables.



There seems to be a linear relationship between CPI and cargo volume. But it is quite unclear for me to discover the relationship between CPI and the unemployment rate as well as wheat market price now. In the simple multiple regression model, I propose the below expression:

$$CPI = \beta_o + \beta_1 * Unemployment\ Rate + \beta_2 * Cargo\ Volume + \beta_3 * Wheat\ Market\ Price + random\ error.$$

The below is the least squares regression result.



Regression Equation

CPI = 98.52 + 0.821 Unemployment Rate - 0.0545 Cargo Volume - 0.1103 Wheat Market Price

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	98.52	2.44	40.38	0.000	
Unemployment Rate	0.821	0.458	1.79	0.080	1.19
Cargo Volume	-0.0545	0.0102	-5.34	0.000	1.11
Wheat Market Price	-0.1103	0.0269	-4.11	0.000	1.08

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.941353	54.43%	51.32%	42.63%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	46.569	15.5229	17.52	0.000
Unemployment Rate	1	2.852	2.8519	3.22	0.080
Cargo Volume	1	25.283	25.2827	28.53	0.000
Wheat Market Price	1	14.951	14.9511	16.87	0.000
Error	44	38.990	0.8861		
Total	47	85.559			

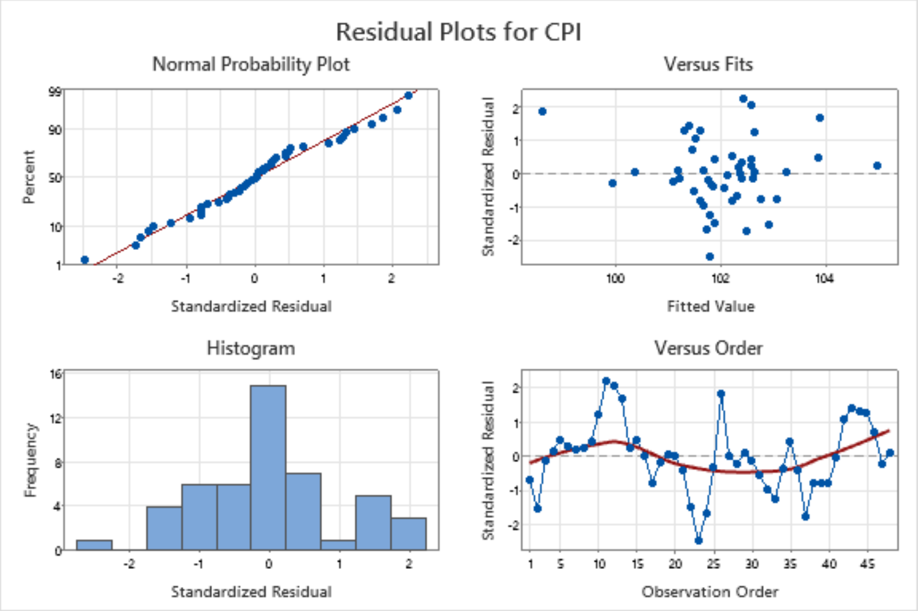
Durbin-Watson Statistic

Durbin-Watson Statistic = 0.557007

Overall, the F-statistic of this regression has a p-value of 0.000, which is statistically significant, and this regression is quite strong as the R-squared value is 54.43% and adjusted R-squared value is 51.32%. In other words, the unemployment rate, cargo volume, and wheat market price are able to account quite well for the observed variability in CPI. The constant coefficient indicates that when the unemployment rate, the year-over-year growth of cargo volume, and the year-over-year increase in wheat market price equal zero, the estimated expected CPI is 98.52. But this point does not have any practical interpretation, since it is not possible to have zero unemployment rate, and there is no such value present in this dataset. As for the unemployment rate coefficient, holding the other two predicting variables fixed, one percentage point increase in the unemployment rate is associated with an estimated expected increase of 0.821 percentage point in CPI. Similarly, for cargo volume coefficient, holding the other two predicting variables fixed, one percentage point increase in the year-over-year growth of cargo volume is associated with an estimated expected decrease of 0.0545 percentage point in

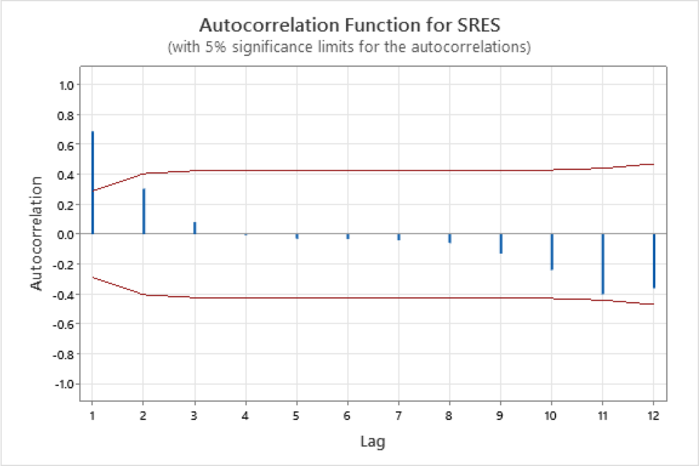
CPI, while for wheat market price coefficient, holding the other two predicting variables fixed, one percentage point increase in the year-over-year increase in wheat market price is associated with an estimated expected decrease of 0.1103 percentage point in CPI. The standard error of the estimate of 0.941353 says that this model could be used to predict CPI within ±1.882706, roughly 95% of the time. All the coefficients in this model are statistically significant as their p-values are smaller than any reasonable significance level. Besides, multicollinearity does not seem to be a problem in this regression, as the VIF values of all the variables are obviously small.

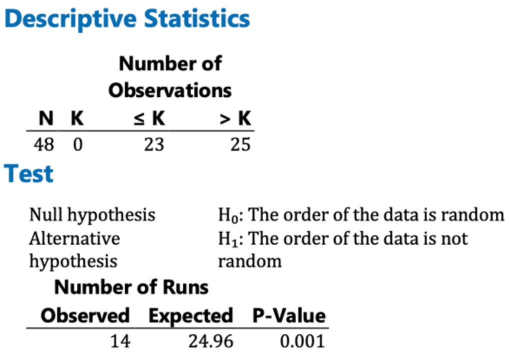
Now, let's take a look at the "four in one" plot to see if our assumptions of least squares regression hold true.



There is a lack of evidence to support the violation of the first assumption and the presence of non-zero expected error since there are no identifiable subgroups exhibiting a consistent deviation from the zero line in terms of residuals. Regarding the second assumption of homoscedasticity, on the plot of standardized residuals against fitted values, the standardized residuals appear to be more tightly clustered around the center, with some points scattered around, suggesting a violation of the assumption of constant variance. The third assumption implies that the errors should have no correlation with one another, but this is only partially fulfilled as a discernible pattern in the plot of standardized residuals versus order indicates a notable autocorrelation between the errors, which is further supported by the smoother. Lastly, the fourth assumption, that errors are normally distributed, is only slightly violated. Although the normal probability plot shows that most of the points are situated along the line, the histogram reveals the presence of a relatively large negative standardized residual with a value less than -2, indicating non-normality.

Let's further check whether there is autocorrelation.





Descriptive Statistics

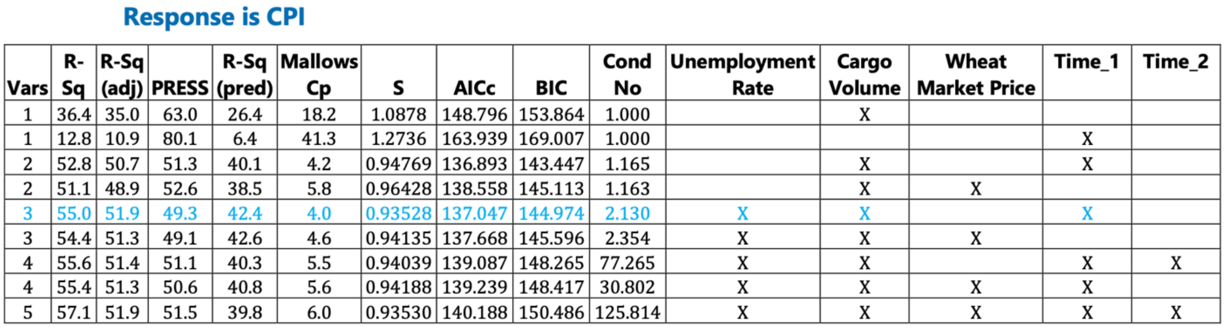
		Number of Observations	
N	K	≤ K	> K
48	0	23	25

Test

Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

Number of Runs		
Observed	Expected	P-Value
14	24.96	0.001

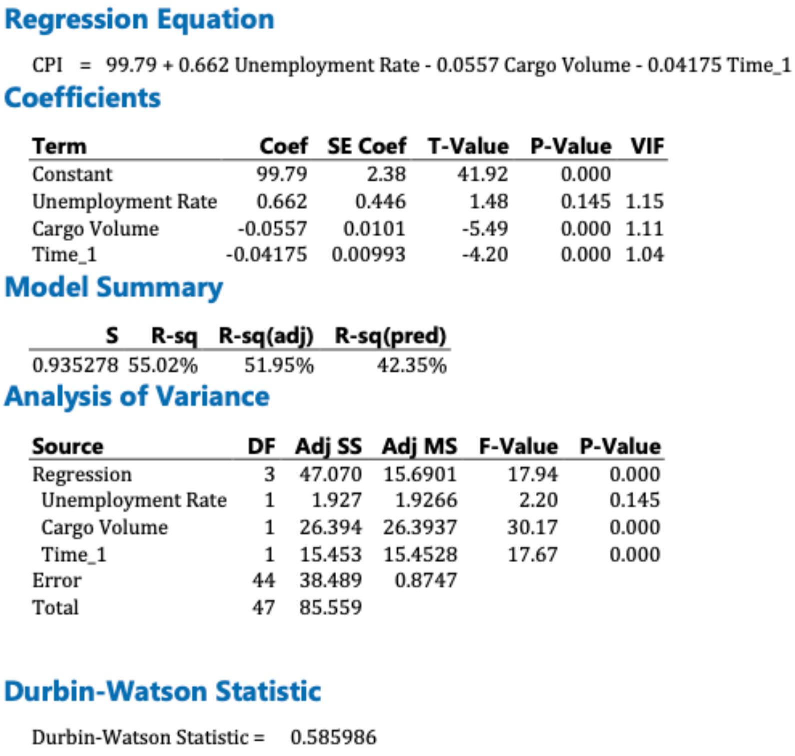
Because the residuals are not well normally distributed, the Durbin-Watson test is not valid here. If it were valid, it would find sufficient evidence to support autocorrelation because the DW statistic of 0.557007 is way smaller than QL = 1.039 at 1% significance level. The ACF plot shows that there is strong autocorrelation at Lag = 1. The runs test also shows strong evidence that there is autocorrelation, with a p-value of 0.001. Thus, it is necessary for me to address the autocorrelation of the errors. I'll first try adding time (i.e., Time_1 in the following analysis) and time-squared (i.e., Time_2 in the following analysis) as new predictors. To select relevant predicting variables more systematically, let's perform the model selection by Minitab.



Response is CPI

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No	Unemployment Rate	Cargo Volume	Wheat Market Price	Time_1	Time_2
1	36.4	35.0	63.0	26.4	18.2	1.0878	148.796	153.864	1.000		X			
1	12.8	10.9	80.1	6.4	41.3	1.2736	163.939	169.007	1.000				X	
2	52.8	50.7	51.3	40.1	4.2	0.94769	136.893	143.447	1.165		X		X	
2	51.1	48.9	52.6	38.5	5.8	0.96428	138.558	145.113	1.163		X	X		
3	55.0	51.9	49.3	42.4	4.0	0.93528	137.047	144.974	2.130	X	X		X	
3	54.4	51.3	49.1	42.6	4.6	0.94135	137.668	145.596	2.354	X	X	X		
4	55.6	51.4	51.1	40.3	5.5	0.94039	139.087	148.265	77.265	X	X		X	X
4	55.4	51.3	50.6	40.8	5.6	0.94188	139.239	148.417	30.802	X	X	X	X	
5	57.1	51.9	51.5	39.8	6.0	0.93530	140.188	150.486	125.814	X	X	X	X	X

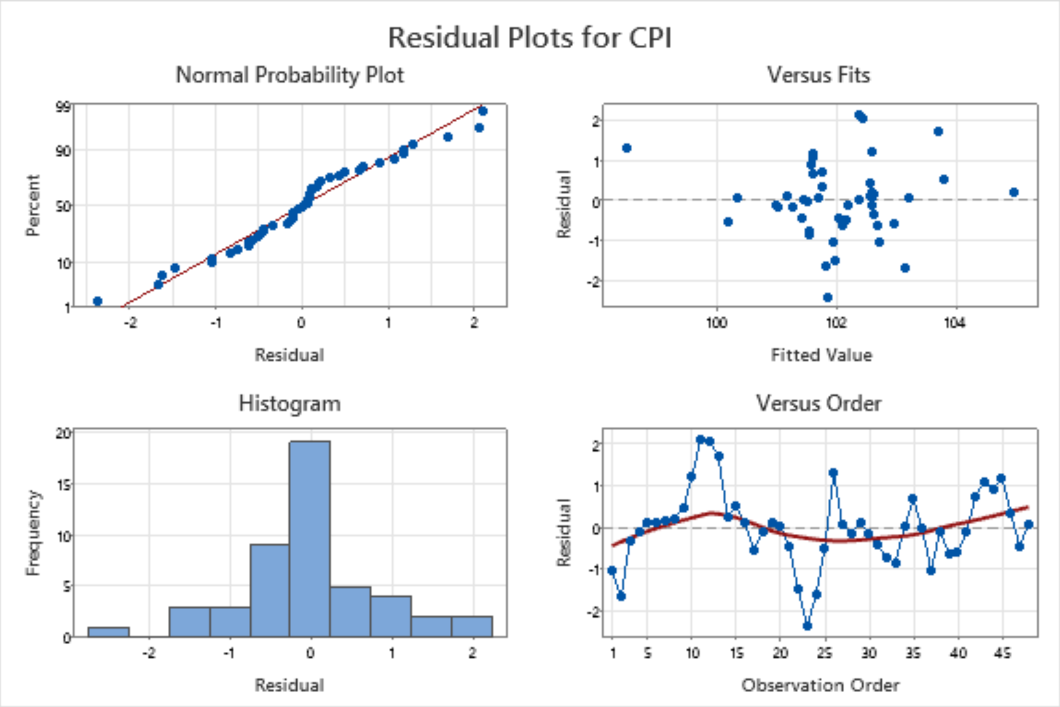
Based on the above model selection output by Minitab and the selection criteria of high R_a^2, small Mallows C_p, high R_{pred}^2, and small AIC_c, the highlighted row is the regression model with three predicting variables that has the best performance among all. It takes the unemployment rate, year-over-year growth of cargo volume, and time as the predicting variables. And the below is the regression output.



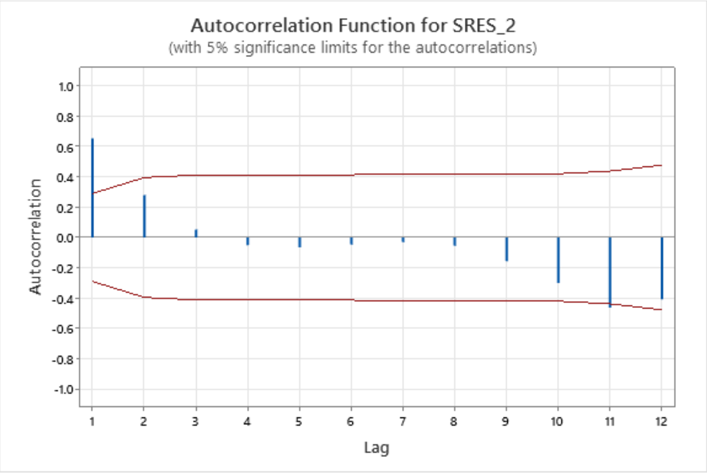
```
Regression Equation

  CPI  =  99.79 + 0.662 Unemployment Rate - 0.0557 Cargo Volume - 0.04175 Time_1

Coefficients

  Term                    Coef   SE Coef   T-Value   P-Value   VIF
  Constant               99.79      2.38     41.92     0.000
  Unemployment Rate      0.662     0.446      1.48     0.145  1.15
  Cargo Volume         -0.0557    0.0101     -5.49     0.000  1.11
  Time_1              -0.04175   0.00993     -4.20     0.000  1.04

Model Summary

         S     R-sq   R-sq(adj)   R-sq(pred)
  0.935278   55.02%      51.95%       42.35%

Analysis of Variance

  Source               DF   Adj SS    Adj MS   F-Value   P-Value
  Regression            3   47.070   15.6901     17.94     0.000
    Unemployment Rate   1    1.927    1.9266      2.20     0.145
    Cargo Volume        1   26.394   26.3937     30.17     0.000
    Time_1              1   15.453   15.4528     17.67     0.000
  Error                44   38.489    0.8747
  Total                47   85.559


Durbin-Watson Statistic

  Durbin-Watson Statistic =    0.585986
```

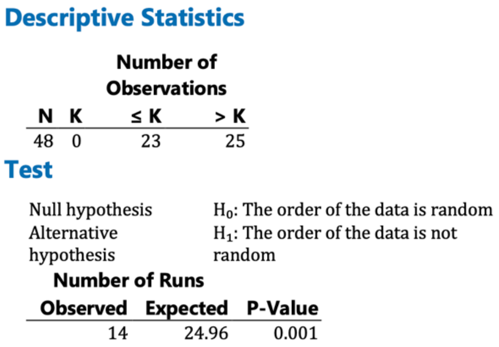
With the R-squared value is 55.02% and adjusted R-squared value is 51.95%, this regression model of CPI versus the unemployment rate, cargo volume, and time seems to work slightly better than the regression model of CPI versus unemployment rate, cargo volume, and wheat market price. However, the unemployment rate coefficient is not statistically significant with a p-value of 0.145, which is greater than any reasonable significance level. Besides, multicollinearity is not a problem in this regression.

The "four in one" plot of the new regression is shown below.



This "four in one" plot is almost identical to the previous one. The normal probability plot and the histogram still show non-normality. The non-constant variance remains, and there appears to be autocorrelation in the plot of standardized residual versus order. As a result, the Durbin-Watson test is not reliable. Let's look at some more tests. The ACF plot and the run test are as follows.

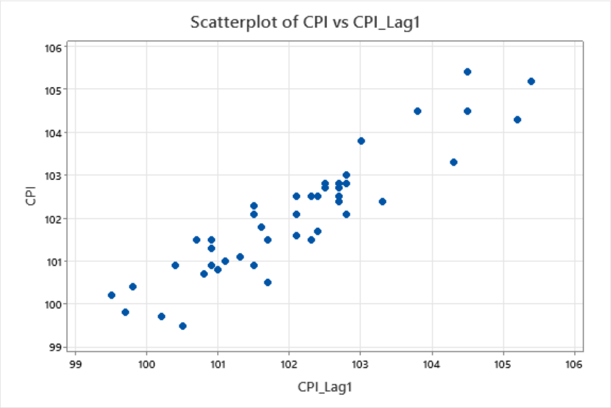


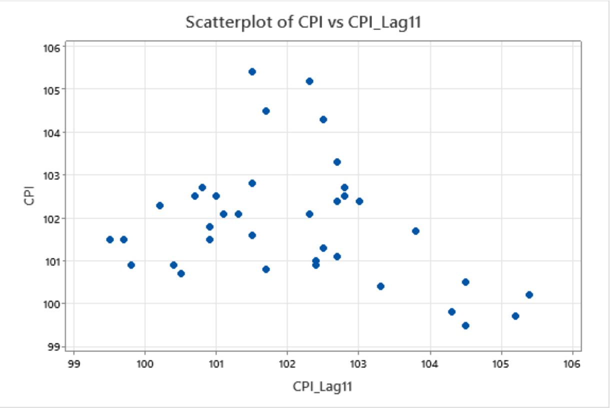


```
Descriptive Statistics

          Number of
         Observations
 N   K     ≤ K     > K
48   0      23      25

Test

Null hypothesis         H₀: The order of the data is random
Alternative             H₁: The order of the data is not
hypothesis              random

     Number of Runs
 Observed  Expected  P-Value
       14     24.96    0.001
```

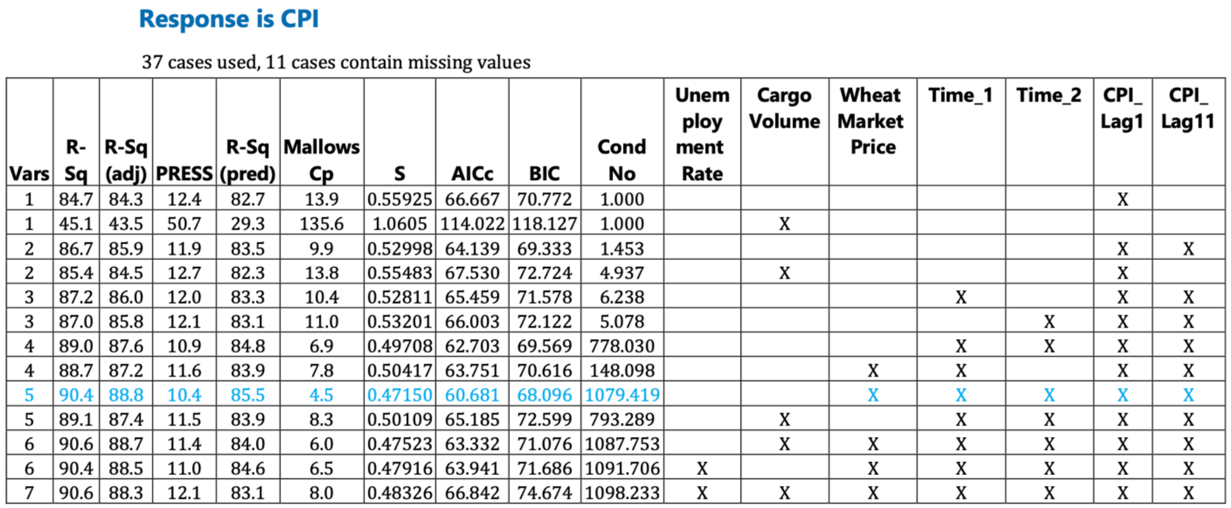
In the ACF plot, there is some reduction in autocorrelation when Lag = 1, but when Lag = 11, this model shows autocorrelation. Furthermore, the runs test shows a p-value of 0.001, which implies that there is sufficient evidence to reject the null hypothesis of no autocorrelation.

To proceed with the autocorrelation problem, I'm going to lag the target variable by 1 and 11, and use the CPI of the previous month (i.e., CPI_Lag1 in the following analysis) and CPI of the eleventh month before (i.e., CPI_Lag11 in the following analysis) as new predicting variables. The below is the scatterplot of CPI versus CPI_Lag1 and CPI_Lag11.





Based on the plots, it appears that there is a positive linear relationship between CPI and CPI_Lag1, while there is a negative linear relationship between CPI and CPI_Lag11. To determine the most appropriate variables for the model, I will use Minitab to perform the model selection.

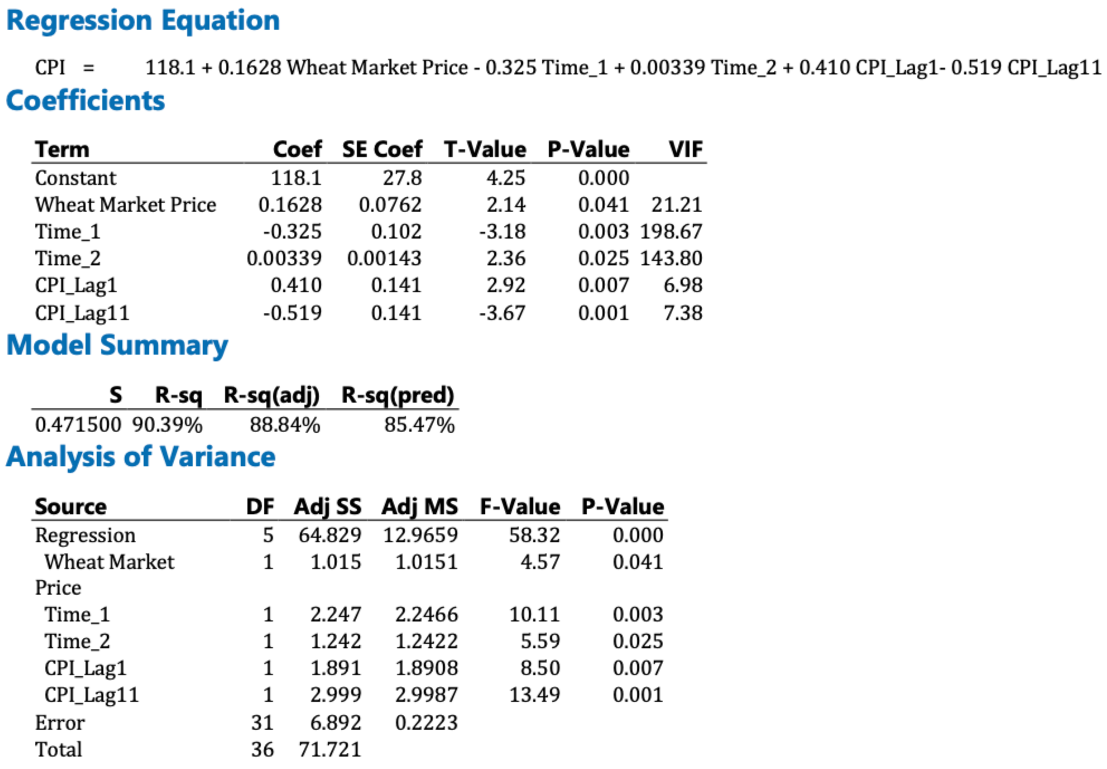


Response is CPI

37 cases used, 11 cases contain missing values

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No	Unemployment Rate	Cargo Volume	Wheat Market Price	Time_1	Time_2	CPI_Lag1	CPI_Lag11
1	84.7	84.3	12.4	82.7	13.9	0.55925	66.667	70.772	1.000						X	
1	45.1	43.5	50.7	29.3	135.6	1.0605	114.022	118.127	1.000		X					
2	86.7	85.9	11.9	83.5	9.9	0.52998	64.139	69.333	1.453						X	X
2	85.4	84.5	12.7	82.3	13.8	0.55483	67.530	72.724	4.937		X				X	
3	87.2	86.0	12.0	83.3	10.4	0.52811	65.459	71.578	6.238				X		X	X
3	87.0	85.8	12.1	83.1	11.0	0.53201	66.003	72.122	5.078					X	X	X
4	89.0	87.6	10.9	84.8	6.9	0.49708	62.703	69.569	778.030				X	X	X	X
4	88.7	87.2	11.6	83.9	7.8	0.50417	63.751	70.616	148.098			X	X		X	X
5	90.4	88.8	10.4	85.5	4.5	0.47150	60.681	68.096	1079.419			X	X	X	X	X
5	89.1	87.4	11.5	83.9	8.3	0.50109	65.185	72.599	793.289		X		X	X	X	X
6	90.6	88.7	11.4	84.0	6.0	0.47523	63.332	71.076	1087.753		X	X	X	X	X	X
6	90.4	88.5	11.0	84.6	6.5	0.47916	63.941	71.686	1091.706	X		X	X	X	X	X
7	90.6	88.3	12.1	83.1	8.0	0.48326	66.842	74.674	1098.233	X	X	X	X	X	X	X

Considering the selection criteria of high R_a^2, small Mallows C_p, high R_{pred}^2, and small AIC_c, the highlighted row represents the regression model with the five predicting variables that exhibit the best performance out of all the models. It takes the year-over-year increase in wheat market price, time, time-squared, CPI_Lag1, and CPI_Lag11 as predicting variables. The output of the regression is as follows.



Regression Equation

CPI = 118.1 + 0.1628 Wheat Market Price - 0.325 Time_1 + 0.00339 Time_2 + 0.410 CPI_Lag1- 0.519 CPI_Lag11

Coefficients

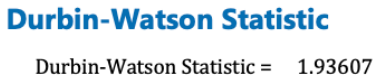
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	118.1	27.8	4.25	0.000	
Wheat Market Price	0.1628	0.0762	2.14	0.041	21.21
Time_1	-0.325	0.102	-3.18	0.003	198.67
Time_2	0.00339	0.00143	2.36	0.025	143.80
CPI_Lag1	0.410	0.141	2.92	0.007	6.98
CPI_Lag11	-0.519	0.141	-3.67	0.001	7.38

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.471500	90.39%	88.84%	85.47%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	64.829	12.9659	58.32	0.000
Wheat Market Price	1	1.015	1.0151	4.57	0.041
Time_1	1	2.247	2.2466	10.11	0.003
Time_2	1	1.242	1.2422	5.59	0.025
CPI_Lag1	1	1.891	1.8908	8.50	0.007
CPI_Lag11	1	2.999	2.9987	13.49	0.001
Error	31	6.892	0.2223		
Total	36	71.721			

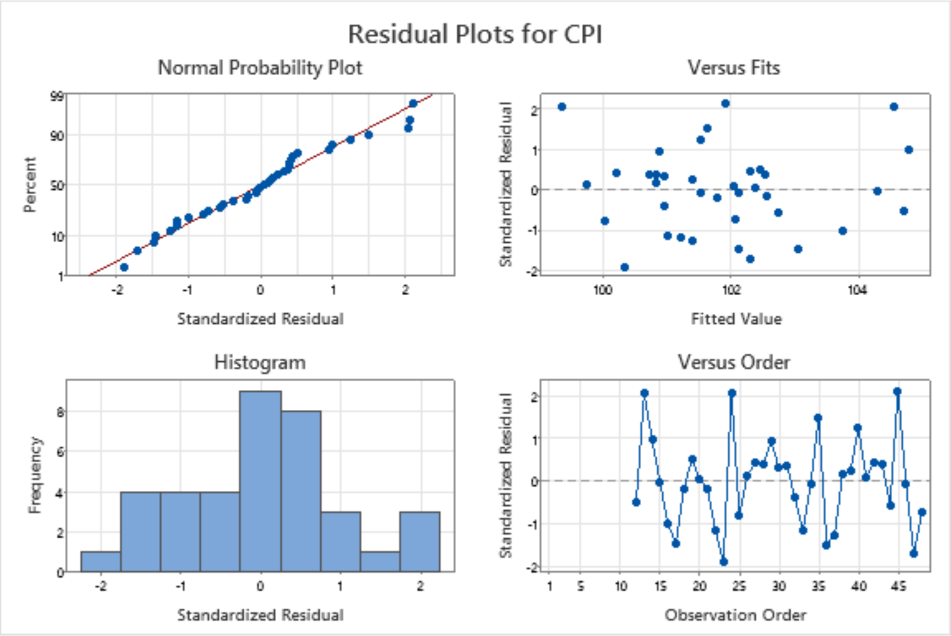


Durbin-Watson Statistic

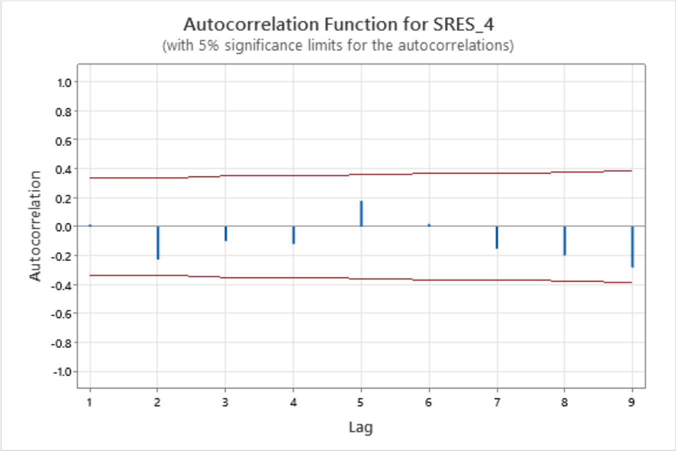
Durbin-Watson Statistic = 1.93607

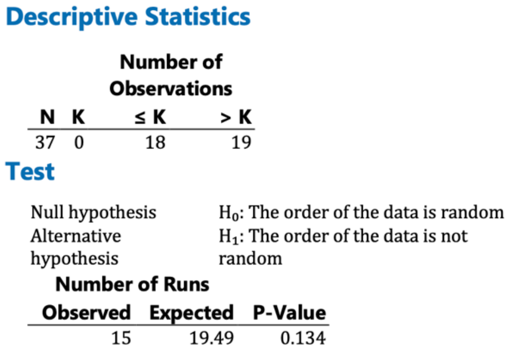
This regression is strong because the R-squared value is 90.39% and the adjusted R-squared value is 88.84%, which is greatly higher than that of the detrending regression model of CPI versus the unemployment rate, cargo volume, and time. This means that wheat market price, time, time-squared, CPI_Lag1, and CPI_Lag11 are able to account for the observed variability in CPI quite well. The F-statistic and t-statistics are all highly statistically significant. Holding all other variables fixed, one percentage point increase in the CPI of the previous month (i.e., CPI_Lag1) is associated with an estimated expected increase of 0.410 percentage point in the current month's CPI, while holding all other variables fixed, one percentage point increase in the CPI of the eleventh month before (i.e., CPI_Lag11) is associated with an estimated expected decrease of 0.519 percentage point in the current month's CPI. Despite the presence of the collinearity problem, the model selection and the t-statistics indicate that all of the predicting variables are statistically significant and therefore needed by the model.

The following is the "four in one" plot for this regression.



As can be seen from the normal probability plot and the histogram, there are still some points distributed outside of the straight line, showing non-normality. The problem of constant variance is largely alleviated and there no longer appears to be an autocorrelation problem in the plot of standardized residuals versus order. The ACF plot and run test are shown below.





Descriptive Statistics

		Number of Observations	
N	K	≤ K	> K
37	0	18	19

Test

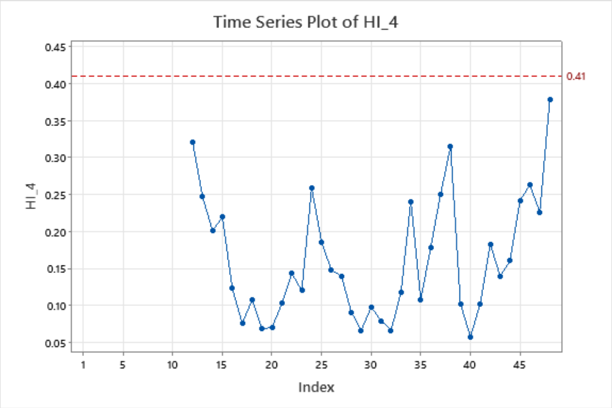
Null hypothesis	H_0: The order of the data is random
Alternative hypothesis	H_1: The order of the data is not random

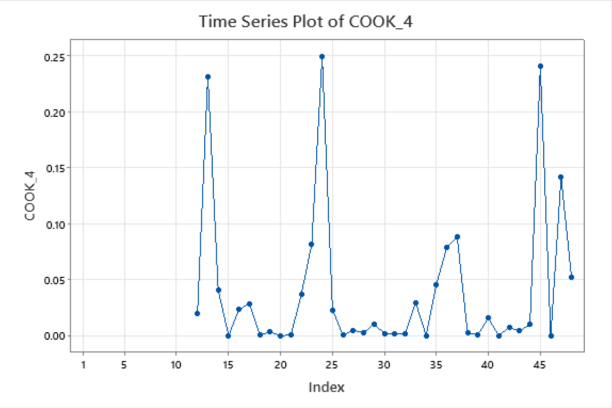
Number of Runs		
Observed	Expected	P-Value
15	19.49	0.134

It is clear that the autocorrelation problem has been resolved in the ACF plot, and the p-value of 0.134 in the run test is higher than any reasonable significance level.

Now, let's look at the unusual observations. The below is the Time Series Plot of HI and Time Series Plot of Cook's Distance. The reference line of the HI plot is

$$2.5 * \frac{p+1}{n} = 2.5 * \frac{5+1}{37} = 0.41.$$





As I have discussed, the plot of standardized residual versus fitted value looks pretty good. Also, there are no extreme leverage values above the reference value of 0.41 in the HI plot, and the Cook's distance of these points is below the suggested value of 1, indicating the absence of extreme unusual observations.

In conclusion, in this report, the optimal model for analyzing this time series data includes the predicting variables of the year-over-year increase in wheat market price, time, time-squared, CPI from the previous month, and CPI of the eleventh month before. The issue of autocorrelation is effectively addressed by using lags. The below is my current best regression model:

$$CPI = 118.1 + 0.1628 * Wheat\ Market\ Price - 0.325 * Time\_1 + 0.00339 * Time\_2 + 0.410 * CPI\_Lag1 - 0.519 * CPI\_Lag11.$$

It should be emphasized that the forecast of CPI is characterized by considerable uncertainty and is affected by a multitude of factors beyond the scope of the current model. Although the model I have presented has shown satisfactory performance in predicting CPI based on the selected predicting variables, there may be scenarios in which its accuracy could be compromised. Therefore, it is crucial to acknowledge the limitations of the current model and explore additional predicting variables that can capture the complexity of CPI dynamics in a more comprehensive way. Additionally, using more advanced and sophisticated prediction models can enhance the accuracy of CPI forecasts and help to mitigate the potential impact of unforeseen events on the inflation rate. Further research is needed to identify and integrate these additional predicting variables, and evaluate their efficacy in improving the performance of the CPI prediction model.