Regression & Multivariate Analysis
Professor Simonoff
11/17/2022
Jessy Huang

Unemployment Rate Regression Model

The unemployment rate measures the share of workers in the labor force who do not currently have a job but are actively looking for work. It is considered one of the most important economic indicators. The COVID-19 pandemic triggered the sharpest downturn in the US economy since the Depression and led to huge increase in unemployment rate. A better understanding of unemployment trends over the past few decades and during the course of Covid-19, and the relation with the factors correlated to the rate could help to predict the future fluctuations in the unemployment rate.

I would like to explore four potential areas which could have relationships with unemployment rate: GDP, interest rate, inflation rate and exchange rate, which are represented in four predictors. The analysis present here is based on the data from the United States from years 2005 and 2021.

The target variable here is "Unemployment Rate", while the potential four predictors identified are "real GDP Percent Change", "Prime Interest Rate in US", "Consumer Price Index less Food and Energy(CPI_less food energy)", "Exchange Rate with Chinese Yuan Renminbi to One U.S. Dollar".

US Prime Interest rate was chosen because it is a core indicator of business investment, and Consumer Price Index less Food and Energy was chosen because it is a measure of the average change over time in the prices paid by urban consumers, which is widely used as an inflation indicator. Exchange rate has direct effects on the real economy through changes in the demand for exports and imports. Because China is the biggest business partner in the world, and therefore likely to be the most impactful, the exchange rate between Chinese yuan and USD was chosen. Real GDP percent change is a measure of GDP growth that is inflation adjusted, making it more accurate than normal GDP percent change. These factors are all believed to be heavily correlated with economic wellbeing, therefore I believe that these four indicators might be good candidates for unemployment rate regression.

Target Variable	Unemployment Rate, %, Quarterly
Predictor Variable	1. Real GDP Percent Change of USA, %, Quarterly
	2. Prime Interest Rate, %, Quarterly, US prime rate.

	3. Consumer Price Index less food and Energy(CPI_less food energy), Quarterly, Ratio of price of all items of consumer goods and services less food and energy in current period to that in basic period in USA.
	4. Exchange Rate with Chinese Yuan Renminbi to One U.S. Dollar, Quarterly

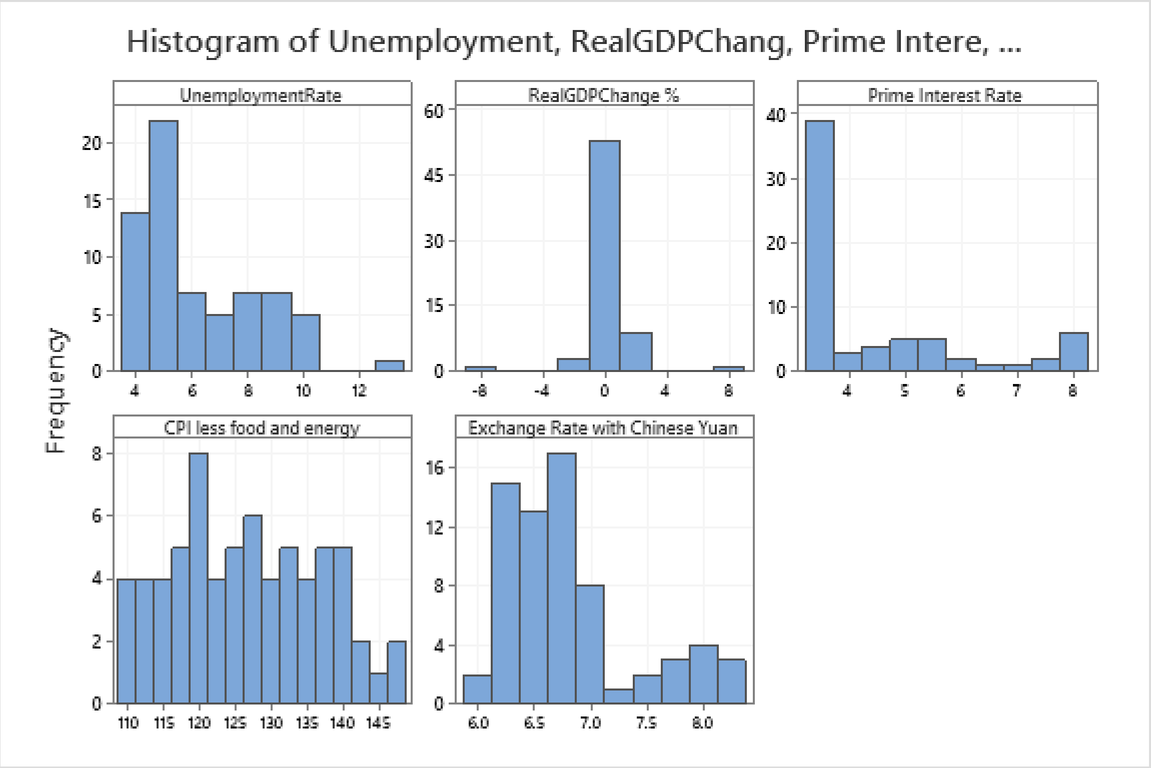
The data I analyzed come from the following sources. They are quarterly data from the United States from 2005 to 2021, totaling up to 68 quarterly time series points for the target and its predictors.

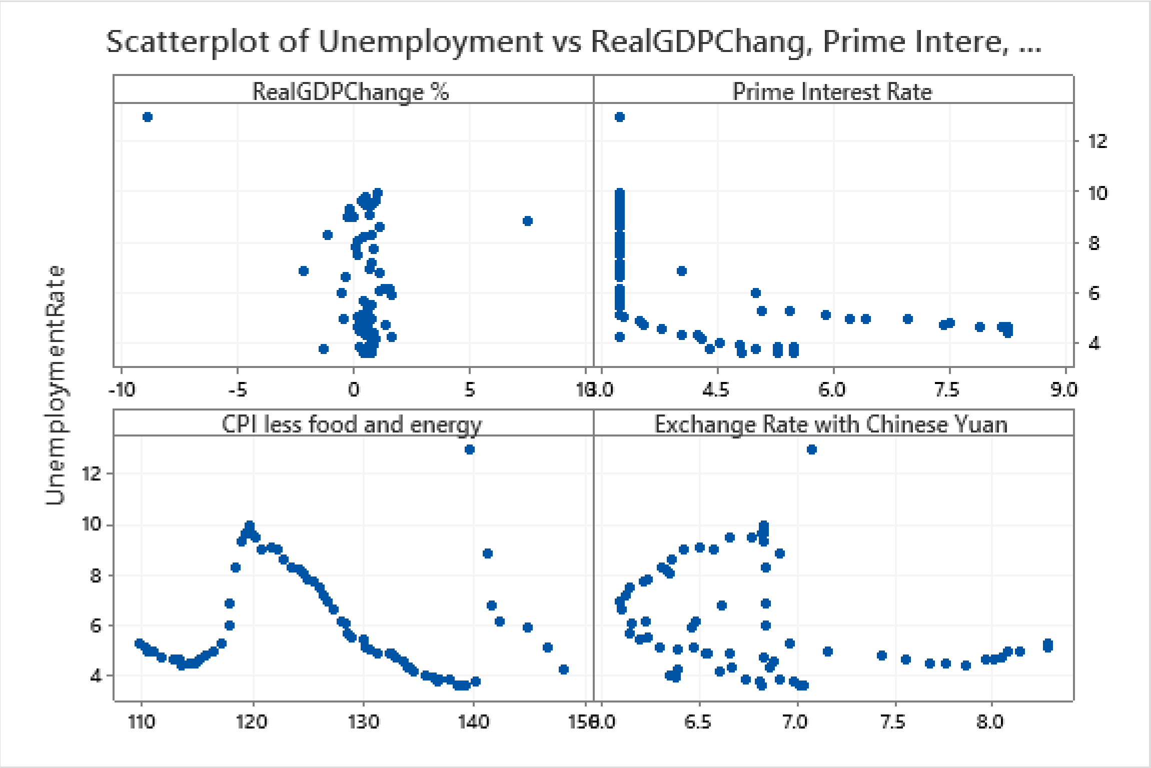
- The unemployment rate and Consumer Price Index data comes from US Bureau of Labor Statistics https://www.bls.gov/lau/lastch20.htm
- GDP data is from https://www.bea.gov/, website of "Bureau of Economic Analysis, US Department of Commerce".
- Exchange Rate data is from https://fred.stlouisfed.org/series/EXCHUS, website of "FRED ECONOMIC DATA ST. LOUIS FED"
- Prime Interest Rate data is from http://www.fedprimerate.com/wall_street_journal_prime_rate_history.htm, website of "FedPrimeRate.com"

A snapshot of the data shows:

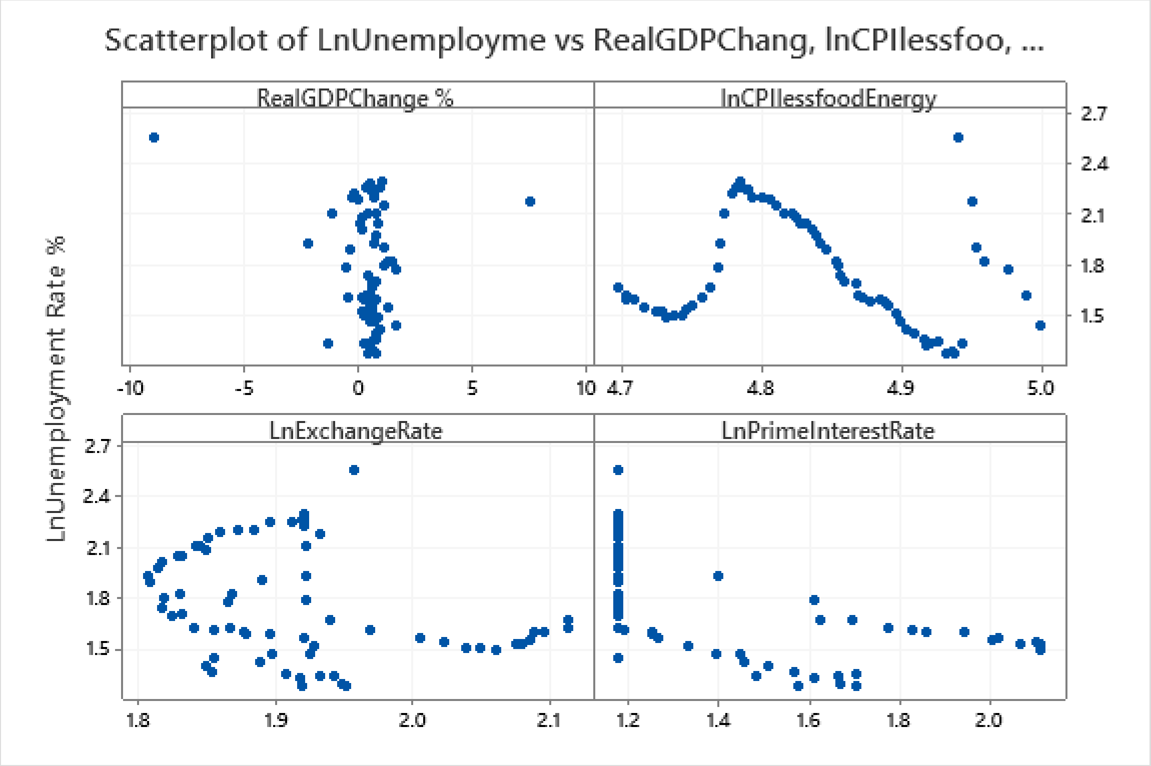
Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3	Maximum
UnemploymentRate	68	0	6.179	0.256	2.115	3.600	4.585	5.315	7.972	12.970
RealGDPChange %	67	1	0.452	0.193	1.580	-8.937	0.246	0.572	0.813	7.548
Prime Interest Rate	68	0	4.390	0.197	1.628	3.250	3.250	3.270	5.230	8.250
Exchange Rate with Chinese Yuan	68	0	6.8206	0.0717	0.5915	6.0900	6.3650	6.7550	6.9750	8.2800
CPI less food and energy	68	0	126.64	1.22	10.10	109.77	118.54	126.21	134.51	148.15

For a quick impression of the factors relationship I plotted the histogram of these variables, and the scatterplots of the response versus each predictor:


Histogram of Unemployment, RealGDPChang, Prime Intere, ...
Frequency
UnemploymentRate
0
5
10
15
20
4
6
8
10
12
RealGDPChange %
0
15
30
45
60
-8
-4
0
4
8
Prime Interest Rate
0
10
20
30
40
4
5
6
7
8
CPI less food and energy
0
2
4
6
8
110
115
120
125
130
135
140
145
Exchange Rate with Chinese Yuan
0
4
8
12
16
6.0
6.5
7.0
7.5
8.0

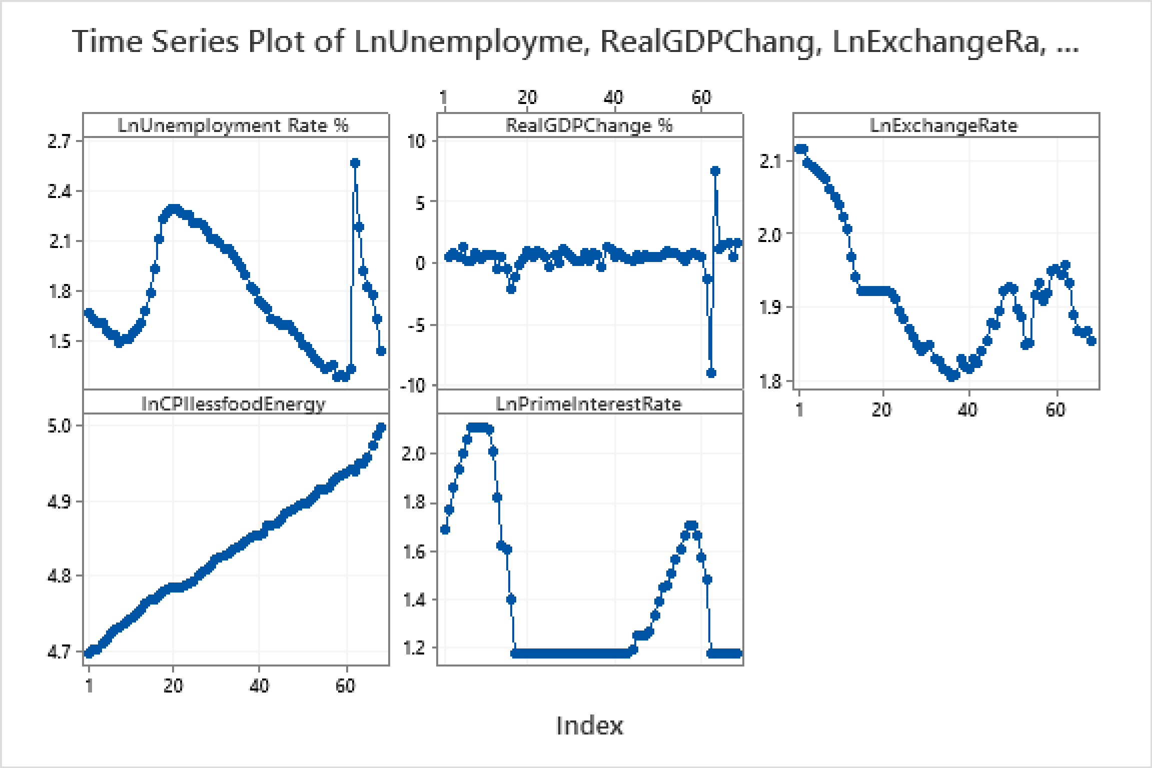

Scatterplot of Unemployment vs RealGDPChang, Prime Intere, ...
UnemploymentRate
RealGDPChange %
-10
-5
0
5
10
Prime Interest Rate
3.0
4.5
6.0
7.5
9.0
4
6
8
10
12
CPI less food and energy
110
120
130
140
150
Exchange Rate with Chinese Yuan
6.0
6.5
7.0
7.5
8.0
4
6
8
10
12

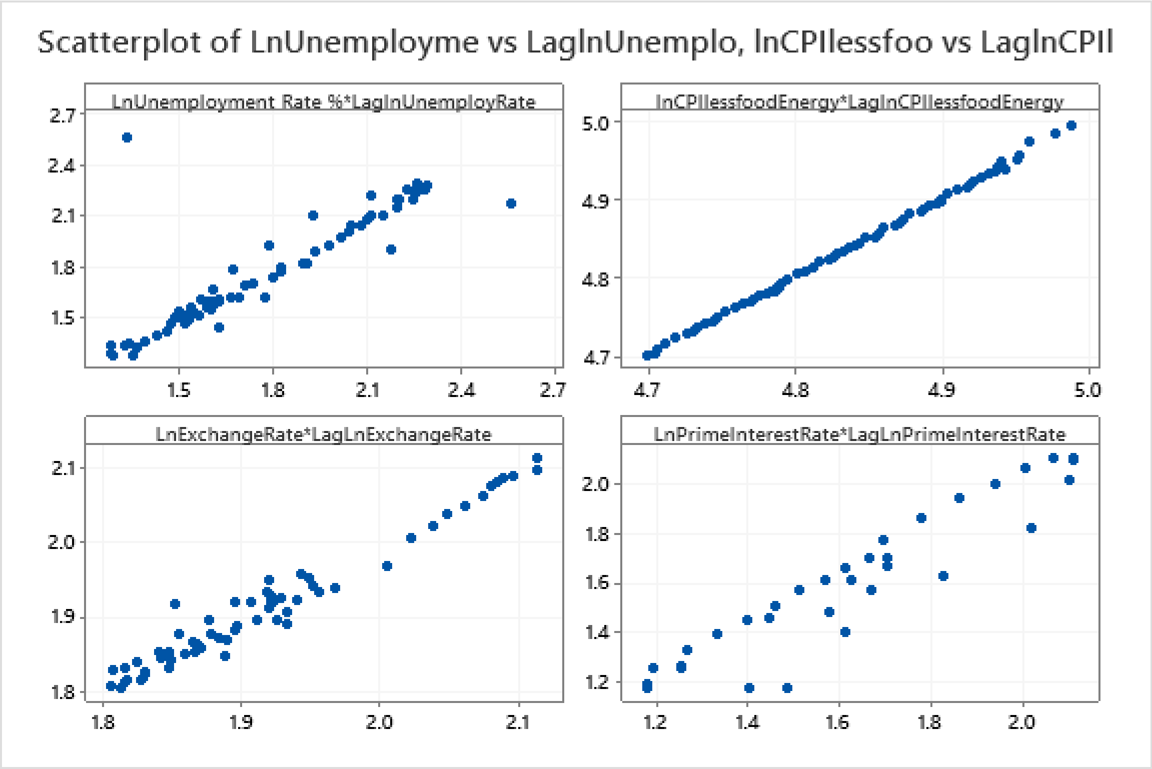
We can see that the histogram of the target variable “Unemployment Rate” and two predictors “Prime Interest Rate”, “Exchange Rate with Chinese Yuan” seems to have a long right-tailed distribution. In addition, the predictor “CPI less Food and Energy” has larger values (in the 100 to 200 range) than all other predictors, which can be seen in the scatterplot above. For these reasons, I decide to use logarithms in the modeling for these four variables. Here are the scatter plots with above four variables taking natural log:



By looking at the above target versus predictor scatterplot, we can see that there sems to be some sort of relationship between the target and predictors although it is not like exactly what we would expect to see, as there appears to be evidence of nonconstant variance. There also seems to be some unusual observations.

I also plotted time series plot for target and the four predictors. All of these variables except “Real GDP Percent Change” seem to exhibit strong time ordering. So I generated lagged version variable for each individual variable, and plotted a scatterplot of variable vs it’s lagged version variable for each of them. We can see that all of these four variable show strong relations with their previous quarter data. Let's ignore this for the moment. I’ll come back later to discuss more about this autocorrelation handling.


Time Series Plot of LnUnemployme, RealGDPChang, LnExchangeRa, ...
LnUnemployment Rate %
RealGDPChange %
LnExchangeRate
lnCPIlessfoodEnergy
LnPrimeInterestRate
2.7
2.4
2.1
1.8
1.5
10
5
0
-5
-10
2.1
2.0
1.9
1.8
5.0
4.9
4.8
4.7
2.0
1.8
1.6
1.4
1.2
1
20
40
60
Index


Scatterplot of LnUnemployme vs LaglnUnemplo, lnCPIlessfoo vs LaglnCPIl
LnUnemployment Rate %*LaglnUnemployRate
lnCPIlessfoodEnergy*LaglnCPIlessfoodEnergy
LnExchangeRate*LagLnExchangeRate
LnPrimeInterestRate*LagLnPrimeInterestRate
2.7
2.4
2.1
1.8
1.5
1.5
1.8
2.1
2.4
2.7
5.0
4.9
4.8
4.7
4.7
4.8
4.9
5.0
2.1
2.0
1.9
1.8
1.8
1.9
2.0
2.1
2.0
1.8
1.6
1.4
1.2
1.2
1.4
1.6
1.8
2.0

Here is the correlation of the variables. First, we can notice that some of the variables are correlated with each other. A regression using all of the variables might exhibit multicollinearity and will presumably include redundant variables.

Correlations

	LnUnemployment Rate %	RealGDPChange %	LnExchangeRate	lnCPIlessfoodEnergy
RealGDPChange %	-0.167			
LnExchangeRate	-0.301	-0.064		
lnCPIlessfoodEnergy	-0.228	0.077	-0.577	
LnPrimeInterestRate	-0.609	0.000	0.825	-0.465

Here is the regression output. Just as suspected, there is one redundant variable, which is indicated by the insignificant t–statistics. The VIF values look ok.

Regression Equation

LnUnemployment Rate % = 12.45 - 0.0210 RealGDPChange % + 1.348 LnExchangeRate - 2.396 lnCPIlessfoodEnergy - 1.163 LnPrimeInterestRate

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	12.45	2.00	(8.46, 16.44)	6.23	0.000	
RealGDPChange %	-0.0210	0.0131	(-0.0471, 0.0052)	-1.60	0.115	1.02
LnExchangeRate	1.348	0.497	(0.354, 2.341)	2.71	0.009	3.79
lnCPIlessfoodEnergy	-2.396	0.315	(-3.025, -1.767)	-7.62	0.000	1.44
LnPrimeInterestRate	-1.163	0.116	(-1.395, -0.930)	-9.99	0.000	3.34

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.166806	75.44%	73.85%	4.36658	37.83%	-41.64	-29.81

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	4	5.2987	75.44%	5.29874	1.32469	47.61	0.000
RealGDPChange %	1	0.1968	2.80%	0.07120	0.07120	2.56	0.115
LnExchangeRate	1	0.6966	9.92%	0.20458	0.20458	7.35	0.009
lnCPIlessfoodEnergy	1	1.6304	23.21%	1.61455	1.61455	58.03	0.000
LnPrimeInterestRate	1	2.7749	39.51%	2.77493	2.77493	99.73	0.000
Error	62	1.7251	24.56%	1.72510	0.02782		
Total	66	7.0238	100.00%				

Fits and Diagnostics for Unusual Observations

Obs	LnUnemployment Rate %	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid
2	1.6292	1.9514	0.0670	(1.8176, 2.0853)	-0.3222	-2.11	-2.17

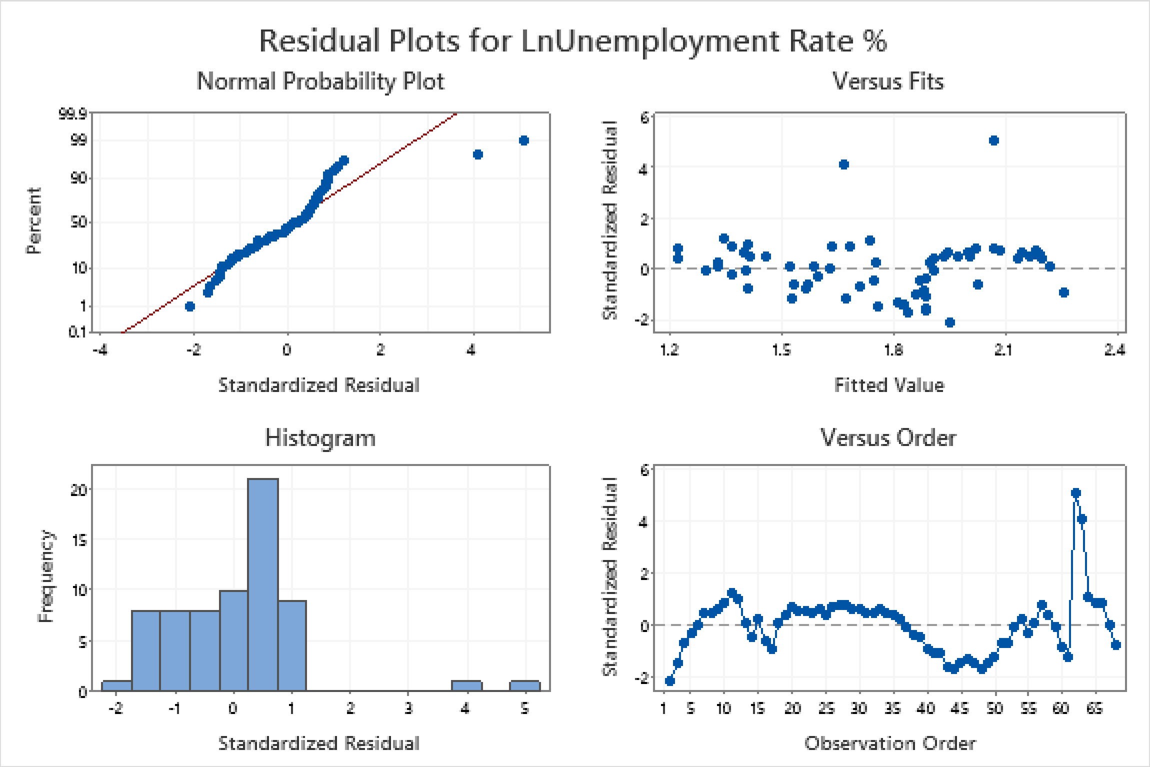
62	2.5626	2.0667	0.1357	(1.7955, 2.3380)	0.4959	5.11	6.66
63	2.1782	1.6632	0.1102	(1.4429, 1.8835)	0.5149	4.11	4.78

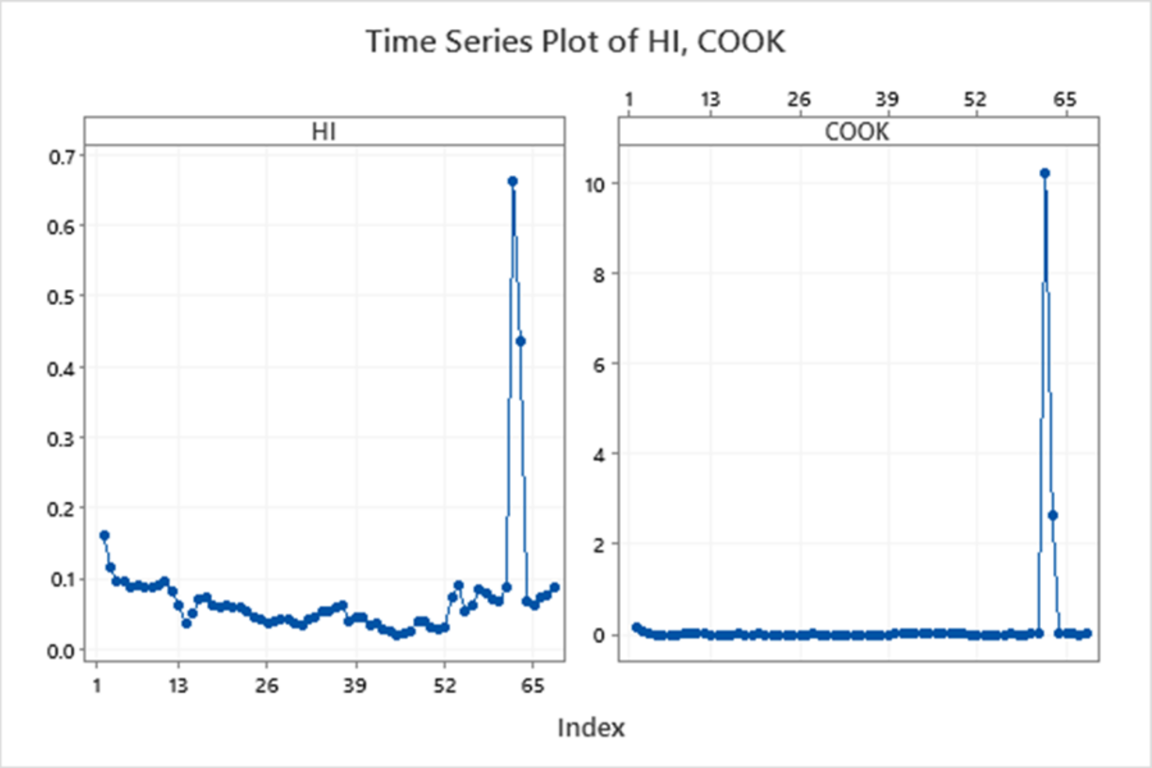
Obs	HI	Cook's D	DFITS		
2	0.161207	0.17	-0.95187	R	
62	0.661587	10.21	9.31631	R	X
63	0.436621	2.62	4.21173	R	X

R Large residual
X Unusual X

Durbin-Watson Statistic

Durbin-Watson Statistic =	0.499702





The diagnosis result doesn’t look good. Apparently Point# 62, #63 are flagged as unusual points, since they show large values on COOK’s distance, Leverage HI and Residual. In context, this isn’t surprising as the points #62 and #63 represent the second and third quarter of 2020, which was Covid-19’s peak period. I decided to omit them through the category indicator “CovidUnusual1”,“CovidUnusual2”, and reran the regression:

Regression Equation

LnUnemployment Rate % = 17.94 + 0.0330 RealGDPChange %
+ 0.129 LnExchangeRate
- 3.125 lnCPIlessfoodEnergy
- 0.9361 LnPrimeInterestRate
+ 1.206 CovidUnusual1 + 0.313 CovidUnusual2

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	17.94	1.72	(14.50, 21.38)	10.43	0.000	
RealGDPChange %	0.0330	0.0257	(-0.0183, 0.0843)	1.29	0.203	6.74
LnExchangeRate	0.129	0.419	(-0.708, 0.966)	0.31	0.759	4.65
lnCPIlessfoodEnergy	-3.125	0.264	(-3.653, -2.598)	-11.85	0.000	1.75
LnPrimeInterestRate	-0.9361	0.0950	(-1.1261, -0.7462)	-9.86	0.000	3.84
CovidUnusual1	1.206	0.285	(0.636, 1.775)	4.23	0.000	4.97
CovidUnusual2	0.313	0.221	(-0.129, 0.755)	1.42	0.162	2.98

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.126891	86.25%	84.87%	*	*	-75.41	-60.25

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	6	6.05776	86.25%	6.05776	1.00963	62.70	0.000
RealGDPChange %	1	0.19677	2.80%	0.02669	0.02669	1.66	0.203
LnExchangeRate	1	0.69662	9.92%	0.00153	0.00153	0.09	0.759
lnCPIlessfoodEnergy	1	1.63043	23.21%	2.26060	2.26060	140.40	0.000
LnPrimeInterestRate	1	2.77493	39.51%	1.56450	1.56450	97.17	0.000
CovidUnusual1	1	0.72667	10.35%	0.28833	0.28833	17.91	0.000
CovidUnusual2	1	0.03235	0.46%	0.03235	0.03235	2.01	0.162
Error	60	0.96608	13.75%	0.96608	0.01610		
Total	66	7.02384	100.00%				

Fits and Diagnostics for Unusual Observations

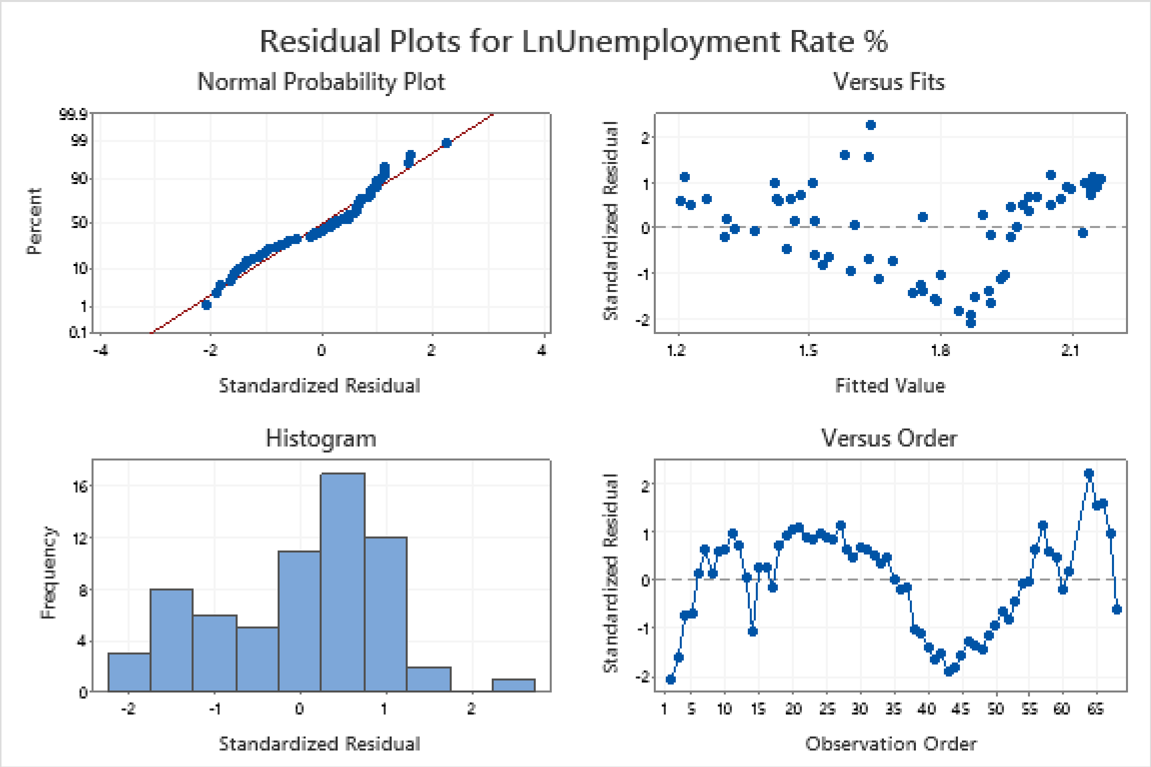
Obs	LnUnemployment Rate %	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI
2	1.6292	1.8674	0.0529	(1.7615, 1.9733)	-0.2381	-2.06	-2.12	0.17402
62	2.5626	2.5626	0.1269	(2.3088, 2.8165)	-0.0000	*	*	1.00000
63	2.1782	2.1782	0.1269	(1.9243, 2.4320)	-0.0000	*	*	1.00000
64	1.9125	1.6399	0.0378	(1.5642, 1.7156)	0.2726	2.25	2.33	0.08894

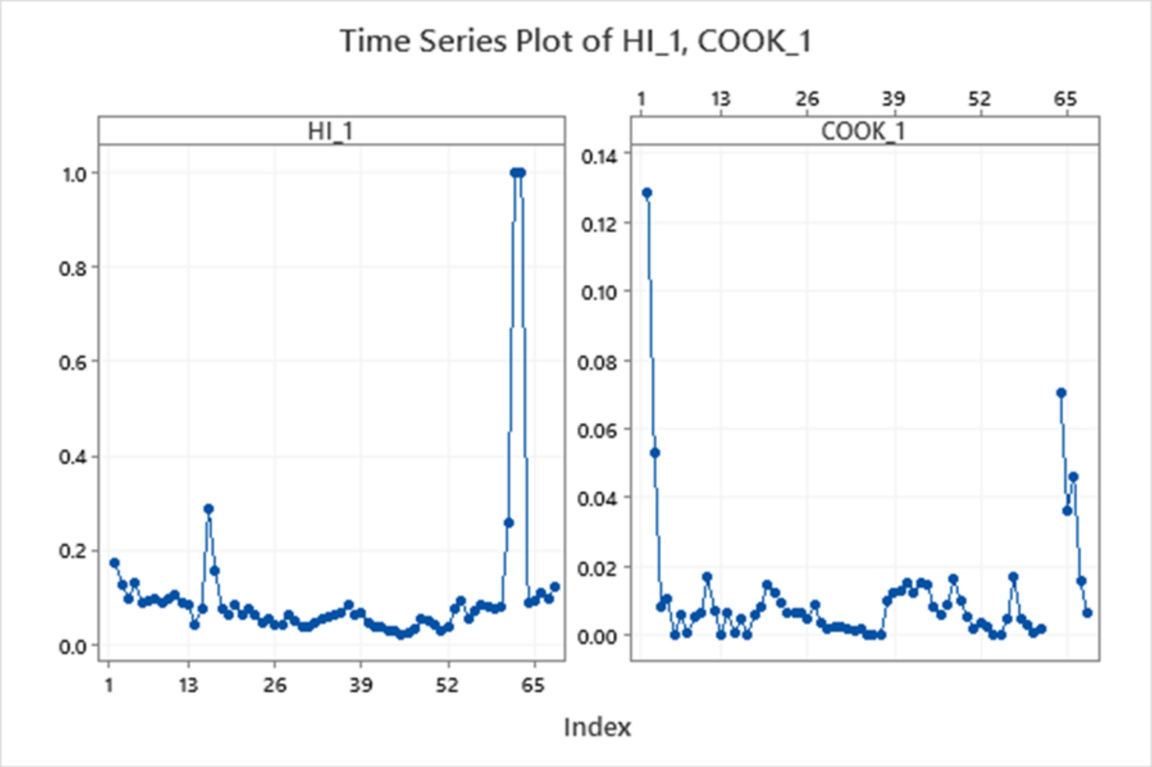
Obs	Cook's D	DFITS	
2	0.13	-0.975158	R
62	*	*	X
63	*	*	X
64	0.07	0.728706	R

R Large residual
X Unusual X

Durbin-Watson Statistic

Durbin-Watson Statistic =	0.295596


Residual Plots for LnUnemployment Rate %
Normal Probability Plot
Percent
99.9
99
90
50
10
1
0.1
-4
-2
0
2
4
Standardized Residual
Versus Fits
Standardized Residual
2
1
0
-1
-2
1.2
1.5
1.8
2.1
Fitted Value
Histogram
Frequency
16
12
8
4
0
-2
-1
0
1
2
Standardized Residual
Versus Order
Standardized Residual
2
1
0
-1
-2
1
5
10
15
20
25
30
35
40
45
50
55
60
65
Observation Order


Time Series Plot of HI_1, COOK_1
HI_1
1.0
0.8
0.6
0.4
0.2
0.0
1
13
26
39
52
65
COOK_1
1
13
26
39
52
65
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Index

The diagnosis test doesn't look good either. First of all it flags Point #61 and #16 as potential unusual points with HI values of ~0.25 , which is little above the guide line value. I examined them further by omitting them and rerunning the regression, but the result doesn't seem to change much, so I decided to keep them.

Besides we can see that several variables with the t–statistics being insignificant, which suggests they might need to be removed. But I'll ignore them for the moment since I need to resolve autocorrelation first, as it is an immediate violation to the regression assumption.

The diagnosis test show that there is apparent nonnormality of the residuals, and also nonconstant variance. The residual seems to be suffering from autocorrelation. As mentioned earlier, the four variables show strong time ordering in the time series plots.

Durbin-Watson Statistic support this as it equals to 0.296. So does the "run" test. It also represents a sign of autocorrelation.

Runs Test: SRES_1

Descriptive Statistics

		Number of Observations	
N	**K**	**≤ K**	**> K**
67	0.0034152	30	37

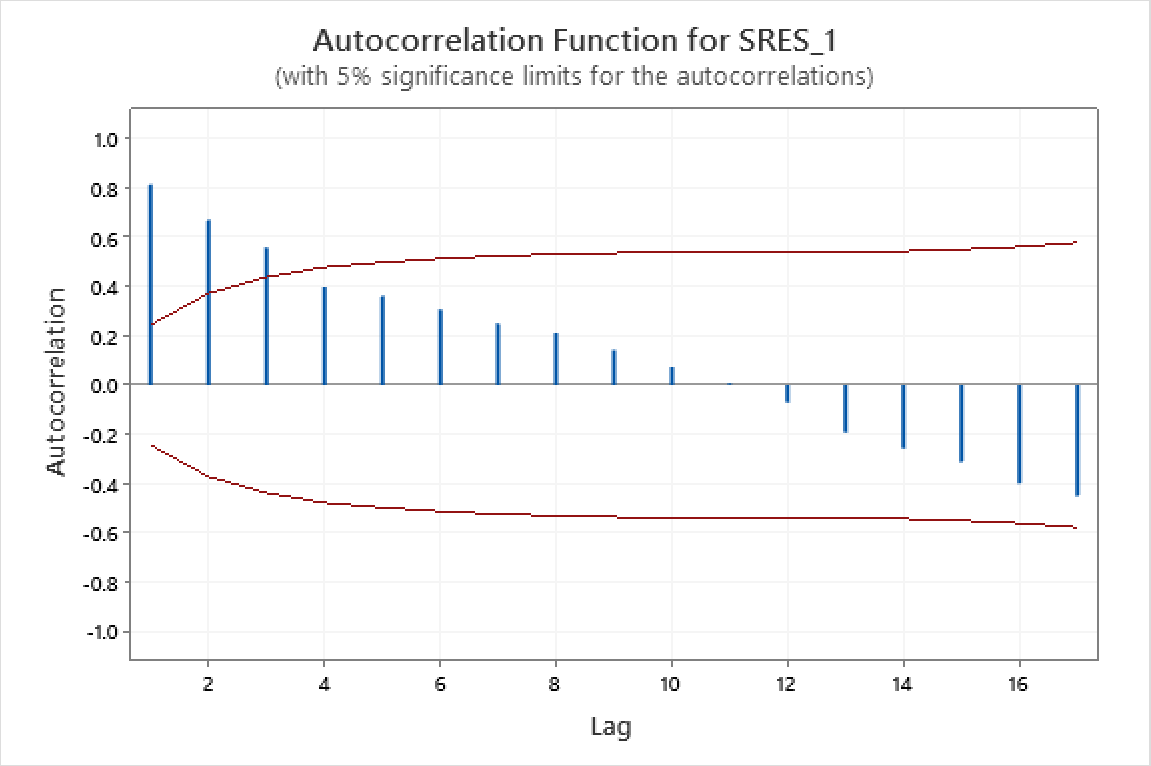
K = sample mean

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**
13	34.13	0.000

The ACF plot of the standardized residuals also indicates autocorrelation:



As learned in the class, one approach for handling autocorrelation is to use a lagged version of the target variable as a predictor (LaglnUnemployRate), saying that the previous quarter unemployment comes to predict this quarter's unemployment due to basic stability in the process. In addition, due to the nature of economics, the impact of the variables interest rate, exchange rate, inflation, and GDP will always be delayed, so it seems reasonable to consider also using a lagged version of the interest rate(LagLnPrimeInterestRate), exchange rate(LagLnExchangeRate), Consumer Price Index (LaglnCPIlessfoodEnergy), and lagged GDP change (LagRealGDPChange). We understand that using lagged versions of predictors is not designed to specifically address autocorrelation (as the use of the lagged target as a predictor often is), but rather based on such use making sense in context.

Here is the list of Lagged version of target and predictors variables which are to be included in the regression model as potential predictor. I reran the regression, here is the regression output:

- LaglnUnemployRate
- LagLnPrimeInterestRate
- LagRealGDPChange
- LaglnCPIlessfoodEnergy
- LagLnExchangeRate

Regression Equation

LnUnemployment Rate % = 1.302 - 0.03339 RealGDPChange % - 0.110 LnExchangeRate
- 0.48 lnCPIlessfoodEnergy - 0.1833 LnPrimeInterestRate
+ 0.8301 CovidUnusual1 - 0.348 CovidUnusual2
+ 0.9593 LaglnUnemployRate + 0.1510 LagLnPrimeInterestRate
- 0.02792 LagRealGDPChange + 0.21 LaglnCPIlessfoodEnergy
+ 0.173 LagLnExchangeRate

Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	1.302	0.885	(-0.473, 3.077)	1.47	0.147	
RealGDPChange %	-0.03339	0.00882	(-0.05106, -0.01571)	-3.79	0.000	10.76
LnExchangeRate	-0.110	0.310	(-0.730, 0.511)	-0.35	0.725	31.12
lnCPIlessfoodEnergy	-0.48	1.62	(-3.73, 2.77)	-0.29	0.770	850.31
LnPrimeInterestRate	-0.1833	0.0998	(-0.3834, 0.0168)	-1.84	0.072	56.27
CovidUnusual1	0.8301	0.0804	(0.6689, 0.9913)	10.32	0.000	5.35
CovidUnusual2	-0.348	0.100	(-0.549, -0.147)	-3.47	0.001	8.30
LaglnUnemployRate	0.9593	0.0403	(0.8785, 1.0401)	23.80	0.000	9.44
LagLnPrimeInterestRate	0.1510	0.0867	(-0.0228, 0.3249)	1.74	0.087	42.90
LagRealGDPChange	-0.02792	0.00567	(-0.03929, -0.01654)	-4.92	0.000	4.42
LaglnCPIlessfoodEnergy	0.21	1.66	(-3.12, 3.54)	0.13	0.900	880.34
LagLnExchangeRate	0.173	0.307	(-0.443, 0.788)	0.56	0.576	33.51

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	AICc	BIC
0.0345050	99.08%	98.90%	*	*	-237.34	-215.88

Analysis of Variance

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Regression	11	6.93965	99.08%	6.93965	0.630877	529.88	0.000
RealGDPChange %	1	0.19642	2.80%	0.01707	0.017073	14.34	0.000
LnExchangeRate	1	0.69263	9.89%	0.00015	0.000149	0.13	0.725
lnCPIlessfoodEnergy	1	1.61714	23.09%	0.00010	0.000103	0.09	0.770
LnPrimeInterestRate	1	2.89641	41.35%	0.00402	0.004016	3.37	0.072
CovidUnusual1	1	0.67931	9.70%	0.12686	0.126858	106.55	0.000
CovidUnusual2	1	0.02460	0.35%	0.01437	0.014374	12.07	0.001
LaglnUnemployRate	1	0.79396	11.34%	0.67430	0.674296	566.35	0.000
LagLnPrimeInterestRate	1	0.00811	0.12%	0.00361	0.003613	3.03	0.087
LagRealGDPChange	1	0.03069	0.44%	0.02883	0.028826	24.21	0.000
LaglnCPIlessfoodEnergy	1	0.00000	0.00%	0.00002	0.000019	0.02	0.900
LagLnExchangeRate	1	0.00038	0.01%	0.00038	0.000377	0.32	0.576
Error	54	0.06429	0.92%	0.06429	0.001191		
Total	65	7.00394	100.00%				

Fits and Diagnostics for Unusual Observations

Obs	LnUnemployment Rate %	Fit	SE Fit	95% CI	Resid	Std Resid	Del Resid	HI
15	1.7918	1.7131	0.0123	(1.6884, 1.7378)	0.0787	2.44	2.56	0.12781
62	2.5626	2.5626	0.0345	(2.4935, 2.6318)	-0.0000	*	*	1.00000

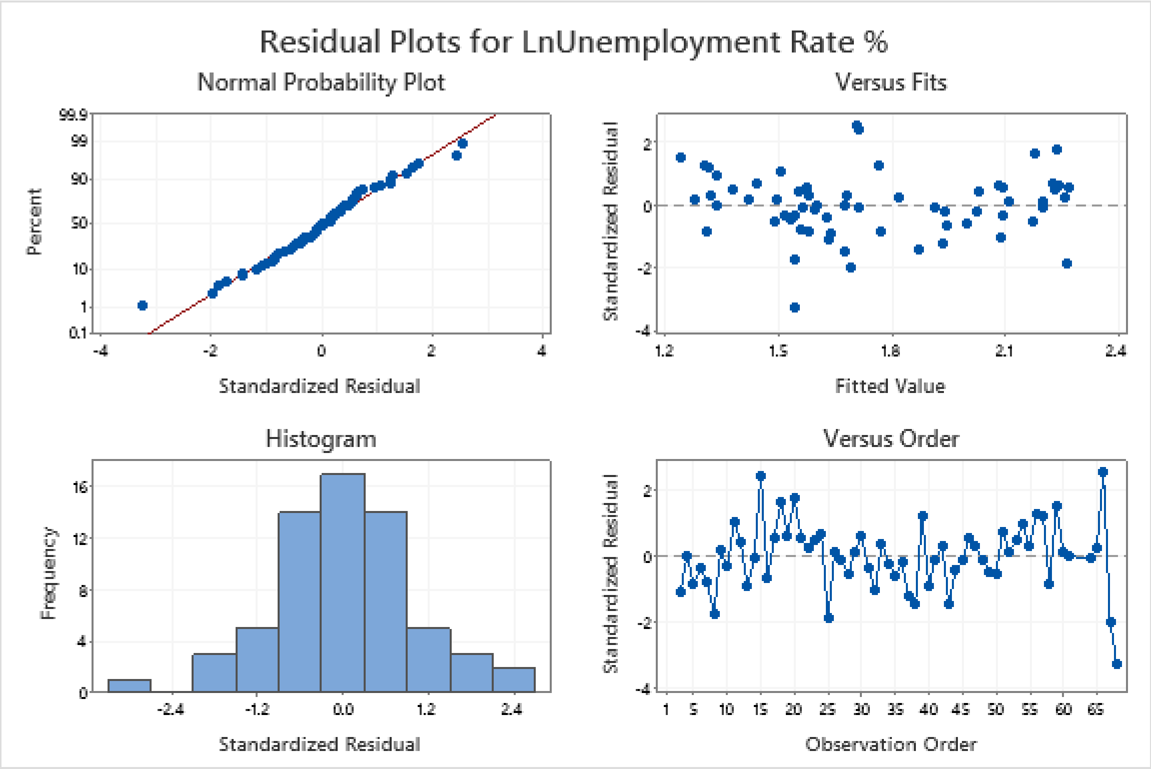
63	2.1782	2.1782	0.0345	(2.1090, 2.2473)	0.0000	*	*	1.00000
64	1.9125	1.9136	0.0304	(1.8527, 1.9745)	-0.0011	-0.07	-0.07	0.77527
66	1.7750	1.7069	0.0218	(1.6632, 1.7505)	0.0681	2.54	2.69	0.39846
68	1.4422	1.5416	0.0155	(1.5105, 1.5727)	-0.0994	-3.22	-3.56	0.20195

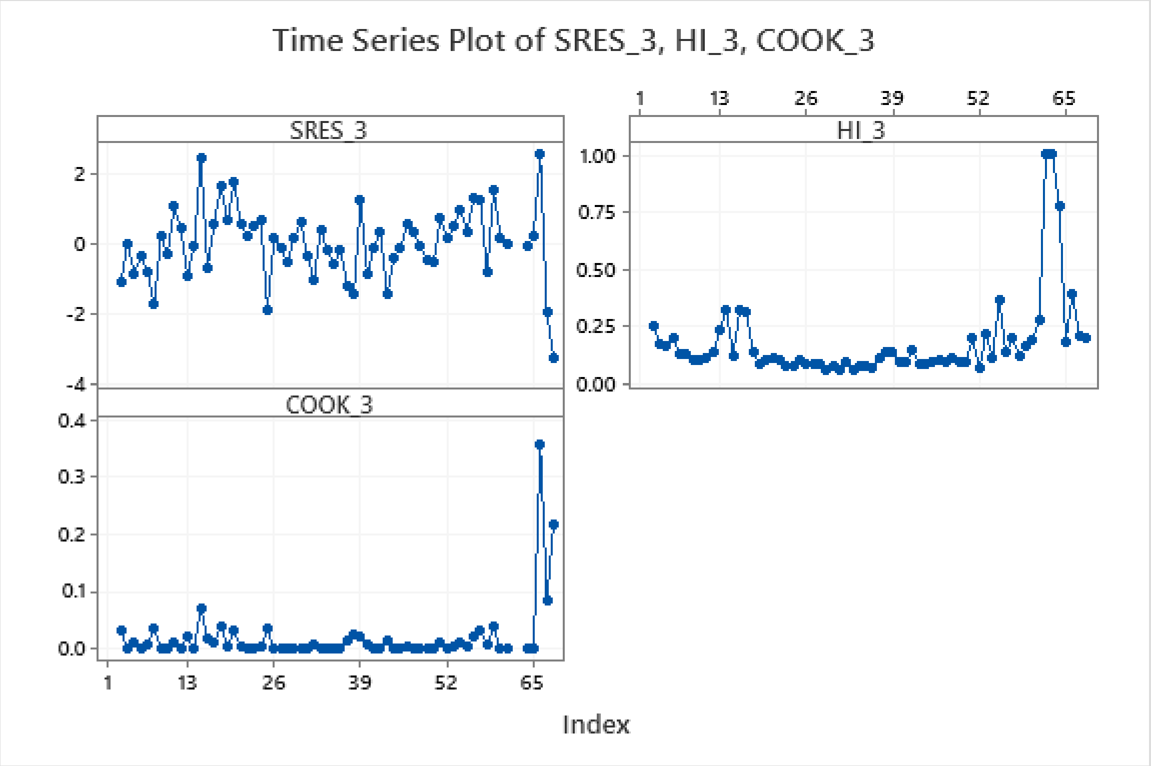
Obs	Cook's D	DFITS	
15	0.07	0.98142	R
62	*	*	X
63	*	*	X
64	0.00	-0.12135	X
66	0.36	2.18680	R
68	0.22	-1.78856	R

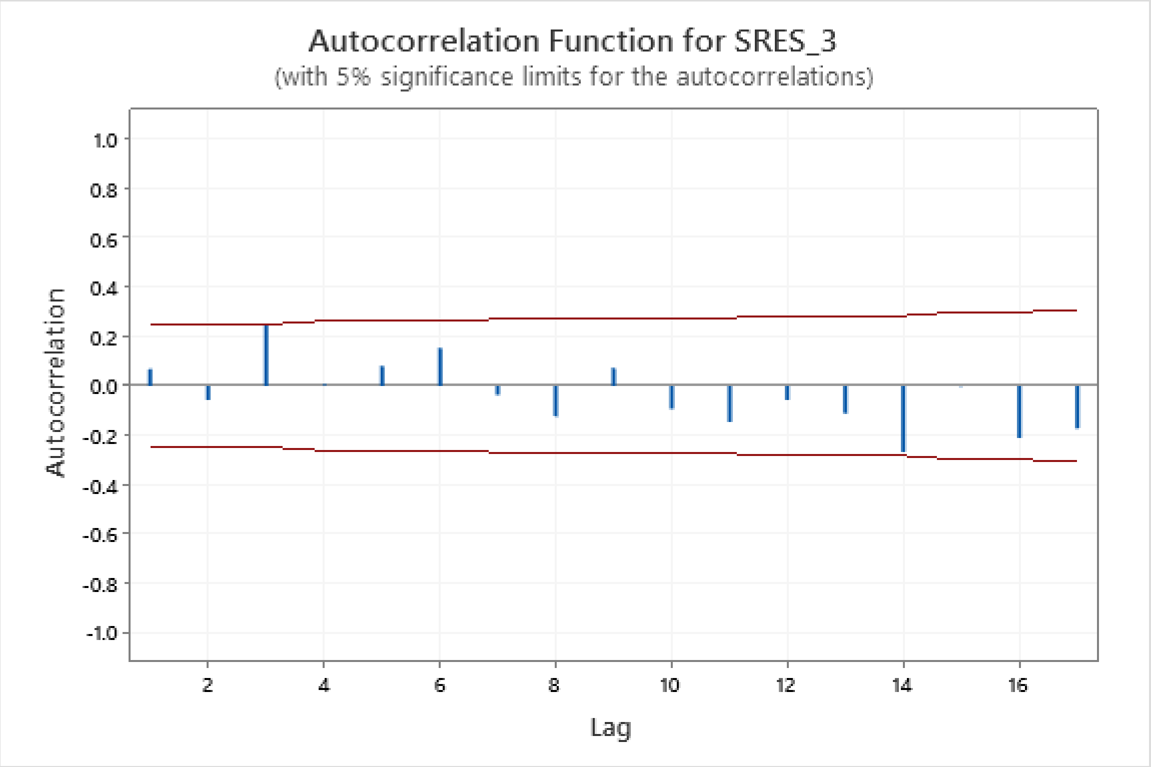
R Large residual
X Unusual X

Durbin-Watson Statistic

Durbin-Watson Statistic =	1.67270







Runs Test: SRES_3

Descriptive Statistics

		Number of Observations	
N	K	≤ K	> K
66	0.0036860	33	33

K = sample mean

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
29	34.00	0.215

Based on "Run Test", ACF result, we can see that the autocorrelation is removed. But the diagnosis results flag potential unusual point #64 with a high HI value of 0.75, which represents the fourth quarter of 2020 during the peak of COVID-19. So it was omitted through category indicator "CovidUnusual3" and the regression was rerun. The regression results changed a lot, so I decide to permanently remove it. Here is the regression result:

Regression Equation

LnUnemployment Rate % = 1.297 + 0.8298 CovidUnusual1 - 0.343 CovidUnusual2
- 0.0048 CovidUnusual3 - 0.185 LnPrimeInterestRate
+ 0.1527 LagLnPrimeInterestRate
- 0.03343 RealGDPChange %
+ 0.9594 LaglnUnemployRate - 0.110 LnExchangeRate
- 0.50 lnCPIlessfoodEnergy
- 0.02742 LagRealGDPChange
+ 0.174 LagLnExchangeRate
+ 0.23 LaglnCPIlessfoodEnergy

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.297	0.897	1.45	0.154	
CovidUnusual1	0.8298	0.0813	10.20	0.000	5.37
CovidUnusual2	-0.343	0.125	-2.75	0.008	12.65
CovidUnusual3	-0.0048	0.0735	-0.07	0.948	4.38
LnPrimeInterestRate	-0.185	0.106	-1.75	0.085	62.08
LagLnPrimeInterestRate	0.1527	0.0913	1.67	0.100	46.68
RealGDPChange %	- 0.03343	0.00892	-3.75	0.000	10.82
LaglnUnemployRate	0.9594	0.0407	23.57	0.000	9.45
LnExchangeRate	-0.110	0.313	-0.35	0.727	31.12
lnCPIlessfoodEnergy	-0.50	1.66	-0.30	0.766	880.87

LagRealGDPChange	-0.02742	0.00949	-2.89	0.006	12.12
LagLnExchangeRate	0.174	0.311	0.56	0.577	33.80
LaglnCPIlessfoodEnergy	0.23	1.71	0.14	0.893	911.30

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0348276	99.08%	98.87%	*

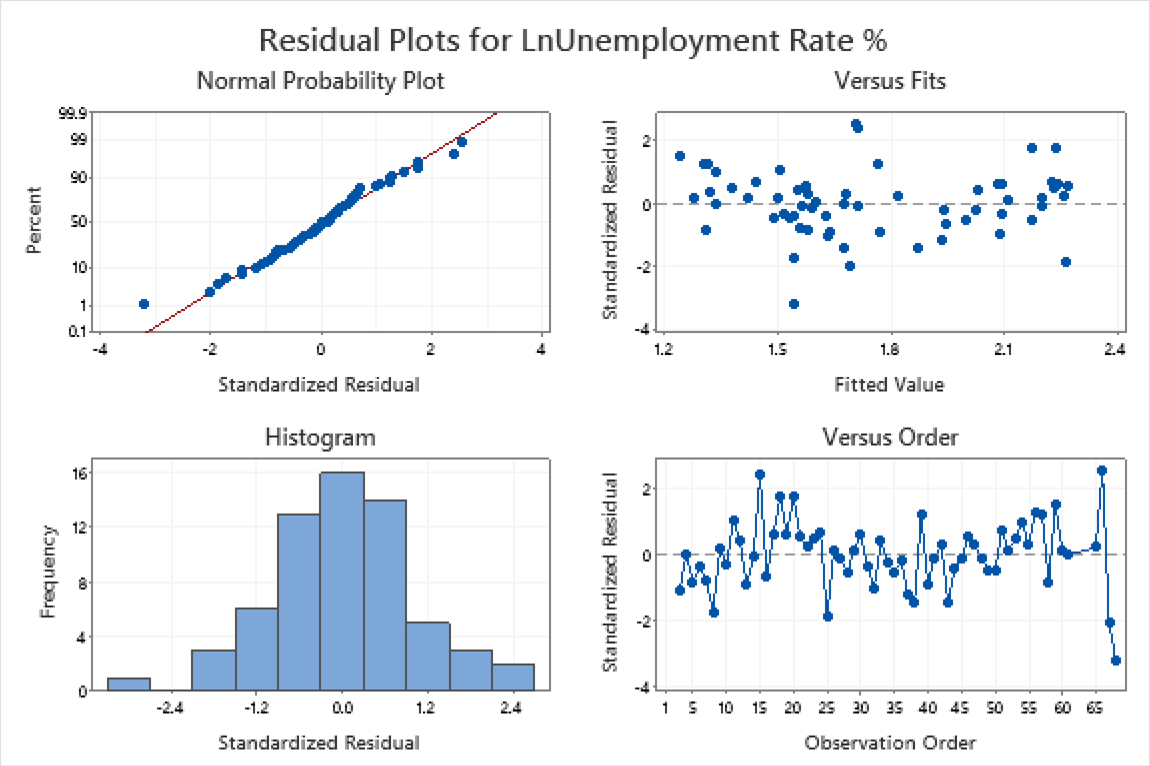
Analysis of Variance

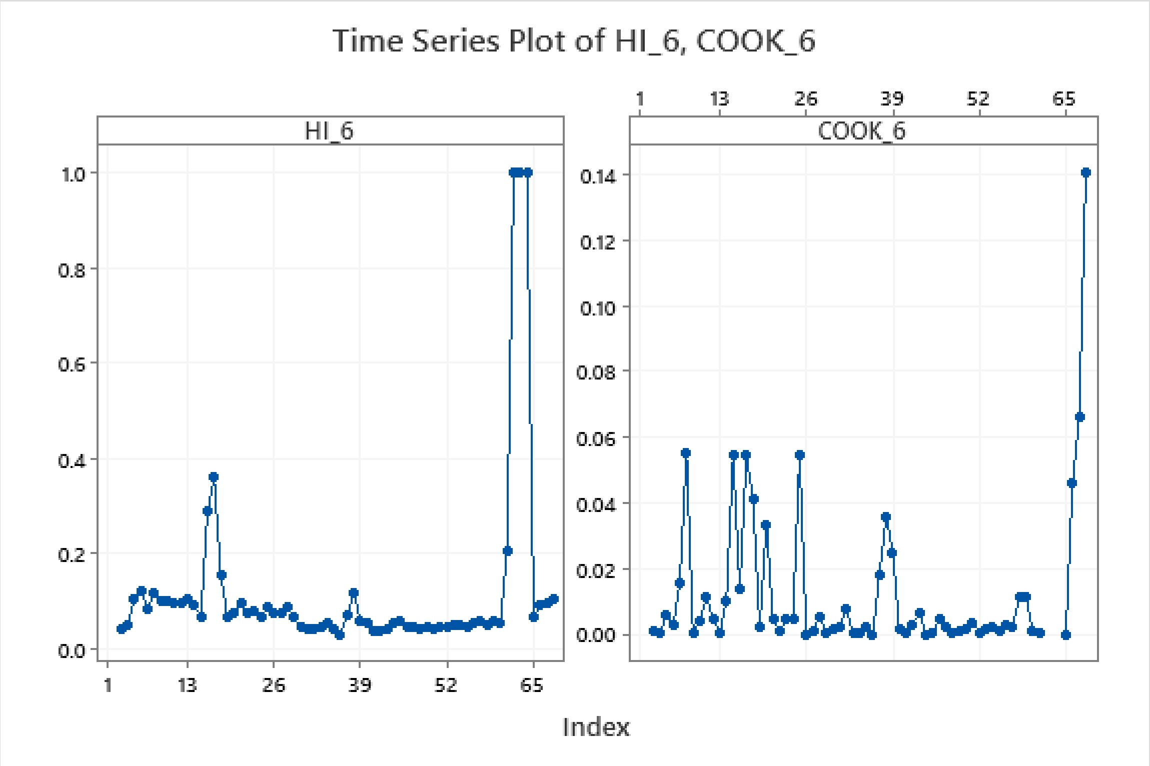
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	12	6.93966	0.578305	476.77	0.000
CovidUnusual1	1	0.12625	0.126248	104.08	0.000
CovidUnusual2	1	0.00917	0.009172	7.56	0.008
CovidUnusual3	1	0.00001	0.000005	0.00	0.948
LnPrimeInterestRate	1	0.00372	0.003725	3.07	0.085
LagLnPrimeInterestRate	1	0.00340	0.003396	2.80	0.100
RealGDPChange %	1	0.01703	0.017030	14.04	0.000
LaglnUnemployRate	1	0.67375	0.673753	555.46	0.000
LnExchangeRate	1	0.00015	0.000149	0.12	0.727
lnCPIlessfoodEnergy	1	0.00011	0.000108	0.09	0.766
LagRealGDPChange	1	0.01013	0.010133	8.35	0.006
LagLnExchangeRate	1	0.00038	0.000382	0.31	0.577
LaglnCPIlessfoodEnergy	1	0.00002	0.000022	0.02	0.893
Error	53	0.06429	0.001213		
Total	65	7.00394			

Fits and Diagnostics for Unusual Observations

	LnUnemployment				
Obs	Rate %	Fit	Resid	Std Resid	
15	1.7918	1.7131	0.0786	2.42 R	
62	2.5626	2.5626	-0.0000	*	X
63	2.1782	2.1782	0.0000	*	X
64	1.9125	1.9125	0.0000	*	X
66	1.7750	1.7071	0.0678	2.54 R	
67	1.6292	1.6897	-0.0605	-2.01 R	
68	1.4422	1.5415	-0.0993	-3.19 R	

R Large residual
X Unusual X


Residual Plots for LnUnemployment Rate %
Normal Probability Plot
Percent
99.9
99
90
50
10
1
0.1
-4
-2
0
2
4
Standardized Residual
Versus Fits
Standardized Residual
2
0
-2
-4
1.2
1.5
1.8
2.1
2.4
Fitted Value
Histogram
Frequency
16
12
8
4
0
-2.4
-1.2
0.0
1.2
2.4
Standardized Residual
Versus Order
Standardized Residual
2
0
-2
-4
1
5
10
15
20
25
30
35
40
45
50
55
60
65
Observation Order


Time Series Plot of HI_6, COOK_6
HI_6
1.0
0.8
0.6
0.4
0.2
0.0
1
13
26
39
52
65
COOK_6
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
1
13
26
39
52
65
Index

After omitting unusual points #64, the diagnosis result showed that points #16, #17, #61 had slightly high HI values, suggesting them to be potential unusual points. The regression result doesn't change much after omitting them, so I believe they aren't really unusual points, and they remain in the data set. I used a similar methodology as before to determine unusual points, so to save space the results are not shown.

To further identify the best model, I choose to use the best subsets regression method. Here is the output:

Best Subsets Regression:

Response is LnUnemployment Rate %

66 cases used, 2 cases contain missing values

Total Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC
4	97.1	96.9	*	*	110.2	0.057491	-181.485	-169.771
4	44.2	40.5	*	*	3168.0	0.25319	14.209	25.923
5	98.4	98.3	*	*	36.2	0.042701	-219.327	-205.931
5	98.2	98.0	*	*	51.1	0.046091	-209.243	-195.847
6	98.8	98.7	*	*	16.3	0.037473	-235.078	-220.087
6	98.7	98.6	*	*	20.7	0.038652	-230.990	-216.000
7	99.0	98.9	*	*	6.9	0.034487	-244.480	-227.988
7	99.0	98.9	*	*	6.9	0.034490	-244.468	-227.975
8	99.0	98.9	*	*	8.7	0.034731	-241.914	-224.017
8	99.0	98.9	*	*	8.7	0.034735	-241.896	-223.999
9	99.1	98.9	*	*	7.5	0.034033	-242.869	-223.672
9	99.1	98.9	*	*	7.5	0.034045	-242.825	-223.628
10	99.1	98.9	*	*	9.1	0.034230	-240.300	-219.911
10	99.1	98.9	*	*	9.2	0.034247	-240.234	-219.845
11	99.1	98.9	*	*	11.0	0.034510	-237.325	-215.859
11	99.1	98.9	*	*	11.1	0.034533	-237.236	-215.771
12	99.1	98.9	13365.3	0.0	13.0	0.034828	-234.112	-211.692

Total Vars	Cond No	LnPrimeIntere	LagLnPrimeInt	RealGDPChange	LaglnUnemploy	LnExchangeRa	lnCPIleftfood	LagLnExchange	LagRealGDPCha	LaglnCPIleftf

		stRate	erestRate	%	Rate	te	Service	Rate	nge	oodService
4	2.216				X					
4	1.385	X								
5	25.726			X	X					
5	29.163				X				X	
6	59.473			X	X				X	
6	27.416			X	X		X			
7	59.694			X	X				X	X
7	59.673			X	X		X		X	
8	66.832		X	X	X		X		X	
8	67.155		X	X	X				X	X
9	279.796	X	X	X	X		X		X	
9	278.472	X	X	X	X				X	X
10	351.750	X	X	X	X		X	X	X	
10	350.293	X	X	X	X			X	X	X
11	486.779	X	X	X	X	X	X	X	X	
11	487.240	X	X	X	X	X		X	X	X
12	8278.066	X	X	X	X	X	X	X	X	X

At your request, the best subsets procedure included these variables in every model:
CovidUnusual1, CovidUnusual2, CovidUnusual3

I looked at Mallows Cp and AICc and chose to minimize Cp, AICc, while considering R-Sq value. It seems that 7 variables models are the "best model" candidates. There are two 7 variables candidates, very close. I examined both of them. The above highlighted 7 variable model seems the "best model". Here is the regression result.

Regression Equation

LnUnemployment Rate % = 1.130 + 0.8070 CovidUnusual1 - 0.357 CovidUnusual2
+ 0.0433 CovidUnusual3 - 0.04126 RealGDPChange %
+ 0.9803 LaglnUnemployRate
- 0.03351 LagRealGDPChange
- 0.2217 LaglnCPIlessfoodEnergy

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.130	0.322	3.52	0.001	
CovidUnusual1	0.8070	0.0764	10.56	0.000	4.84

CovidUnusual2	-0.357	0.115	-3.11	0.003	10.96
CovidUnusual3	0.0433	0.0669	0.65	0.520	3.71
RealGDPChange %	-0.04126	0.00753	-5.48	0.000	7.85
LaglnUnemployRate	0.9803	0.0150	65.21	0.000	1.31
LagRealGDPChange	-0.03351	0.00832	-4.03	0.000	9.52
LaglnCPIlessfoodEnergy	-0.2217	0.0649	-3.41	0.001	1.35

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0344871	99.02%	98.90%	*

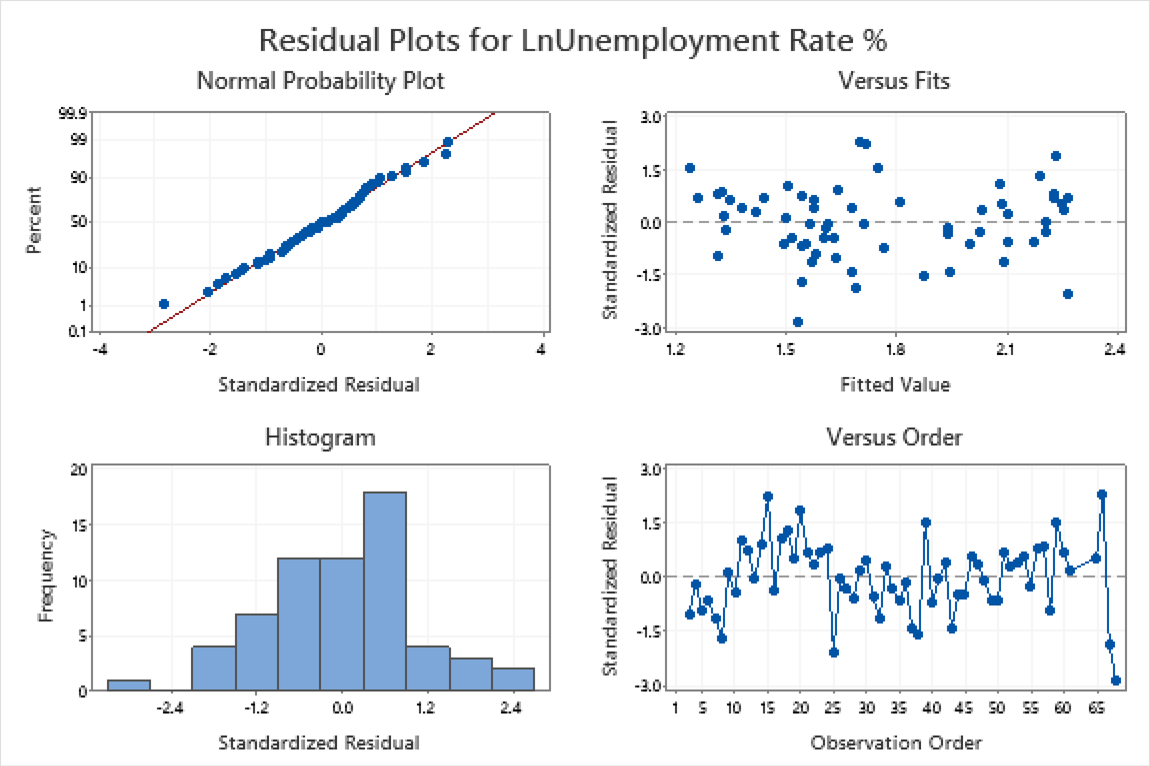
Analysis of Variance

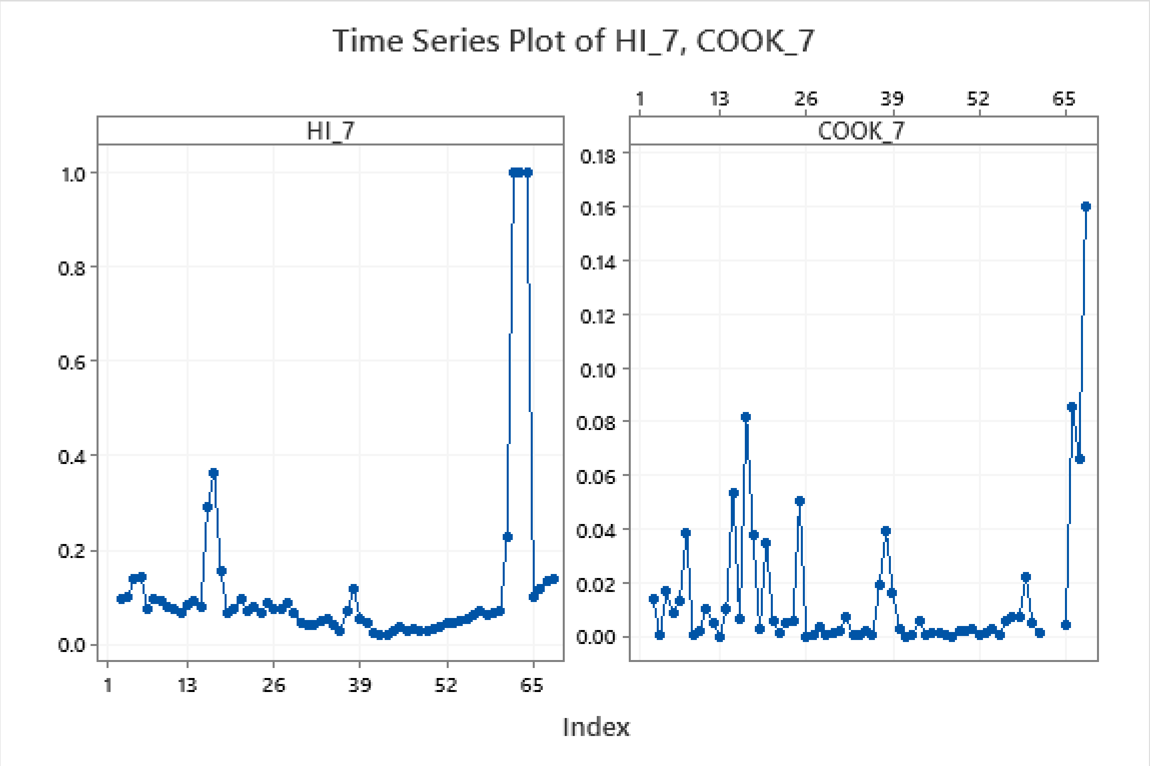
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	6.93496	0.99071	832.98	0.000
CovidUnusual1	1	0.13253	0.13253	111.43	0.000
CovidUnusual2	1	0.01148	0.01148	9.65	0.003
CovidUnusual3	1	0.00050	0.00050	0.42	0.520
RealGDPChange %	1	0.03575	0.03575	30.06	0.000
LaglnUnemployRate	1	5.05828	5.05828	4252.94	0.000
LagRealGDPChange	1	0.01927	0.01927	16.20	0.000
LaglnCPIlessfoodEnergy	1	0.01387	0.01387	11.66	0.001
Error	58	0.06898	0.00119		
Total	65	7.00394			

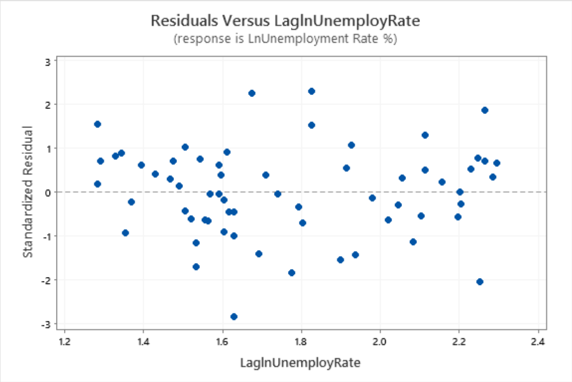
Fits and Diagnostics for Unusual Observations

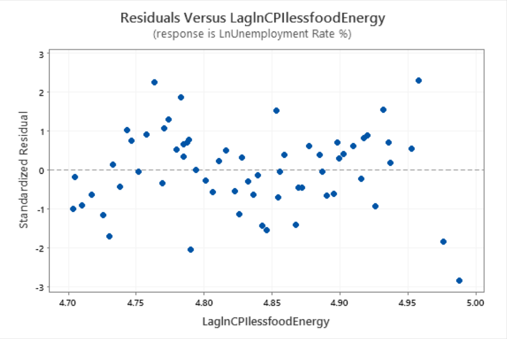
Obs	LnUnemployment Rate %	Fit	Resid	Std Resid	
15	1.7918	1.7173	0.0745	2.25	R
17	2.1126	2.0832	0.0294	1.07	X
25	2.2006	2.2679	-0.0673	-2.04	R
62	2.5626	2.5626	-0.0000	*	X
63	2.1782	2.1782	0.0000	*	X
64	1.9125	1.9125	0.0000	*	X
66	1.7750	1.7006	0.0743	2.29	R
68	1.4422	1.5331	-0.0909	-2.84	R

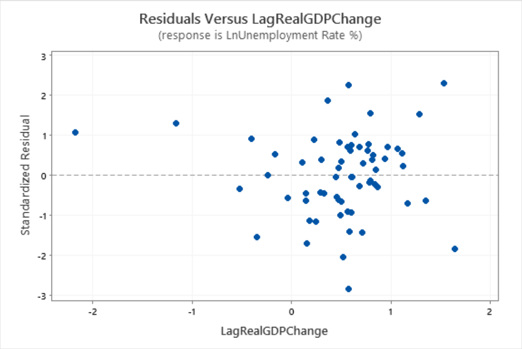
R Large residual
X Unusual X

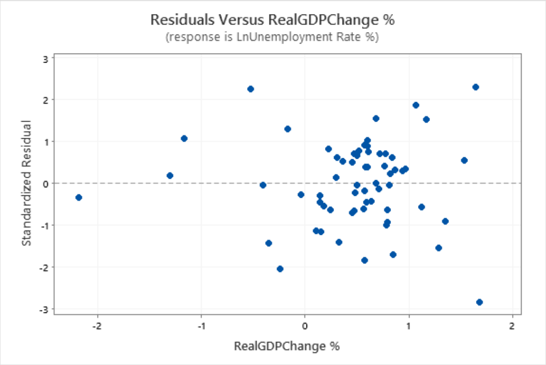

Residual Plots for LnUnemployment Rate %
Normal Probability Plot
Percent
99.9
99
90
50
10
1
0.1
-4
-2
0
2
4
Standardized Residual
Versus Fits
Standardized Residual
3.0
1.5
0.0
-1.5
-3.0
1.2
1.5
1.8
2.1
2.4
Fitted Value
Histogram
Frequency
20
15
10
5
0
-2.4
-1.2
0.0
1.2
2.4
Standardized Residual
Versus Order
Standardized Residual
3.0
1.5
0.0
-1.5
-3.0
1
5
10
15
20
25
30
35
40
45
50
55
60
65
Observation Order


Time Series Plot of HI_7, COOK_7
HI_7
1.0
0.8
0.6
0.4
0.2
0.0
1
13
26
39
52
65
COOK_7
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
1
13
26
39
52
65
Index









Runs Test: SRES_7

Descriptive Statistics

		Number of Observations	
N	K	≤ K	> K
66	0.0004350	34	32

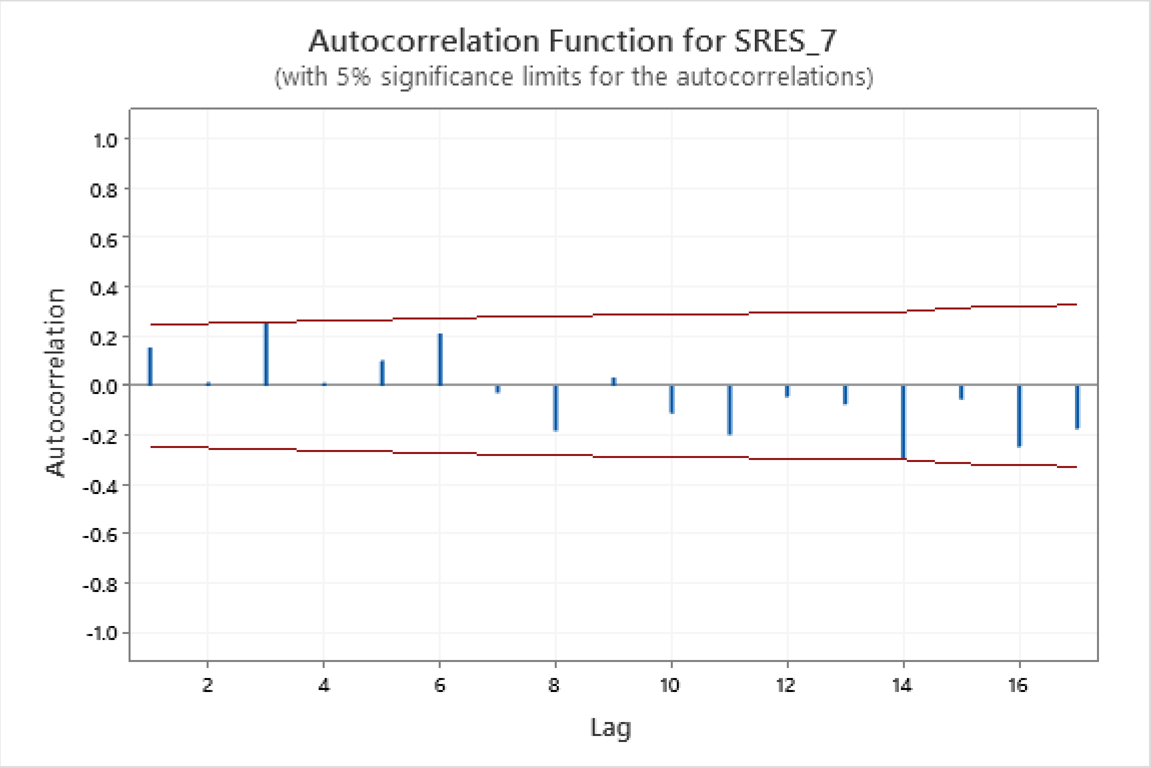
K = sample mean

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
29	33.97	0.217



We can therefore see that this model with four predictors: LaglnUnemployRate, RealGDPChange %, LagRealGDPChange, LaglnCPIlessfoodEnergy, is the "best model". Each of the variables is highly statistically significant, the model has a strong fit of R-sq= 99.02%. The statistical significance of variable CovidUnusual3 can be ignored since it is only used for omitting point #64. There does not seem to be a problem with collinearity as VIFs are fairly small. Diagnostics and residual plots look okay here. There is no obvious pattern on the plots, Regression assumptions are not violated. Most importantly, the autocorrelation has apparently been removed based on "Run Test" and ACF chart.

Regression Equation

LnUnemployment Rate % = 1.130 + 0.8070 CovidUnusual1 - 0.357 CovidUnusual2
+ 0.0433 CovidUnusual3 - 0.04126 RealGDPChange %
+ 0.9803 LaglnUnemployRate - 0.03351 LagRealGDPChange
- 0.2217 LaglnCPIlessfoodEnergy

The coefficient for LaglnUnemployRate" in the equation is 0.9803. what this means: given that RealGDPChange %, LagRealGDPChange, LaglnCPIlessfoodEnergy are held fixed, a one unit increase in the LaglnUnemployRate, or 2.718 percentage point increase in previous quarter unemployment rate (e^1=2.718) is associated with a 0.9803 unit increase in

LnUnemployment Rate % of current quarter, in other words, or 2.66 percentage point increase in Unemployment Rate%(e^0.9803=2.66).

The coefficient for RealGDPChange % in the equation is - 0.04126. what this means: given that LaglnUnemployRate, LagRealGDPChange, LaglnCPIlessfoodEnergy are held fixed, a one percentage point increase in RealGDPChange %, is associated with a 0.04126 unit decrease in LnUnemployment Rate %, in other words, or 0.9595 percentage point increase in Unemployment Rate% (e^(-0.04126)=0.9595).

The coefficient for LagRealGDPChange in the equation is -0.03351. what this means: given that LaglnUnemployRate, RealGDPChange %, LaglnCPIlessfoodEnergy are held fixed, a one percentage point increase in previous quarter RealGDPChange %, is associated with a 0.03351 unit decrease in LnUnemployment Rate %, in other words, 0.967 percentage point increase in Unemployment Rate% (e^(-0.03351)=0.967).

The coefficient for LaglnCPIlessfoodEnergy" in the equation is -0.2217. what this means: given that RealGDPChange %, LagRealGDPChange, LaglnUnemployRate, are held fixed, a one unit increase in the LaglnCPIlessfoodEnergy, or 2.718 index point increase in previous quarter CPIlessfoodEnergy, Customer Price Index less food and energy(e^1=2.718) is associated with a 0.2217 unit decrease in LnUnemployment Rate % of current quarter, in other words, 0.801 percentage point increase in Unemployment Rate%(e^(-0.2217)=0.801).

There are only a total 68 points available in the original data set (total 63 data point after omitting three unusual points, lagging the variable, calculating real GDP percent change), and the most recent 7 quarters are covid-19 data points, so it's difficult to validate the model selection process and chosen model on new data. To validate the predictive power of the chosen model, I chose to use leave-one-out cross-validation method in Minitab, which estimates the standard deviation of the errors using the standard deviation of the set of prediction errors of the observations as each is left out. Here is the result:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)	63-fold S	63-fold R-sq
0.0344871	98.88%	98.80%	98.63%	0.0366607	98.63%

"63-fold S", the leave-one-out cross-validation estimate of standard error σ is 0.0366607, roughly 6% larger than S=0.0344871, the value given in the regression output. Thus, a better rough 95% prediction interval for "LnUnemployment Rate %" when using this model is ±(2)(0.0366607) = ±0.0733214, rather than ±(2)(0.0344871) = ±0.06897, in other words, rough 95% prediction interval for "Unemployment Rate %" when using this model is e^±(2)(0.0366607) =e^ ±0.0733214= [0.929, 1.076] percentage point, rather than e^ ±(2)(0.0344871) = e^±0.06897= [0.933, 1.714] percentage point, with the former roughly 6% wider than the latter.

Overall this regression equation has an strong intuitive justification, and reflects the causal implication between unemployment and the chosen predictors. It also reflects the delayed impact of these indicators. This makes senses by thinking the character of the economic indicators unemployment rate, Interest rate, CPI, and Exchange Rate. By nature their impact is always somewhat delayed.

However, this analysis is limited by my knowledge of regression. The residual normality is a little off the line, and slightly shows some structure. I hope to further analyze this topic and improve my understanding of regression as the course continues.