Mengjie Shen

Jeffrey Simonoff

Regression and Multivariate Data Analysis

April 13, 2021

Regression on Time Series: Estimating the Death Rate in China

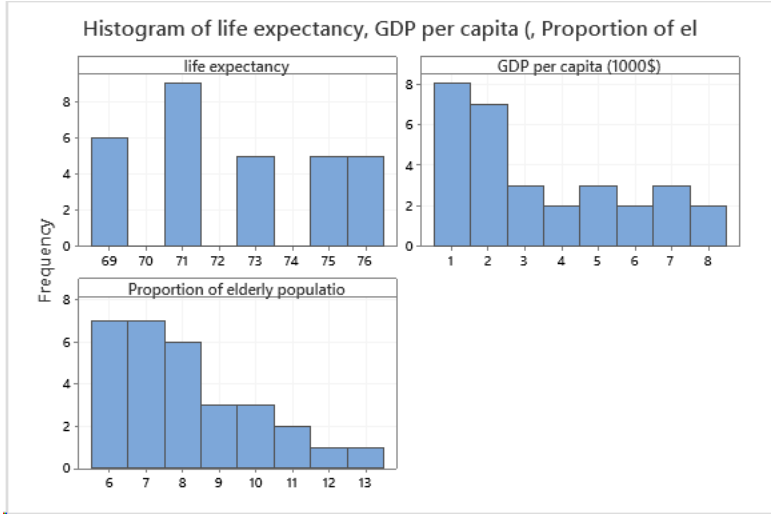
The death rate is an important indicator of the health status of the whole population. A relatively lower death rate usually indicates a more developed living condition. When the government makes policies, for example, the birth control policy in China, the death rate is an important factor that should be taken into account. Therefore, it will be very interesting to know what factors influence the death rate as well as predict the future death rate for better policymaking.

From my point of view, the death rate might be related to the health level of the whole population, society's degree of aging and level of economic development (I used GDP per capita to represent here). So, I will use the average life expectancy, the proportion of the population aged 65 years and over and GDP per capita in my report to model the death rate. I acquired the GDP per capita (constant 2010 US$) data from the website of the world bank(https://data.worldbank.org/indicator/NY.GDP.PCAP.KD?locations=CN). I use the version that converted to the constant dollar in order to make sure that the inflation won't take effect. I acquired the data of the population aged 65 years and over from the Chinese National Bureau of Statistics (https://data.stats.gov.cn/easyquery.htm?cn=C01). In order to make sure that population growth doesn't make any effect, I divided each value of observation by the number of the total population and get the percentage of the population aged 65 years and over. I also

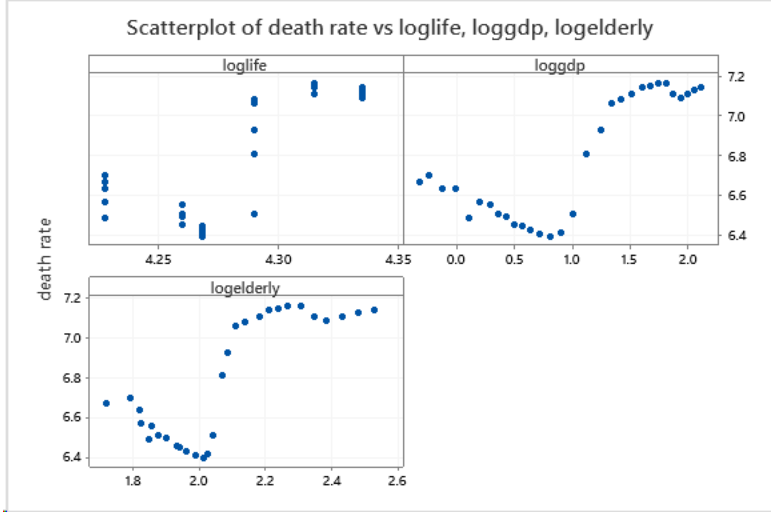
acquired the data on life expectancy and death rate from the Chinese National Bureau of Statistics. The graph down below is the head of the data.

year	Life Expectancy	GDP per capita(1000$)	Proportion of Elderly Population	Death Rate
1990	68.55	0.729282386	5.569695538	6.67
1991	68.55	0.786035813	5.990174663	6.7
1992	68.55	0.886909436	6.16022736	6.64
1993	68.55	0.998500351	6.150172549	6.64
1994	68.55	1.115987788	6.359616187	6.49
1995	68.55	1.224851893	6.200411159	6.57
1996	70.8	1.33235016	6.400084975	6.56
1997	70.8	1.440596917	6.539886432	6.51
1998	70.8	1.538788933	6.70001042	6.5

We can see from the histogram that there is a pattern of right-tailed distribution in all three of the predicting variables (life expectancy, GDP per capita and proportion of elderly population), we will take natural logs on these variables and fit the semi-log model.



Here are the scatter plots of each logged variable versus death rate.



We witness a strong pattern in all three scatter plots that as each variable goes up, the death rate goes down at first and then goes up.

It's a very strange trend. This actually corresponds to two time periods, before and after 2005. Under the trend of the development of the economy and medical level, the death rate usually goes down. The reason for the rise in mortality here may be the age structure of the population. However, there aren't any resources online indicating why it's the year 2005 that is the turning point of the death rate. My guess is that in 2005, the average life expectancy of the Chinese population was around 73. The death in 2005 was born in roughly the 1930s. If we look at the historical data of Chinese population growth, we can see that from 1932 and 1933, the births went from more than 5 million before to more than 7 million, and then stabilized at around 8 million, so the mortality rate was bound to rise. As the number of births jumped into the tens of millions starting in 1945. Especially after the founding of the country, it soared to 20 million, the mortality rate is very likely to keep climbing in the following year. Worth noticing that the above is just my interpretation which might not be totally correct.

Let's then fit the regression and see the result:

Regression Equation

death rate = 9.5 - 0.94 loglife + 0.209 loggdp + 0.534 logelderly

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	9.5	20.2	0.47	0.642	
loglife	-0.94	4.82	-0.20	0.847	26.34
loggdp	0.209	0.293	0.71	0.482	39.45
logelderly	0.534	0.741	0.72	0.478	22.43

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.189302	64.50%	60.40%	54.55%

Analysis of Variance

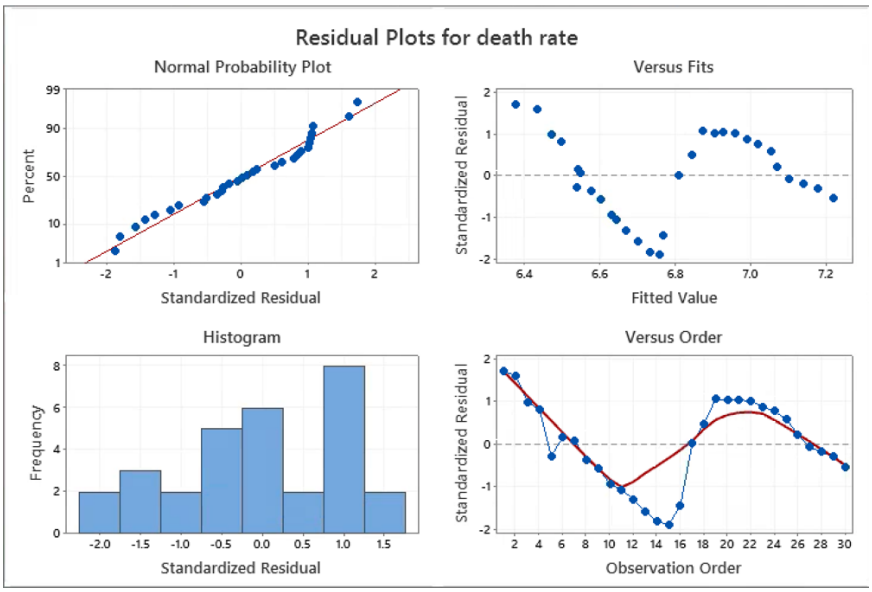
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	1.69283	0.564276	15.75	0.000
loglife	1	0.00136	0.001364	0.04	0.847
loggdp	1	0.01826	0.018257	0.51	0.482
logelderly	1	0.01861	0.018606	0.52	0.478
Error	26	0.93172	0.035835		
Total	29	2.62455			

Durbin-Watson Statistic

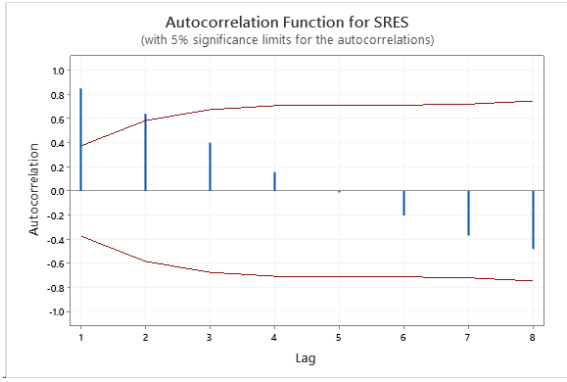
Durbin-Watson Statistic = 0.187591

The regression fit is very strong, but VIF indicates the issue of collinearity. The overall F-test is highly statistically significant. However, the t-test for each variable is not significant. This is probably due to the collinearity issue. We will use model selection in the following process to see whether we can address the issue, but now, let's focus on identifying the autocorrelation issue.

There is a pattern of non-normality in the residual plot. It is worth noticing that it also indicates a time trend left in the residuals. If we add a smooth curve to the plot, we can observe a non-linear pattern. The Durbin-Watson test is highly statistically significant, also indicating the time trend.



An ACF plot of the residuals and the runs test also agrees that there is autocorrelation present.



Descriptive Statistics

Number of Observations			
N	K	≤ K	> K
30	0	14	16

Test

Null hypothesis H₀: The order of the data is random

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

Number of Runs		
Observed	Expected	P-Value
6	15.93	0.000

Therefore, let's add the linear and quadratic terms in time and run the regression again:

Regression Equation

death rate = 13.0 - 0.24 loglife + 2.912 loggdp - 2.55 logelderly - 0.2638 time
+ 0.003687 timesq

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	13.0	10.2	1.28	0.213	
loglife	-0.24	2.31	-0.10	0.918	27.52
loggdp	2.912	0.468	6.23	0.000	455.44
logelderly	-2.55	1.09	-2.33	0.028	221.03
time	-0.2638	0.0526	-5.02	0.000	785.27
timesq	0.003687	0.000554	6.65	0.000	78.59

Model Summary

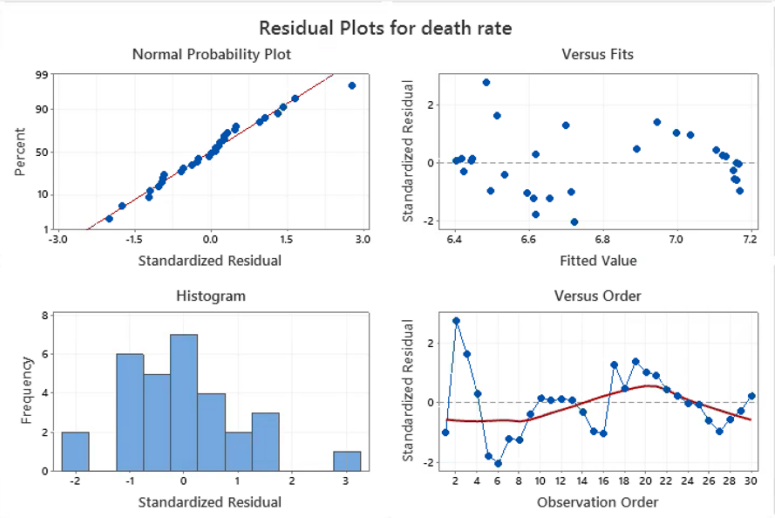
S	R-sq	R-sq(adj)	R-sq(pred)
0.0889234	92.77%	91.26%	86.91%

Durbin-Watson Statistic

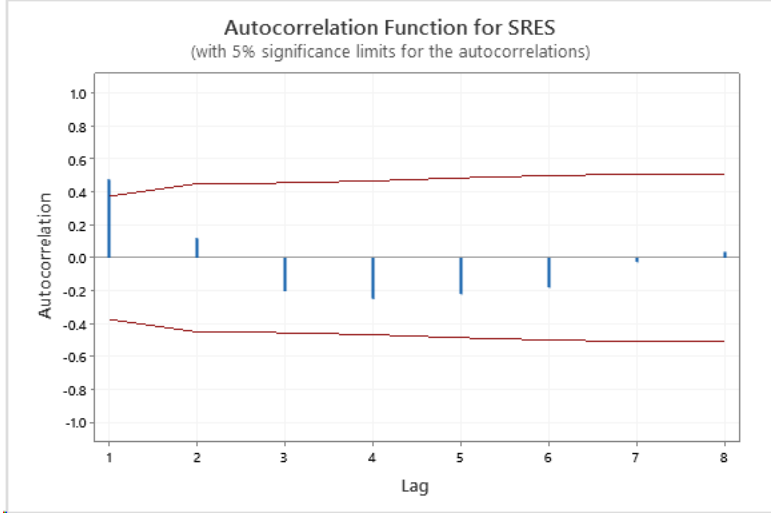
Durbin-Watson Statistic = 0.943321

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	2.43477	0.486954	61.58	0.000
loglife	1	0.00009	0.000086	0.01	0.918
loggdp	1	0.30669	0.306688	38.79	0.000
logelderly	1	0.04301	0.043010	5.44	0.028
time	1	0.19911	0.199110	25.18	0.000
timesq	1	0.35016	0.350160	44.28	0.000
Error	24	0.18978	0.007907		
Total	29	2.62455			



R^2 improves a lot, with an increase of roughly thirty percentage points. VIF shows that collinearity still exists. The t-test indicates that the variable log of life expectancy is no longer necessary, but here we will continue to concentrate on the autocorrelation issues. The Durbin-Watson indicates that auto correlation issue still exists and so does the runs test and the ACF plot. There is a pattern of non-normality and right-tailed distribution of the residuals and the pattern of non-constant variance shows in the residual versus fits plot.



Descriptive Statistics

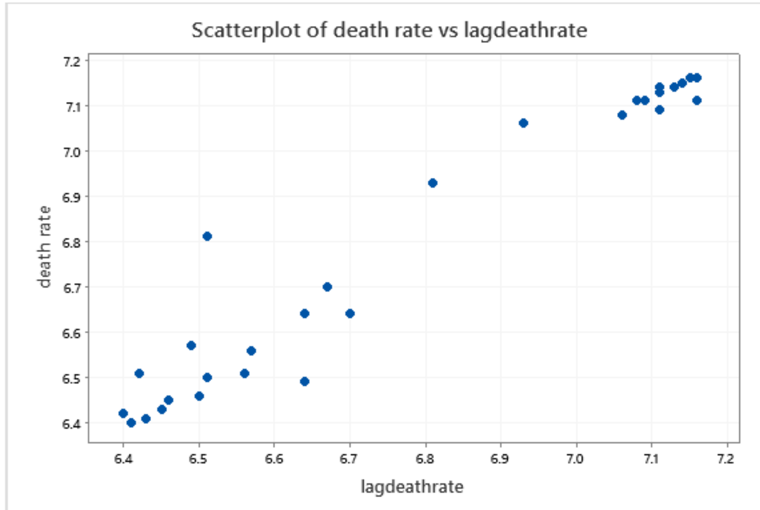
N K	Number of Observations ≤ K	> K
30 0	14	16

Test

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

Number of Runs Observed	Expected	P-Value
8	15.93	0.003

Since detrending can't account for all the autocorrelation effect, we next take the approach of adding a lag variable into the model. We can see from the scatterplot down below that there is a direct relationship between lagged death rate and the death rate.



We first perform the best subsets regression to simplify and choose the most reasonable model:

Response is death rate

29 cases used, 1 cases contain missing values

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No
1	93.6	93.4	0.2	92.8	15.0	0.078606	-60.326	-57.184	1.000
1	71.2	70.1	0.9	67.3	155.6	0.16700	-16.618	-13.477	1.000
2	94.6	94.1	0.2	93.1	11.0	0.073863	-62.324	-58.521	8.381
2	94.4	94.0	0.2	93.0	11.9	0.074812	-61.583	-57.780	7.859
3	95.2	94.6	0.2	93.5	9.0	0.070770	-62.999	-58.772	184.436
3	95.2	94.6	0.2	93.4	9.1	0.070801	-62.974	-58.747	108.945
4	96.5	95.9	0.1	94.8	3.1	0.061930	-68.713	-64.327	3049.978
4	96.5	95.9	0.1	94.9	3.2	0.062068	-68.583	-64.198	3306.040
5	96.5	95.7	0.2	93.8	5.0	0.063196	-65.258	-61.021	18936.622
5	96.5	95.7	0.1	94.3	5.1	0.063228	-65.229	-60.991	3856.569
6	96.5	95.5	0.2	93.1	7.0	0.064574	-61.430	-57.691	22791.443

Vars	time	timesq	loglife	loggdp	logelderly	lagdeathrate
1						X
1		X				
2				X		X
2	X					X
3	X				X	X
3				X	X	X
4		X		X	X	X
4	X	X			X	X
5	X	X		X	X	X
5		X	X	X	X	X
6	X	X	X	X	X	X

We will go with the four-predictor model (time squared, logged GDP per capita, logged proportion of the elderly population and lagged death rate), since it's the best model under the criteria of maximizing Adjusted R^2, minimizing Mallows C_p and minimizing AIC_C.

Regression Equation

death rate = 9.03 + 0.3867 loggdp - 3.83 logelderly + 0.002467 timesq + 0.6807 lagdeathrate

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	9.03	2.59	3.49	0.002	
loggdp	0.3867	0.0946	4.09	0.000	34.42
logelderly	-3.83	1.15	-3.32	0.003	460.12
timesq	0.002467	0.000838	2.95	0.007	354.91
lagdeathrate	0.6807	0.0993	6.85	0.000	6.42

Model Summary

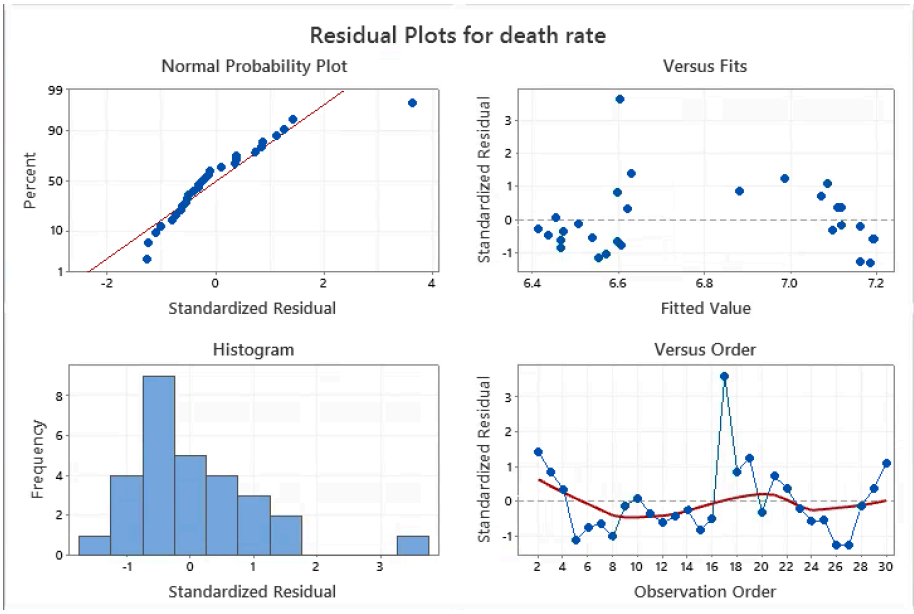
S	R-sq	R-sq(adj)	R-sq(pred)
0.0619301	96.47%	95.89%	94.76%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	2.51890	0.629724	164.19	0.000
loggdp	1	0.06406	0.064056	16.70	0.000
logelderly	1	0.04239	0.042393	11.05	0.003
timesq	1	0.03327	0.033270	8.67	0.007
lagdeathrate	1	0.18012	0.180118	46.96	0.000
Error	24	0.09205	0.003835		
Total	28	2.61094			

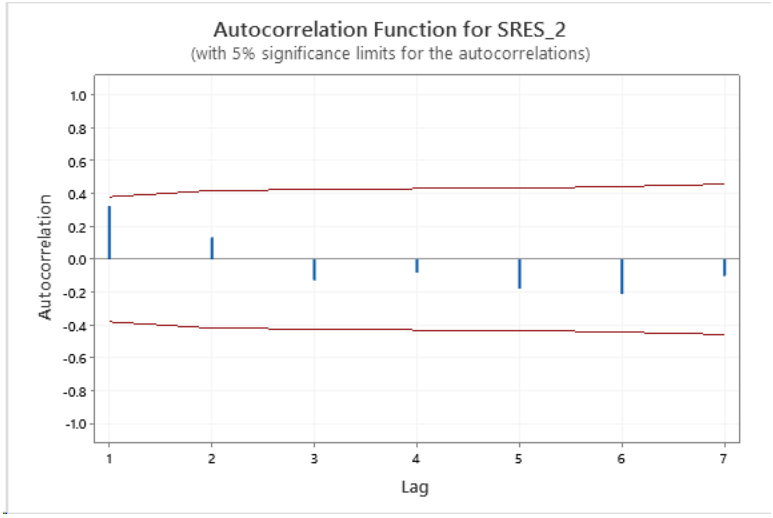
VIF indicates that the issue of collinearity still exists after model selection. Although there is collinearity, but the t-test indicates that each variable has significant predicting power. The t-test for the lag death rate is highly statistically significant. The coefficient for logged GDP per capita means that multiplying GDP per capita by e is associated with an expected increase of 0.3867 percentage point in the death rate, holding all else fixed. The coefficient for the lag death rate means holding all else fixed, 1 percentage point increase in the death rate last year is associated with 0.68 percentage point increase in death rate this year. The coefficient for time squared means that given all else is held fixed, the estimated expected time-related death rate

growth is the derivative of $(\beta x^2), which\ is\ 2\beta x$. Thus, given all else is fixed, the estimated expected time-related death rate growth is 0.004 percentage point $(2 \times 0.002 \times 1 = 0.004)$ in 1990. Therefore, there will be an increasing growth rate in death rate if the current condition persists.



There is a pattern of non-normality and right-tailed distribution of the residuals and the pattern of non-constant variance shows in the plot of residual versus fits. There seems to be an outlier on the top right corner. We will look into this observation later.

An ACF plot shows that autocorrelation is no longer the issue, but worth noticing that when lag equals one, it almost exceeds the boundary of 5% significance limits for the autocorrelation.



The p-value for the runs test increases, but it still shows that autocorrelation exists.

Descriptive Statistics

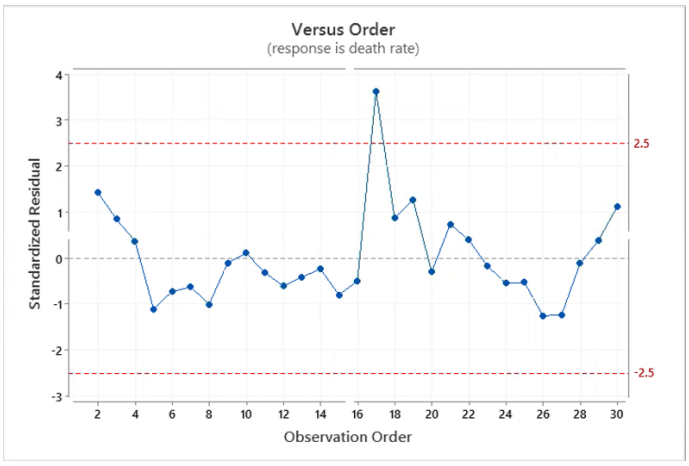
	Number of Observations		
N	K	≤ K	> K
29	0	18	11

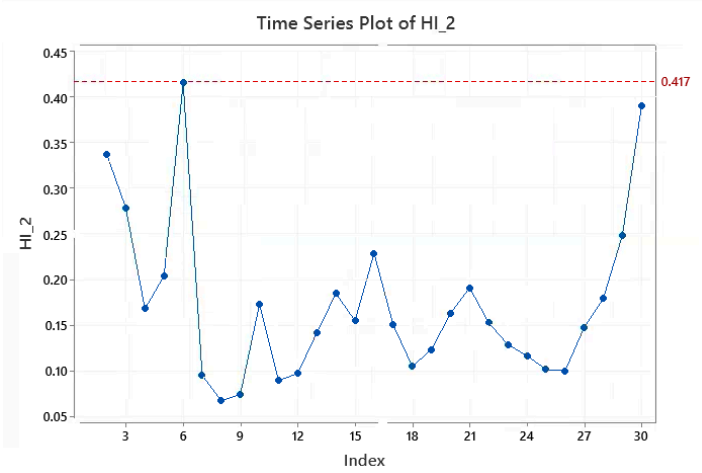
Test

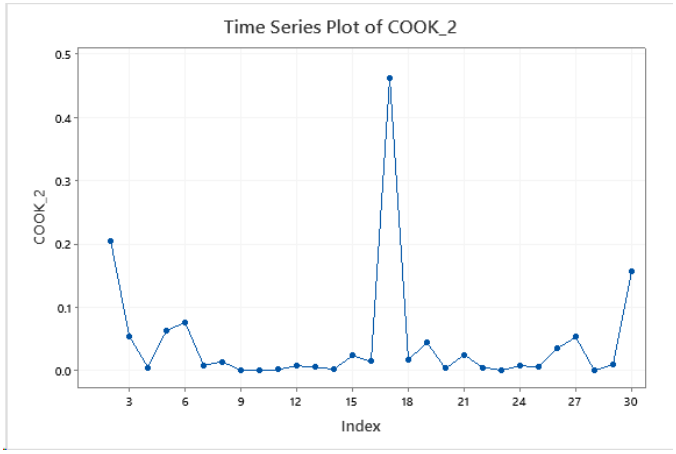
Null hypothesis H₀: The order of the data is random
Alternative hypothesis H₁: The order of the data is not random

Number of Runs		
Observed	Expected	P-Value
9	14.66	0.023

Next, we calculate the standardized residuals, leverages and Cook's distance to identify the outlier and leverage point.







We can see from the plot of standardized residuals that the 17th observation appears to be an outlier, which corresponds to the year 2006. This point also seems unusual in the plot of the Cook's distance. There is a point on the plot of leverages that almost touches the red line, which

corresponds to the year 1995. However, since it doesn't significantly influence the regression, we won't dive deeper into this border point.

Now, let's pay attention to the outliers of the year 2006. There wasn't any major natural disaster that happened in 2006 and I didn't find any concrete evidence or specific research online that explains this outlier. My guess for the reason is the change in the birth rate because birth rate affects the total population and thus affects the death rate. After looking into the birth-controlled policy in China, I found that from 1978 to 2001, the birth-controlled policy was strictly enforced, and from 2002 to the present, the birth-controlled policy has been gradually relaxed because of the current situation of the aging society. As a result, the number of births began to decline in 1991 and bottomed out in around 2006 and after 2006, the birth rate became stabilized and went up again. Since total population growth hits the bottom in 2006, the death rate is very likely to be relatively high comparing to adjacent years.

Since the death rate of 6.81% is relatively too high, we substitute it with the average value of the year 2015 and the year 2017, which is 6.72(%). It is worth noticing that we didn't adjust the corresponding value in lagged death rate.

We will run the regression again. Since we have modified data, so we performed model selection again.

Response is death rate adj

29 cases used, 1 cases contain missing values

Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC	Cond No
1	95.3	95.1	0.1	94.7	20.0	0.067719	-68.972	-65.830	1.000
1	71.3	70.3	0.8	67.5	247.8	0.16668	-16.730	-13.589	1.000
2	96.0	95.7	0.1	94.9	15.4	0.063707	-70.902	-67.100	8.381
2	95.9	95.5	0.1	94.9	16.4	0.064542	-70.147	-66.344	7.859
3	96.6	96.1	0.1	95.3	11.8	0.060047	-72.530	-68.302	184.436
3	96.5	96.1	0.1	95.2	12.1	0.060301	-72.285	-68.057	108.945
4	97.7	97.3	0.1	96.6	3.1	0.050333	-80.739	-76.353	3306.040
4	97.6	97.2	0.1	96.5	3.4	0.050707	-80.309	-75.923	3049.978
5	97.7	97.2	0.1	96.3	5.0	0.051299	-77.355	-73.118	4149.333
5	97.7	97.2	0.1	95.9	5.1	0.051404	-77.236	-72.998	18936.622
6	97.7	97.1	0.1	95.4	7.0	0.052438	-73.505	-69.766	22791.443

Vars	logelderly	laggedeathrate	time	timesq
1		X		
1				X
2	X	X		
2		X	X	
3		X	X	X
3	X	X	X	
4		X	X	X
4	X	X	X	
5	X	X	X	X
5	X	X	X	X
6	X	X	X	X

We will go with the four-predictor model (time, time squared, the logged proportion of the elderly population and lagged death rate), because it is the best model under the criteria of maximizing Adjusted R^2, maximizing R^2, minimizing Mallows C_p and minimizing AIC_C.

Here is the result of the four-predictor model:

Regression Equation

death rate adj = 8.53 - 3.984 logelderly + 0.002307 timesq + 0.7817 lagdeathrate
+ 0.03894 time

Coefficients

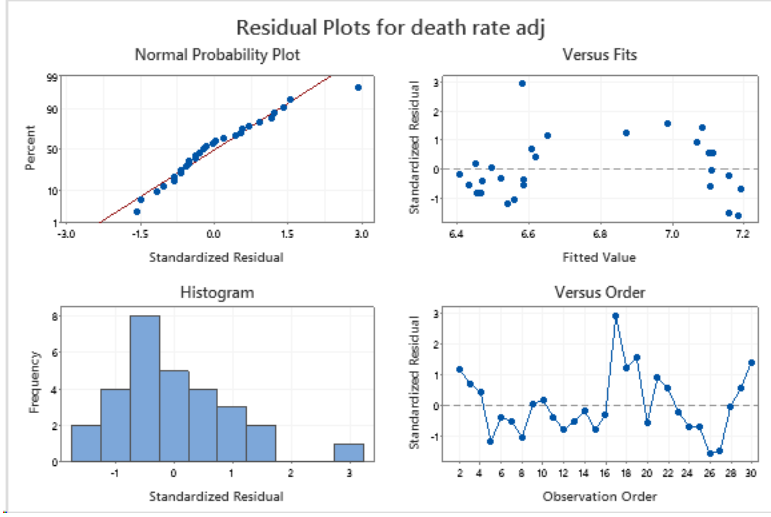
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	8.53	2.13	4.01	0.001	
logelderly	-3.984	0.999	-3.99	0.001	524.73
timesq	0.002307	0.000678	3.40	0.002	351.87
lagdeathrate	0.7817	0.0750	10.43	0.000	5.53
time	0.03894	0.00831	4.69	0.000	55.35

Model Summary

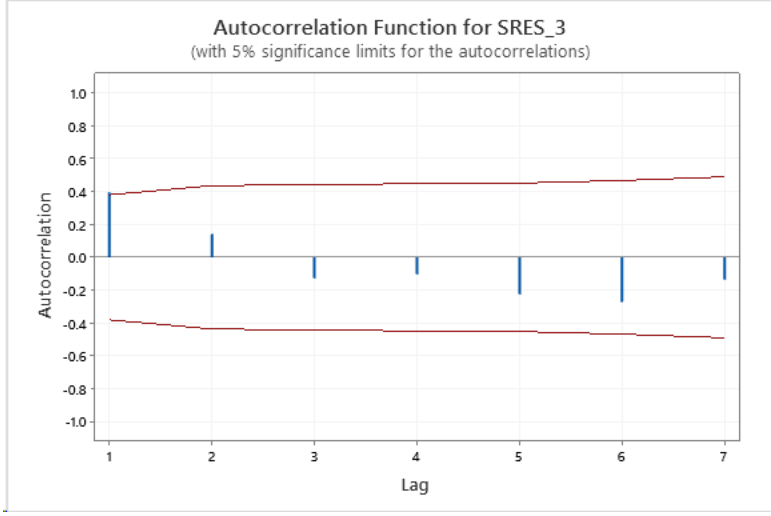
S	R-sq	R-sq(adj)	R-sq(pred)
0.0503333	97.67%	97.29%	96.60%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	2.55411	0.638529	252.04	0.000
logelderly	1	0.04034	0.040337	15.92	0.001
timesq	1	0.02934	0.029338	11.58	0.002
lagdeathrate	1	0.27540	0.275402	108.71	0.000
time	1	0.05561	0.055610	21.95	0.000
Error	24	0.06080	0.002533		
Total	28	2.61492			



The model slightly improved. However, even though we substitute the outlier observation with the average value of 6.72%. It still identifies as an outlier. The non-normality and right-tailed distribution persist as well as heteroscedasticity. The pattern of autocorrelation becomes slightly more significant in the ACF plot. The runs-test also shows that the autocorrelation is significant.



Descriptive Statistics

N K	Number of Observations ≤ K	> K
29 0	17	12

Test

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

Number of Runs Observed	Expected	P-Value
9	15.07	0.018

Next, we take a different approach to deal with the outlier. We add a 0/1 variable (Year 2006) to indicate whether the observation is in the year 2006.

Response is death rate

29 cases used, 1 cases contain missing values

Total Vars	R-Sq	R-Sq (adj)	PRESS	R-Sq (pred)	Mallows Cp	S	AICc	BIC
2	96.8	96.6	*	*	23.4	0.056346	-78.024	-74.221
2	71.3	69.1	*	*	398.3	0.16977	-14.054	-10.251
3	97.1	96.8	*	*	21.3	0.054884	-77.744	-73.516
3	97.1	96.7	*	*	22.1	0.055365	-77.238	-73.010
4	97.6	97.2	*	*	16.1	0.051015	-79.958	-75.573
4	97.5	97.1	*	*	17.1	0.051710	-79.174	-74.788
5	98.5	98.2	*	*	4.9	0.041183	-90.095	-85.857
5	98.4	98.1	*	*	6.5	0.042600	-88.133	-83.895
6	98.5	98.1	*	*	6.4	0.041623	-86.900	-83.162
6	98.5	98.1	116145.5	0.0	6.5	0.041679	-86.823	-83.084
7	98.6	98.1	41421.9	0.0	8.0	0.042173	-83.215	-80.383

Total Vars	Cond No	logelderly	lagdeathrate	timesq	time	Year2006
2	1.407		X			
2	1.059				X	
3	9.497	X	X			
3	8.830		X	X		
4	187.480		X	X	X	
4	111.180	X	X	X		
5	3368.700		X	X	X	X
5	3124.940	X	X	X		X
6	19752.659	X	X	X	X	X
6	4230.245	X		X	X	X
7	23790.847	X	X	X	X	X

At your request, the best subsets procedure included these variables in every model: Year2006

We will go with the five-predictor model (Year 2006, time, time squared, logged proportion of the elderly population and lagged death rate) under the criteria of maximizing Adjusted R^2, maximizing R^2, minimizing Mallows C_p and minimizing AIC_C.

Here is the result of the regression:

Regression Equation

death rate = 7.44 - 3.586 logelderly + 0.002076 timesq + 0.8398 lagdeathrate + 0.03414 time + 0.2490 Year2006

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	7.44	1.77	4.21	0.000	
logelderly	-3.586	0.825	-4.35	0.000	534.47
timesq	0.002076	0.000558	3.72	0.001	356.60
lagdeathrate	0.8398	0.0635	13.24	0.000	5.92
time	0.03414	0.00693	4.93	0.000	57.50
Year2006	0.2490	0.0444	5.61	0.000	1.12

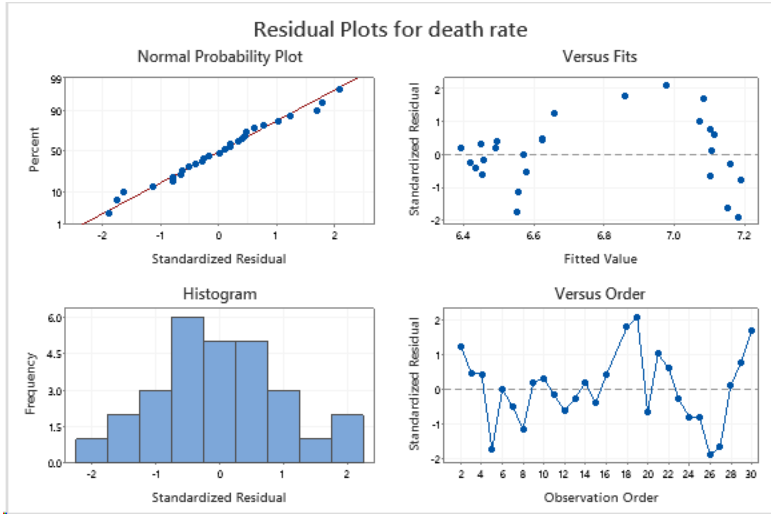
Model Summary

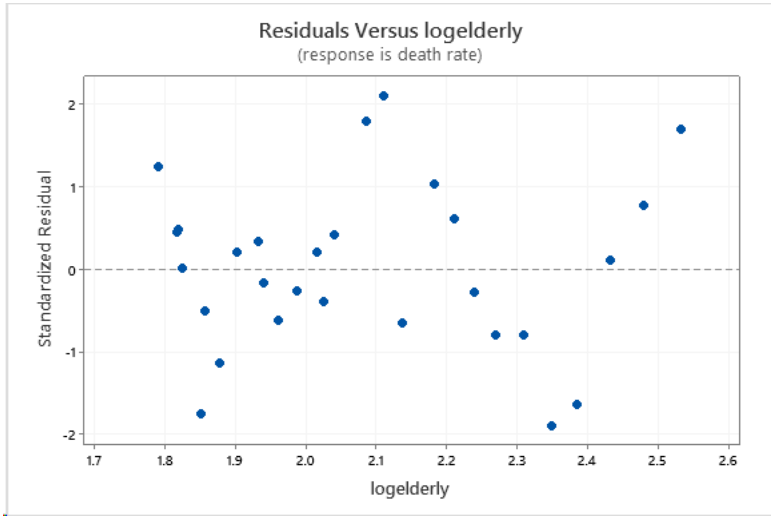
S	R-sq	R-sq(adj)	R-sq(pred)
0.0411828	98.51%	98.18%	*

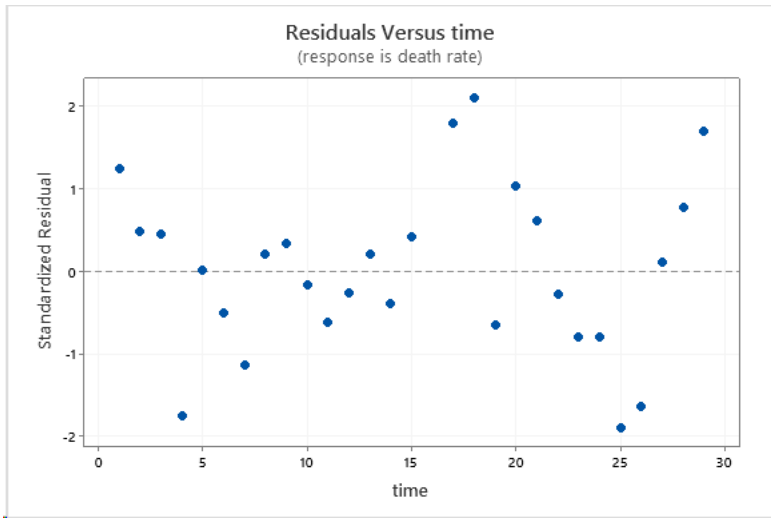
Analysis of Variance

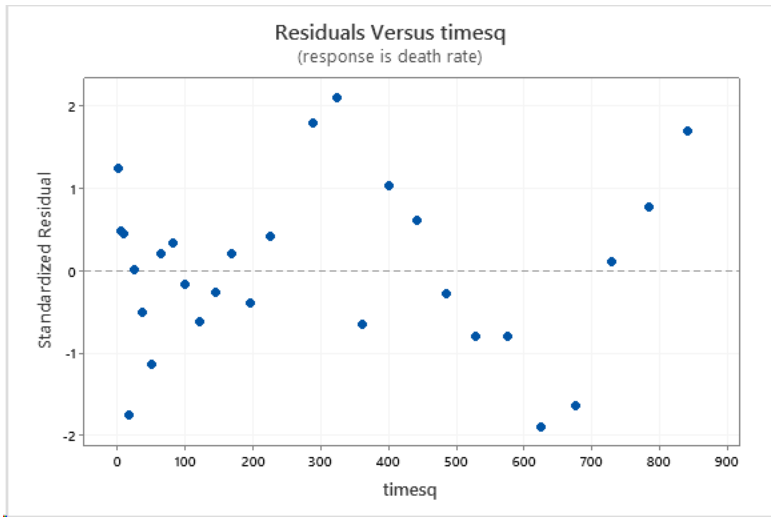
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	5	2.57194	0.514387	303.29	0.000
logelderly	1	0.03207	0.032069	18.91	0.000
timesq	1	0.02345	0.023452	13.83	0.001
lagdeathrate	1	0.29712	0.297123	175.19	0.000
time	1	0.04114	0.041144	24.26	0.000
Year2006	1	0.05345	0.053451	31.52	0.000
Error	23	0.03901	0.001696		
Total	28	2.61094			

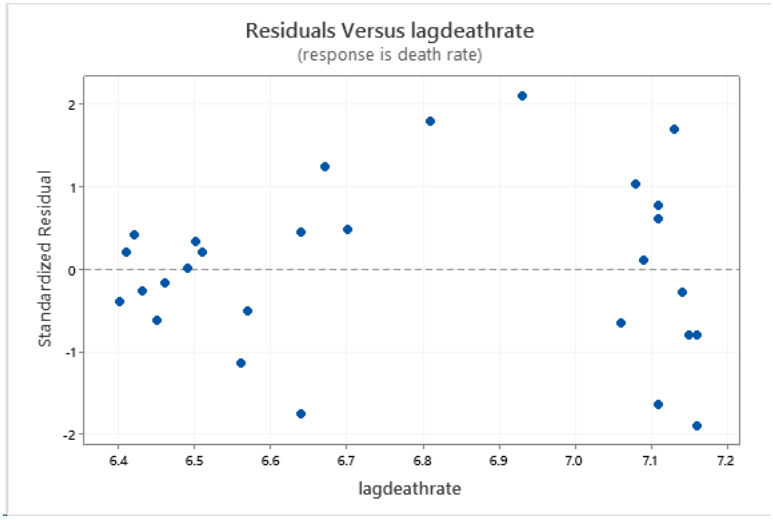
The coefficient for the proportion of the elderly population means that multiplying the proportion of elderly population by e is associated with an expected increase of -3.586 percentage point in y, holding all else fixed. The coefficient for the lag death rate means holding all else fixed, 1 percentage point increase in the death rate last year is associated with a 0.84 percentage point increase in death rate this year. The coefficient for time and time squared means that given all else is held fixed, the estimated expected time-related death rate growth is the derivative of $(\beta_1 x + \beta_2 x^2), which\ is\ \beta_1 + 2\beta_2 x$. Thus, given all else is fixed, the estimated expected time-related death rate growth is 0.034 percentage point $(0.03 \times 1 + 2 \times 0.002 \times 1 = 0.034)$ in 1990. Therefore, there will be an increasing growth rate in death rate unless there are some unexpected changes. The t-test for Year 2006 indicates that it is an outlier. The coefficient for Year2006 indicates that given every other variable fixed, the observed death rate is 0.25 percentage points higher than expected that year.


Residual Plots for death rate
Normal Probability Plot
Percent
Standardized Residual
Versus Fits
Standardized Residual
Fitted Value
Histogram
Frequency
Standardized Residual
Versus Order
Standardized Residual
Observation Order


Residuals Versus logelderly
(response is death rate)
Standardized Residual
logelderly

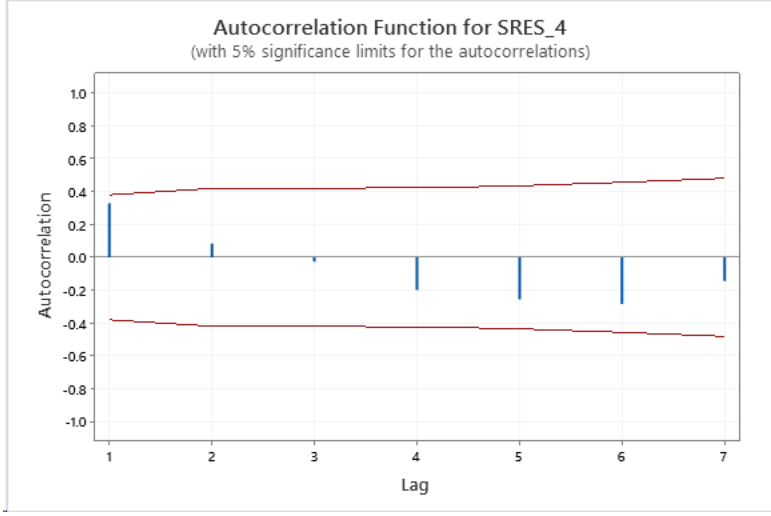

Residuals Versus time
(response is death rate)
Standardized Residual
time





The pattern of non-normality and right-tailed distribution seems to be weakened (but still exists). However, heteroscedasticity still occurs in the residual versus fitted value. In the residuals versus each predictor plot, we can also observe heteroscedasticity. It is especially distinguishable in the plot of residuals versus lagged death rate and the plot of residuals versus time squared.

The ACF plot and runs test indicate the autocorrelation disappears.



The p-value for the runs-test is not significant. So, the runs-test also agrees that there is no autocorrelation.

Descriptive Statistics

N	K	Number of Observations ≤ K	> K
29	0	14	15

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	15.48	0.855

Since our model has already achieved a good fit, it is unlikely that any action will make much of a difference. So, we will stop here for a conclusion.

To conclude, the model which time, time squared, the logged proportion of the elderly population, lagged death rate and indicator variable Year 2006 to eliminate the outlier is the best model. Moreover, it eliminates the autocorrelation in the end. I was hoping for a stronger effect of life expectancy and GDP per capita on the death rate. However, it turns out that they have little predicting power, comparing to the time and proportion of the elderly population. The time and time squared variable indicate that if the current situation precedes, the death rate is going to

keep climbing at an increasing speed. This is consistent with the fact that China is stepping into an aging society. In terms of practical use, since this model can be used to estimate the future death rate, it can better help the government to adjust the childbirth policy in order to adapt to the situation of the aging society.